The
University
Of
Sheffield.

NAMED ENTITY RECOGNITION – CHALLENGES IN DOCUMENT ANNOTATION, GAZETTEER CONSTRUCTION AND DISAMBIGUATION

ZIQI ZHANG

March 2013

Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

Department of Computer Science

Supervisor: Professor Fabio Ciravegna

DEDICATION

This thesis is dedicated to my brilliant wife, Yaxin Liu, for her infinite love and support throughout the course of this work.

ABSTRACT

The 'information explosion' has generated unprecedented amount of published information that is still growing at an astonishing rate. As the amount of information grows, the problem of managing the information becomes challenging. A key to this challenge rests on the technology of Information Extraction, which automatically transforms unstructured textual data into structured representation that can be interpreted and manipulated by machines. It is recognised that a fundamental task in Information Extraction is Named Entity Recognition, the goals of which are identifying references of named entities in unstructured documents, and classifying them into pre-defined semantic categories. Further, due to the polysemous nature of natural language, name references are often ambiguous. Resolving ambiguity concerns recognising the true referent entity of a name reference, essentially a further named entity 'recognition' step and often a compulsory process required by tasks built on top of NER.

This research presents a body of work aimed at addressing three research questions for NER. The first question concerns effective and efficient methods for training data annotation, which is the task of creating essential training examples for machine learning based NER methods. The second question studies automatically generating background knowledge for NER in the form of gazetteers, which are often critical resources to improve the performance of NER methods. The third question addresses resolving ambiguous name references, a further 'recognition' step that ensures the output of NER to be usable by many complex tasks and applications.

For each research question, the related literature has been carefully studied and their limitations have been identified and discussed. New hypotheses and methods have been proposed, leading to a number of contributions:

- an approach to training data annotation for supervised NER methods, based on the study of annotator suitability and suitability based task allocation;
- a method of automatically expanding existing gazetteers of pre-defined semantic categories exploiting the structure and knowledge of Wikipedia;
- a method of automatically generating untyped gazetteers for NER based on the "topic-representativeness" of words in documents;
- a method of named entity disambiguation based on maximising the semantic relatedness between candidate entities in a text discourse;

- a review of lexical semantic relatedness measures; and a new lexical semantic relatedness measure that harnesses knowledge from different resources.

The proposed methods have been evaluated by carefully designed experiments, following the standard practice in each related research area. The results have confirmed the validity of their corresponding hypotheses, as well as the empirical effectiveness of these methods. Overall it is believed that this research has made solid contribution to the research of NER and related areas.

LIST OF PUBLICATIONS RELATED TO THIS WORK

Zhang, Z., Cohn, T., and Ciravegna, F. 2013**.** Topic-oriented Words as Features for Named Entity Recognition**.** In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*. (Chapter 7)

Zhang, Z., Gentile, A., and Ciravegna, F. 2012. Recent Advances in Methods of Lexical Semantic Relatedness. In *Journal of Natural Language Engineering (Available on CJO doi:10.1017/S1351324912000125)* (Chapter 8)

Zhang, Z., Gentile, A., and Ciravegna, F. 2011. Harnessing Different Knowledge Sources to Measure Semantic Relatedness under a Uniform Model. In *Proceedings of the 2011 International Conference on Empirical Methods in Natural Language Processing (EMNLP), Edinburgh, UK*, pages 991-1002. (Chapter 8)

Zhang, Z. and Ciravegna, F. 2011. Named Entity Recognition for Ontology Population using Background Knowledge from Wikipedia, In Wang, W., Liu, W., Bennamoun, M. (Eds.), *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*, page 26, IGI Global. (Chapter 6)

Zhang, Z. 2011. jatetoolkit – Java Automatic Term Extraction toolkit. Last retrieved on 08 Sep 2012. URL: http://code.google.com/p/jatetoolkit/. (Chapter 7)

Zhang. Z., Chapman, S., and Ciravegna, F. 2010. A Methodology towards Effective and Efficient Manual Document Annotation: Addressing Annotator Discrepancy and Annotation Quality. In *Proceedings of the 17th International Conference on Knowledge Engineering and Management by the Masses (EKAW), Lisbon, Portugal*, pages 301-305. (*Shortlisted for the best paper award*). (Chapter 5)

Gentile, A., **Zhang, Z.,** Xia, L. and Iria, J. 2010. Cultural Knowledge for Named Entity Disambiguation: a Graph-based Semantic Relatedness Approach. In the *Serdica Journal of Computing*, *Volume 4, Issue 2,* pages: 217-242. (Chapter 9)

Zhang, Z., Iria, J. and Ciravegna, F. 2010. Improving Domain-specific Entity Recognition with Automatic Term Recognition and Feature Extraction. In *Proceedings of the 7th International conference on Language Resources and Evaluation (LREC), Valletta, Malta.* (Chapter 7)

Zhang, Z., Gentile, A., Xia, L., Iria, J. and Chapman, S. 2010. A Random Graph Walk based Approach to Compute Semantic Relatedness Using Knowledge from Wikipedia. In *Proceedings of the 7th International conference on Language Resources and Evaluation (LREC), Valletta, Malta.* (Chapter 8)

Gentile, A., **Zhang, Z.,** Xia, L. and Iria, J. 2010. Semantic Relatedness Approach for Named Entity Disambiguation. In *Proceedings of the 6th Italian Research Conference on Digital Libraries (IRCDL), Padua, Italy,* pages 137-148. (Chapter 9)

Zhang, Z. and Iria, J. 2009. A Novel Approach to Automatic Gazetteer Generation using Wikipedia. In *Proceedings of the Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources in the Joint conference of the 47th Annual Meet-*

ing of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing, (IJCNLP), Singapore, pages 1-9. (Chapter 6)

Gentile, A., **Zhang, Z.**, Xia, L. and Iria, J. 2009. Graph-based Semantic Relatedness for Named Entity Disambiguation. In *Proceedings of the 1st International Conference on Software, Services and Semantic Technologies (S3T)*, Sofia, Bulgaria, pages 13-20. (*Winner of the best student paper award*). (Chapter 9)

ACKNOWLEDGEMENT

It is a great pleasure to express my respect to the many people who have supported me throughout my doctoral study at the University of Sheffield.

First and foremost, I would like to express my deep and sincere gratitude to my supervisor, Professor Fabio Ciravegna for his invaluable advice, constant guidance and encouragement that has made this work possible. He has generously and patiently spent much time discussing research ideas with me, pointing me to important research questions while giving me freedom to explore my own interests. His vision and unique perspective at research have always inspired me, and have provoked more research question I wish to pursue in the future.

I would also like to thank all the other thesis committee members, Dr. Trevor Cohn and Dr. Victoria Uren, for their constructive comments and suggestions that made this thesis complete. Special thanks to Dr. Trevor Cohn for his generous time spent on discussions with me and his expertise in Machine Learning and Natural Language Processing that has particularly helped many parts of this work.

During my study, I have also received help from many colleagues of the Organisations, Information and Knowledge group (OAK) at the Department of Computer Science of The University of Sheffield. I would like to thank all the members of OAK, particularly for their feedback to my work at various stages of my doctoral study.

Finally I am grateful to my wife Yaxin and my parents, who always supported me with unceasing love and encouragement, and helped me through all the difficulties throughout this venture.

DECLARATION

I declare that this report was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification.

(*Ziqi Zhang*)

LIST OF ABBREVIATIONS

ACE	Advanced Content Extraction evaluation conference
ATR	Automatic Term Recognition
CRF	Conditional Random Field
HMM	Hidden Markov Model
IAA	Inter Annotator Agreement
IE	Information Extraction
IR	Information Retrieval
KA	Knowledge Acquisition
ME	Maximum Entropy
MEMM	Maximum Entropy Markov Model
MUC	Message Understanding Conference
MUC6	The Sixth Message Understanding Conference
MUC7	The Seventh Message Understanding Conference
NE	Named Entity
NED	Named Entity Disambiguation
NER	Named Entity Recognition
QA	Question Answering
SVM	Support Vector Machine
WSD	Word Sense Disambiguation

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

PREFACE

This chapter provides an overview of the research questions and objectives of the work in this thesis. It is divided into five sections. Section 1 introduces the motivation to this research and a brief introduction to the research area. Section 2 discusses the research questions that this thesis aims to address. Section 3 introduces the research hypothesises. Section 4 discusses the main contributions of this research and Section 5 outlines the structure of this thesis.

1.1 Motivation

We live in the Information Age. In every moment, an enormous amount of information is generated on the Internet, adding to its already gigantic size. Access to such a massive amount of information has totally changed the way we work and study. For organisations, possession and effective utilisation of information is deemed as a key part of strategic competitiveness. On the other hand, the scale and the scope of the information that one has to deal with at a time are also unprecedented, which makes locating useful pieces of information extremely difficult. The amount of accessible information would not be of much use if there were no suitable techniques to process it and extract knowledge from it.

The answer to this challenge is the technology of **Information Extraction (IE)**, the technique for transforming unstructured textual data into structured representation that can be understood by machines. IE has been an active research field for decades, involving many sub-topics that are addressed by rigorous communities. It originates from a set of earlier competitions organised within the Natural Language Processing (NLP) community. One of the most important is the Message Understanding Conference (MUC) (Grishman and Sundheim, 1996) where an earlier primary goal was to identify **mentions** or **names** of **entities** from unstructured news articles and classify them into predefined **semantic categories**. In brief, an **entity** is a unique real word object, such as ‘George Walker Bush, born July 6, 1946, an American politician who served as the 43rd President of the United States, from 2001 to 2009’; a **mention** or **name** is a lexicalised expression used to designate an entity, such as ‘George Bush’; and a **semantic category** is a high level concept that groups same types of entities, such as ‘people’, ‘place’, ‘organisation’, and ‘temporal’ and ‘numerical expressions’. This task is called **Named Entity Recognition (NER)**, a term first coined at the sixth Message Understanding Conference (MUC6) (Grishman and Sundheim, 1996), which was hosted to encourage research for Information Extraction (IE) from unstructured texts. It was recognised at the time that an essential step to enable other IE tasks was to identify these important information units from texts. To name a few, named entities and their semantic categories must be identified before identifying relations (Giuliano et al., 2006; Giuliano et al., 2007; Thahir et al., 2011) between them and extracting events (Smith, 2002; Zhang et al., 2007) involving entities. In populating knowledge bases such as ontologies (Cimiano, 2006; Giuliano and Gliozzo, 2008), named entities must be extracted from texts and classified into fine-grained ontological concepts. In addition to IE, NER is also an important technology for many other applications and research areas. In Information Retrieval (IR) and Question Answering (QA), named entities are identified to locate important information and facts (Lee et al., 2007; Srihari and Peterson, 2008). For example, in the question answering competition in TREC-8 (TREC-8 QA Data, 2002), 80% of the evaluation questions ask for a named entity (Nadeau, 2007a). In enabling the Semantic Web, NER is used to improve semantic search (Caputo et al., 2009). In machine translation, accurate translation of named entities plays an important role in the translation of the overall text (Babych and Hartley, 2003). Also in many domain specific contexts, domain specific NER is the key technology for constructing terminology resources (Nenadić et al., 2003; Byrne, 2007; Saha et al., 2009).

The techniques for NER can be divided into two branches: **handcrafted rules** and **learning based** methods (Sarawagi, 2007). Methods based on handcrafted rules require developers to manually create extraction rules usually expressed as lexico-syntactic patterns and semantic constraints that hypothesize the occurrences of similar named entities. Learning based methods automatically induce extraction patterns or sequence labelling algorithms from a collection of training examples. Learning based methods have proved to be more flexible and robust than handcrafted rules, since they lessen the requirements on linguistic knowledge and reduces human effort to only providing sufficient amount of examples. Therefore, they have become the more popular approach to NER (Nadeau, 2007a).

An essential input to learning based methods are **training data**, which usually take the form of documents containing **annotations** that are labelled instances of example named entities. Therefore training data are also called **labelled data** or **training annotations**. Training data have to be manually created by humans, a process that is often time consuming and costly. To address this, recent research has branched out to study methods that require less training data, which has created the stream of semi-supervised methods that learn using both annotated (training) and unannotated data (Chung et al., 2003; Kozareva et al., 2005; Olsson, 2008), and unsupervised methods that learn without or with very few training data (Da Silva et al., 2004; Cimiano and Völker, 2005). To contrast these areas, learning methods that use only annotated training data will be referred to as supervised learning methods in the remainder of this thesis.

Although supervised learning methods have been criticised for their overly dependence on training data, they remain the primary choice in many research and application areas.

In particular, supervised learning methods still dominate in many resource-poor languages (AbdelRahman et al., 2010; Duan and Zheng, 2011; Singh, 2011) and technical domains (Byrne, 2007; Iria, 2009a; Ju et al., 2011). Although there is the lack of comparative evaluation of supervised learning methods against others on the same datasets, some studies of semi-supervised learning methods have shown compromised accuracies when compared against the best results reported for the supervised learning competitors (Gu et al., 2007).

One important problem neglected by NER is the ambiguity in the extracted names. A name is **ambiguous** if it can be used to refer to different entities. For example, 'Washington' can refer to different locations or persons. While NER recognises the mention and assigns general semantic categories or labels, it does not answer what exact entity it refers to. Resolving ambiguities is traditionally a sense disambiguation task and treated separately from NER. However, theoretically, the two carry similar goals – NER can be considered as disambiguation at a higher level (Wacholder et al., 1997) while name disambiguation can be considered as a further step of 'recognition' where the true identity of a name mention is uncovered. Practically disambiguation is often an essential post-process to enable NER output to be useful for other complex NLP applications. This has been widely recognised and major evaluation campaigns have proposed to deal with the two tasks simultaneously (TAC KBP Track, 2009; TAC KBP Track, 2010).

1.2 Research Questions

The above problem setting motivates the work explored in this thesis. This thesis focuses on supervised learning methods for NER due to their significance in this field of research. Central to this thesis is the following research question:

How to effectively recognise named entities from texts?

This thesis views 'recognition' essentially a process fulfilling three goals: *identifying named entities in text, assigning semantic categories to these named entities,* and *assigning referent entities to them (i.e.,* **disambiguation***)*. Therefore, this question is further divided into three related research questions each contributing to the overall objective. The three questions are:

1. *How to create training data effectively and efficiently to enable supervised NER?*

2. *How to automatically generate background knowledge in the form of gazetteers to improve the accuracy of NER?*
3. *How to resolve ambiguities in the extracted names and recognise the unique referent entities, which makes the NER output useful to other applications?*

1.2.1 Training Data Annotation

As mentioned before, annotated training data are essential input to supervised learning methods. These act as examples to the learning algorithm, which induces a model able to predict similar instances of the same types in new data. Training data are annotated by domain experts, and must be of good quality and sufficient quantity in order to ensure an effective model to be learnt.

Creating high-quality annotations is a difficult task due to many reasons. The most important of which that has been widely studied is **Inter Annotator Agreement** (**IAA**), also called **inter annotator consistency**, or **inter annotator discrepancy**. IAA evaluates the problem that annotators can never agree completely or exactly on what and how to annotate. It is a major indicator of the usefulness of training data to a learning method (Brants, 2000). Essentially, discrepancies and inconsistencies among annotators are primarily caused by the differences in their knowledge and experiences (Hripcsak and Wilcox, 2002). To reach a reasonable level of IAA, the traditional annotation process involves multiple domain experts working on the same annotation task in an iterative and collaborative manner to identify and resolve discrepancies progressively, to eventually produce an output that best matches the subtly varying viewpoints across a community. However, such a detailed process is often ineffective despite taking significant time and effort – typically months, and even years in rare cases (Brants, 2000; Wilbur et al., 2006). Unfortunately, discrepancies can never be eliminated and remain high in some cases (Morante et al., 2009).

The tremendous cost involved in such an often ineffective and inefficient practice means that it is difficult to introduce NER to new domains and particularly inapplicable in many practical situations such as industries due to resource limitations such as finance, time and personnel (Iria, 2009a). For this reason, a better approach to training data annotation must be sought. *Addressing this research problem will help tackle one of the major bottlenecks in developing supervised NER systems*.

1.2.2 Gazetteer Generation

In addition to the training data as an essential input, NER often benefits from additional background knowledge, which is most commonly encoded in the form of **gazetteers**. Gazetteers are reference lists used to map terms to certain categories or types. For NER, in the most commonly adopted sense, they contain reference entity names that are labelled by pre-defined categories relevant to the task. For example, a person gazetteer may be used as background knowledge to support recognising person entities. From the more general learning point of view, a gazetteer will be useful as long as it returns consistent labels even if these are not the required named entity types by the task, since the correspondence between the labels and the types can be learnt automatically (Kazama and Torisawa, 2008). A common approach is using word clusters as untyped gazetteers (Kazama and Torisawa, 2008; Saha et al., 2009). For example, the observation that 'Microsoft' and 'AT&T' are often found in the same clusters suggests that they are likely to share certain degree of similarity. With an additional piece of evidence that 'Microsoft' is a 'company', one can infer that 'AT&T' is also a company. In this case, the semantic category represented by the cluster is unknown a-priori; however, it provides additional learning evidence equivalent to a gazetteer. In this thesis, the first type of gazetteers will be referred to as '**type-oriented**' or '**typed**' gazetteers, while the second type will be referred to as '**alternative**' or '**untyped**' gazetteers.

It has been shown that gazetteers play an important role in improving the accuracy of NER systems and often lead to crucial improvement in domain specific applications (Roberts et al., 2008; Sasaki et al., 2008). Unfortunately, gazetteers are not always available and often incomplete, especially in technical domains. Building and maintaining gazetteers by hand is a laborious process and can be very expensive (Kazama and Torisawa, 2008). Therefore, effective methods are needed to support automatic generation of gazetteers – either in the type-oriented or alternative forms. *Addressing this research problem will help improve the learning accuracies of NER methods*.

1.2.3 Resolving Ambiguities

As mentioned before, due to the polysemy of natural language, names can be used for different entities, a problem that is referred to as ambiguity. For example, NER will identify 'Bush' as a *person* entity from 'President Bush attended the opening ceremony of the Olympic Games in Beijing', but is unable to recognise whether it is the '43rd president, George. W. Bush', or 'the 41st president, George H.W. Bush'. While most of the time the problem is not a concern for humans, for machines' interpretation of human language it is necessary to resolve the ambiguities and recognise the true entity that the name refers to.

Traditionally, resolving ambiguities is not considered as part of the NER process, but to be dealt with by **Named Entity Disambiguation** (**NED**), the task of resolving ambiguous name mentions to entities in a reference inventory. It is a field closely related to **Word Sense Disambiguation** (**WSD**), where meanings of ambiguous words are resolved based on their context, usually according to a sense inventory such as a dictionary that lists all possible word senses (Navigli, 2009). NED is often treated as a post-processing task for NER. However, theoretically, the two tasks serve similar goals and address 'recognition' at different levels. While NER recognises named entities in text and their semantic categories, NED recognises the true identity the name refers to. From a different perspective, NER is also a 'disambiguation' process, where the boundaries between name mentions and other text units and the semantic categories for name mentions are disambiguated (Vlachos and Gasperin, 2006). From a practical point of view, NED is often an important step before the output of NER can be used for many advanced tasks. For example, when searching for person names, search engines can benefit by disambiguating different identities and group results referring to the same person entity. When the NER output is used to populate a knowledge base of cities and countries one has to disambiguate the name 'Manchester' to the UK or US cities, or others. The need to combine NER with NED is also acknowledged by some well-known evaluation campaigns in the IE community. For example, the Knowledge Base Population (KBP) track in the Text Analysis Conference (TAC KBP Track, 2009; TAC KBP Track, 2010) has been hosting an entity linking task on an annual basis since 2009. The goal is to identify named entities from a query (NER), and link the named entity mentions to unique entities defined in a knowledge base (NED).

This thesis takes the view that both NER and NED addresses the 'recognition' task of named entities from different but complementary levels, essentially enabling the 'learning' of named entities. NER should be followed by a process of resolving name ambiguities, which essentially assigns unique identifies to the output of NER. *Addressing this problem will truly enable 'recognition' of named entities, i.e., from identification of names, to semantic category classification, to identity recognition. Ultimately this enables NER output to be used for many other applications.*

1.3 Research Hypothesis

To answer the above research questions, this thesis has studied existing work related to each question and identified their limitations. Based on these findings, several hypothesises are proposed corresponding to the three research questions outlined above.

> **H1. Training data annotation: the discrepancies among annotators, caused by the difference in their knowledge and experiences, indicate different levels of annotator's *suitability* for an annotation task. It is possible to assess such suitability and define suitability-based tasks so as to ensure annotations to be generated in a more effective and efficient way.**

As discussed in Section 1.2.1, the discrepancies or inconsistencies among annotators is a major concern in creating training data. Essentially, the majority of discrepancies among annotators are caused by the differences in their knowledge and experiences (Hripcsak and Wilcox, 2002). The traditional annotation process identifies these differences and aims to minimise them iteratively, eventually producing an output that best matches the subtly varying viewpoints across a community. This thesis takes a different point of view that these differences result in different levels of annotators' suitability for an annotation task or sub-tasks. This is inspired by the real life experiences that people typically specialise in one or several areas and no one is perfectly suited for all. Analogously in a document annotation practice, not every candidate is perfectly suitable for all tasks; however, one can be more suitable for particular tasks than others (e.g., based on the entity types in NER). Therefore, the key to improving annotation quality is not correcting the differences revealed by the repetitive checking process at the maximum effort, but rather identifying annotators' suitability and defining suitability-based annotation tasks. Annotators should be allocated to tasks they are most suitable for, thus ensuring both quality and efficiency.

This hypothesis is further discussed in Chapter 5. A document annotation exercise is conducted to reveal and study different levels of discrepancy for annotating different entity types. Based on the findings the annotator's suitability is analysed and annotators are selected for annotating particular types of entities in a further set of documents, which are then used as the final training data to a supervised NER system. Experiments are designed to evaluate the proposed methodology to further justify the validity of the hypothesis.

H2.1 Type-oriented gazetteer: Wikipedia can be used as a knowledge base of named entities. An existing gazetteer of predefined types can be automatically expanded using Wikipedia by defining gazetteer hypernyms using the structure and content of Wikipedia, and extracting similar entities that share similar hypernyms with the seed gazetteer.

This hypothesis particularly addresses methods that automatically generate type-oriented gazetteers, i.e., gazetteers of pre-defined types. Literature in this area has largely assumed that an initial 'seed' gazetteer or domain-specific extraction patterns must be available to bootstrap the automatic generation process (Riloff and Jones, 1999; Thelan and Riloff, 2002). While these methods have predominantly built gazetteers using unstructured documents, recent research has shifted the focus to exploiting collaborative knowledge sources on the Web.

Collaborative knowledge resources have gained substantial popularity in the wider research communities of NLP and IE. The most representative of these is Wikipedia[1], possibly the largest encyclopaedia ever built and maintained by collaborative efforts. It has been used to support a wide range of NLP and IE related tasks, such as NER (Toral and Munoz, 2006; Kazama and Torisawa, 2008), NED (Bunescu and Pasca, 2006; Cucerzan, 2007), and document classification (Gabrilovich and Markovitch, 2006). The majority of Wikipedia articles are descriptions of entities and concepts. The sheer size of Wikipedia and its broad coverage of various topics make it a vast knowledge base of named entities, which has great potential for building comprehensive gazetteers and updating the resources for specific NER tasks.

This hypothesis has been partially justified by a number of existing studies, including Toral and Munoz (2006) and Kazama and Torisawa (2008), which will be further discussed in Chapter 6. This chapter also introduces a domain-independent method of gazetteer expansion, which utilises various structures and contents from Wikipedia to expand existing gazetteers of pre-defined types. This method is then evaluated in a gazetteer expansion task for NER in the Archaeology domain to empirically justify the hypothesis.

H2.2 Alternative gazetteer: named entities are highly related to topic-oriented words specific to a document. The topicality of words can be evaluated based on

[1] Wikipedia, http://www.wikipedia.org/, last retrieved on 14 Mar 2012.

the relevance measures widely used for Information Retrieval. It can be used for generating alternative gazetteers for NER.

This hypothesis particularly addresses alternative gazetteers, i.e., the broader sense of gazetteers that are not explicitly typed but simply as groupings of related terms. Literature in this area has primarily taken word-clustering based approaches (Freitag, 2004; Miller et al., 2004; Jiang et al., 2006; Kazama and Torisawa, 2008; Saha et al., 2009; Finkel and Manning, 2009; Chrupała and Klakow, 2010), while some (Kazama and Torisawa, 2007a; Kazama and Torisawa, 2008) have used automatically extracted hypernyms that group semantically similar concepts or instances.

The link between the topicality of words and named entities was initially discussed in Clifton et al. (1999) and Hassel (2003). Few studies have exploited this feature in NER related research (Rennie and Jaakkola, 2005; Gupta and Bhattacharyya, 2010). These have suggested that topicality of words can be quantified by the property of **informativeness**. Although a formal definition is lacking, it is generally agreed that informative words are those that often demonstrate a ‘peaked’ frequency distribution over a collection of documents, such that the majority of their occurrences are found in only a handful of documents in the collection (Church and Gale, 1995b). Informativeness measures are typically based on *global* word distributional characteristics observed in the entire corpus (e.g., document frequency, word frequency in the corpus), while ignoring the distinctive distributional patterns of words within individual document (*local*) contexts (e.g., frequency within documents). However, in practice, topics can vary by documents even if they belong to the same domain. This may translate to different distributional characteristics of a word observed at individual document basis. Global informativeness scores can mis-represent the strength of topicality of words in different document contexts and harm learning accuracy.

Instead, this thesis hypothesizes that topic-oriented words should be defined specifically to document context and they can be useful indicators of named entities in the same document context. Following this hypothesis, within a specific document context, words can be grouped based on their level of topicality and the intuition is that those falling under the highly topic-oriented groups can be useful features to NER. Essentially this has inspired the creation of document-level, alternative gazetteers. To extract topic-oriented words for each document, this thesis proposes to use relevance measures widely used in the Information Retrieval tasks as a proxy for topicality.

This hypothesis is discussed in Chapter 7. The link between topicality of words and named entities is partly justified by the literature review, which also discusses a number of methods based on the similar ground. The hypothesis leads to a novel approach which, when submitted to a comprehensive comparative evaluation, has shown to be effective and generalisable across domains.

> **H3.1 Resolving ambiguities: an ambiguous entity name can be resolved based on the semantic relatedness between its referent entities and other named entities it co-occurs with in its context, because contextually co-occurring named entities are semantically related.**

Given a coherent text discourse that contains multiple (ambiguous) names of entities, the true referent entities of each name are usually semantically related. For instance, in the previous example ‘President Bush attended the opening ceremony of the Olympic Games in Beijing’, to a human reader it is clear that ‘President Bush’ refers to the 43rd US President George W. Bush and ‘Olympic Games’ refers to the 2008 summer Olympic Games held in Beijing. The underlying logic is that these are the only solutions to maximise the semantic connections among the three names ‘Bush’, ‘Olympic Games’ and ‘Beijing’ in a single discourse.

This hypothesis is inspired by Cucerzan (2007), who argued that ambiguous entity names can be resolved by maximising the agreement among the data held for candidate entities in the same discourse. This thesis argues that this agreement can be measured by lexical semantic relatedness. Then disambiguation can be achieved based on the idea of ‘agreement (as determined by relatedness) maximisation’. This is discussed in Chapter 9 and justified by a method of NED based on a lexical semantic relatedness measure.

The key to manifest this hypothesis is capturing the **semantic relatedness** between entities, a task that can be achieved by lexical semantic relatedness methods that determine the semantic association strength between terms or concepts based on certain background information about them. Although literature on lexical semantic relatedness is particularly abundant, a thorough review reveals that existing methods typically employ background information of terms and entities from a single resource. However, different background information resources may contain information about the same entities or concepts, while having different focuses in terms of the type and amount of knowledge encoded. This suggests a complementary nature of different background information resources. This motivated a study that leads to a novel lexical semantic relatedness measure based on the following hypothesis:

> **H3.2 Lexical semantic relatedness: lexical semantic relatedness measures can benefit from combining different background information resources since they complement each other in certain ways.**

This argument is discussed in Chapter 8. The literature is thoroughly reviewed and compared including an analysis of the characteristics of different background information resources widely used in this task. A novel method is proposed based on the principle of combining knowledge of terms and concepts from different resources. This method is then evaluated in both the general and technical domains, which further justifies the hypothesis it builds on. It is then adapted to the Named Entity Disambiguation task that is further discussed and evaluated in Chapter 9.

1.4 Contributions

This thesis presents a body of work exploring methodologies and techniques to enable effective learning of named entities. The main contributions of this thesis are distinct techniques each addressing an essential task in the 'recognition' of named entities.

1.4.1 Training Data Annotation

An effective and efficient approach to manual document annotation

Creating training data is an essential process but also the bottleneck in supervised NER. This thesis studies the standard methodology for document annotation and analyses its limitations. An alternative approach is introduced based on the hypothesis of annotator's suitability in a task. The approach firstly studies Inter Annotator Agreement for the annotation of each entity type based on a sample of the domain corpus following the standard annotation practice. Next, a set of experiments are carried out to evaluate machine learning accuracy using these annotations. The results together with the IAA studies are used to evaluate annotators' suitability for annotating each type of named entities. Lastly, to create the final training data for supervised NER, each annotator is only required to annotate the documents for the entity types they are most suitable for, and the work load is equally distributed among all annotators. Experiments show that this approach leads to reduced overall annotation time and improved annotation quality. Details of this are presented in Chapter 5, which addresses the hypothesis H1.

1.4.2 Gazetteer Generation

A method of automatically expanding type-oriented gazetteers for NER using Wikipedia

Due to the evolutionary nature of human knowledge, existing gazetteers often need to be updated and expanded in order to be adapted to related domains or simply to be up-to-date. Such a task, if done manually, can cause significant cost. Therefore, the ability to automatically update and expand gazetteers is also an important feature to an NER system. This thesis introduces a novel approach to automatically expanding existing gazetteers using knowledge in Wikipedia. Unlike previous work, the method exploits various kinds of content and structural elements of Wikipedia, and does not rely on domain-specific knowledge. Briefly, given an existing seed gazetteer containing named entities that are described by Wikipedia articles, it firstly extracts hypernyms of the entities in the initial gazetteer using their Wikipedia article contents and structures. Next, related entities are identified as the links on these articles. If a related entity shares the hypernyms of the entities in the seed gazetteer, they are added to the expanded set. The method is empirically tested in the Archaeology domain, where three existing gazetteers are automatically expanded following the proposed method. The resultant gazetteers are then used in an NER task, where the results have shown that they have contributed to further improvement in NER learning accuracy. Details of this are presented in Chapter 6, which answers the hypothesis H2.1.

A method of automatically generating alternative gazetteers for NER by exploiting the association between word topicality and named entities

Based on the hypothetical association between named entities and topic-oriented words within specific document context as outlined in H2.2, this thesis proposes to measure word topicality with respect to specific document contexts by the relevance measures widely used for Information Retrieval tasks, and transfers the scores to useful features for learning NER. Briefly, for each unique word in a document the method firstly computes a topicality score using a relevance measure, such as tf.idf (Spark Jones, 1973). It then hypothesizes that highly topic oriented words are indicative of named entities. They are rare but can be used by many named entities (including multiple occurrences) in the document and as the scores drop, their usefulness drop disproportionately faster. This creates a non-linear distribution of topic-oriented words over named entities. To capture this nature and also to normalise document-specific topicality to a uniform scale such that they are comparable across documents, the words are ranked by the scores and a simple equal interval binning technique is applied to segment the list into a handful of sections. Effectively, this is equivalent to creating a handful of untyped gazetteers, which are then used for a statistical NER model. In addition, other methods of exploiting word informativeness in NER are also studied and compared. Details of these are presented in Chapter 7.

1.4.3 Lexical Semantic Relatedness

As discussed before, methods of lexical semantic relatedness are the enabling technique for the proposed disambiguation approach. To gain sufficient understanding of the field, a thorough review of the literature on lexical semantic relatedness has been carried out, which further led to a novel approach. Two contributions are made in this domain:

A comprehensive review of lexical semantic relatedness methods covering multiple domains and resources, with an objective to connect different methods in terms of their rationale, and contrast different methods in terms of their advantages and disadvantages

A careful study of the literature shows that there is a need for an up-to-date comprehensive review of state-of-the-art. It has been noted that, a great number of methods has been introduced in the last few decades in different domains, and based on different background information resources. Efforts on summarising these studies are rare, and are limited in scope since they generally target on specific areas (e.g., domains, rationales, resources). Work across such area boundaries is insufficiently communicated, and it has been noted by this study that near-identical methods have been introduced in different contexts, costing expensive research effort.

Therefore, one contribution of this thesis is to present a comprehensive literature review that addresses these limitations. Different methods are discussed from a generic perspective and their rationales and connections are analysed. Conclusive remarks are also drawn regarding the research and application of lexical semantic relatedness. It is believed that this will be a valuable reference for researchers and practitioners of lexical semantic relatedness. This part of work is presented in Section 8.2 of Chapter 8.

A lexical semantic relatedness measure that harnesses different knowledge sources under a uniform framework

Following the literature review, a novel lexical semantic relatedness measure is introduced in Chapter 8. This in particular, addresses hypothesis H3.2. As opposed to the majority of existing work that are based on a single source of background knowledge, the method harnesses knowledge from three resources in computing lexical semantic relatedness: Wikipedia, WordNet (Fellbaum, 1998) and Wiktionary[2]. Firstly, given a polysemous term and its corresponding entries found in each of the three resources, an entry in Wikipedia is mapped with the closest entry from WordNet and Wiktionary that are likely to refer to the same meaning using a simple feature overlap based method. Next, different kinds of features (lexical and semantic content) are extracted from each resource, and features of similar types across different resources are mapped. Based on the cross-mapped entries and features, a joint feature vector representation is created for each mapped entry. The semantic relatedness between two polysemous terms is then computed based on the joint feature vectors of their underlying sense entries. Compared to the previous work, the proposed method combines knowledge from different resources and improves the accuracy of measuring semantic relatedness in both general and specific domains.

1.4.4 Resolving Ambiguities

A method of NED based on lexical semantic relatedness measure

The lexical semantic relatedness measure introduced in Chapter 8 is then adapted for resolving ambiguous entity names based on the hypothesis of H3.1 ‘agreement maximisation’. To do so, entity names from a single discourse are firstly extracted to form the context to each other. Next, candidate referent entities for each name are identified from Wikipedia. Then, the lexical semantic relatedness measure proposed before is adapted to compute pairwise relatedness between the candidate entities to derive a semantic relatedness matrix. The final step is choosing a single referent entity for each entity name, the process of which aims to maximise the agreement in terms of the semantic relatedness

[2] Wiktionary, http://www.wiktionary.org/, last retrieved on 14 Mar 2012

scores among all entity names. Several techniques are introduced and experimented for this purpose, which is presented in Chapter 9. The proposed method largely outperforms a baseline model and outperformed the best method in the literature on the larger dataset.

1.5 Thesis Structure

The remainder of this thesis is divided into five parts and organised as follows.

1.5.1 Part I. Background

Chapter 2 presents an overview of NER required for the understanding of the subsequent parts of this thesis. To be consistent with the literature, the discussion focuses on the traditional sense of named entity ‘recognition’, which will be formally defined. The methods for NER and evaluation approaches are briefly introduced. Literature concerning specific research questions will be discussed in details in the subsequent parts of this thesis.

Chapter 3 details the three research questions related to NER outlined above, i.e., training data annotation and gazetteer generation for NER, and resolving ambiguities – where the need for sense disambiguation for NER (Named Entity Disambiguation) is discussed and the view of a complementary nature between NED and NER is introduced.

Chapter 4 presents a supervised learning model for NER. This is a uniform model that makes the fundamental NER system used in the experiments of later chapters.

1.5.2 Part II. Training Data Annotation

Chapter 5 presents the proposed method for training data annotation. It begins with a literature review of the standard practices for document annotation, where the limitations of the standard approaches are analysed. It then presents a case study of a real document annotation exercise conducted in the archaeology domain, in which the details of the proposed annotation method are discussed. This process generates a set of annotations for the archaeology domain, which are also used later in this thesis. The proposed annotation method is then evaluated both for efficiency and effectiveness.

1.5.3 Part III. Gazetteer Generation

Chapter 6 presents the proposed method for expanding typed gazetteers using Wikipedia. It firstly discusses existing studies on automatic generation of typed gazetteers, which usually start with certain seed data. Next, the novel method of expanding gazetteers using Wikipedia is introduced. It is then tested in the archaeology domain, an example of a domain specific application that is rarely addressed in the literature. Three existing gazetteers are expanded using the proposed method, and the expanded gazetteers are evaluated in an NER task.

Chapter 7 presents the proposed method for generating alternative gazetteers based on word topicality. It begins with a review of related work on generating alternative gazetteers. A particular focus will be placed on studies based on the similar ground, against which the proposed method is compared. Next the method is discussed in details, followed by a comprehensive evaluation using several datasets from different domains and comparing against several other methods based on similar hypotheses of word topicality. An in-depth analysis follows to uncover the link between topic-oriented words and named entities, and discusses how it can be used properly to support NER.

1.5.4 Part IV. Resolving Ambiguities

Chapter 8 presents the study of lexical semantic relatedness methods, which lays the foundation for the study on Named Entity Disambiguation in Chapter 9. It begins with a comprehensive review of the state-of-the-art aimed at bridging the gap identified in the existing surveys in this field. A novel method is then proposed to measure lexical semantic relatedness based on the combination of multiple knowledge sources. The method is thoroughly evaluated on both general and specific domain datasets. It comprises the main component for the NED method to be discussed in Chapter 9.

Chapter 9 introduces a method to NED based on the hypothesis of ‘agreement maximisation’, which is assessed using the lexical semantic relatedness measure introduced in Chapter 8. The literature on NED is firstly presented, followed by a discussion of the hypothesis and details of the proposed method of NED. The method is then evaluated on standard benchmarking datasets and compared against state-of-the-art.

1.5.5 Part V. Conclusion

Chapter 10 concludes this thesis and discusses how the work explored in the previous chapters has contributed to proving the hypotheses outlined in Chapter 1. It also discusses how work carried out in this thesis can be extended in the future.

Part I - Background

This part presents the background knowledge that is essential to the understanding of this thesis. Chapter 2 introduces NER in general; Chapter 3 details the three research questions related to NER to be addressed by this thesis; Chapter 4 presents a uniform model of NER that lays a common ground for the individual studies in later chapters.

2 Background of NER

PREFACE

This chapter introduces the Named Entity Recognition task from a general point of view, focusing on basic concepts and principles that are required for the understanding of the subsequent parts of this thesis. Section 1 describes the NER task in detail with supporting examples. Section 2 presents a brief summary of the applications of NER to illustrate its important role to other related research and application areas. Section 3 outlines methods and techniques commonly used for NER. Section 4 describes the evaluation methodologies. Section 5 summarises this chapter.

2.1 Defining Named Entity Recognition

The task of Named Entity Recognition was formally defined in MUC6 as the task of ‘identifying the names of all the people, organisations and geographic locations in a text’, as well as ‘time, currency and percentage expressions’ (Grishman and Sundheim, 1996). An example of such is shown in Figure 2.1, in which names of entities are annotated using mark-up tags. ‘ENAMEX’ and ‘NUMEX’ are both tags introduced in MUC6, where the former stands for ‘entity name expression’ and the latter stands for ‘numeric expression’.

```
Mr. <ENAMEX:TYPE=‘PERSON’>Dooner</ENAMEX> met with
<ENAMEX: TYPE= ‘PERSON’>Martin Puris</ENAMEX>, presi-
dent and chief executive officer of <ENAM-
EX:TYPE=‘ORGANIZATION’>Ammirati & Puris</ENAMEX>,
about <ENAMEX: TYPE=‘ORGANIZATION’>McCann</ENAMEX>’s
acquiring the agency with billings of <NUMEX:
TYPE=‘MONEY’>$400 million</NUMEX>, but nothing has ma-
terialised.
```

Figure 2.1. Example of named entities in the MUC6 dataset

Since MUC6 there has been increasing interest in this topic and extensive effort has been devoted into its research. Major computational linguistic conferences hosted special tracks for the task and there has been steady growth of publications throughout the years. Several events made the attempt to enrich the definition of the task. For example, MUC7 (Chinchor, 1998) included date and time entities, and introduced the multi-lingual named entity recognition. The Automatic Content Extraction (ACE) program introduced several new entity types and a more fine-grained structure of entity sub-types in an attempt to achieve more precise classification of entities, such as distinguishing government, educational and commercial organisations from each other, which all belong to the coarse-grained entity type ‘organisation’ (Doddington et al., 2004).

The task has also been extended to technical domains to recognise domain-specific entities, typically in the domain of biomedical science to recognise domain-specific entities such as gene and protein names. Large amount of resources have been created for the purpose of evaluating biomedical entity recognition such as the Genia corpus (Ohta et al., 2002), and successive events have been hosted to motivate the research such as the BioNLP/JNLPBA shared task on entity recognition (Kim et al., 2004). Figure 2.2 illustrates an example sentence from the Genia corpus annotated by domain-specific entity types (as defined by ‘<cons sem=>‘).

In <cons sem='G#cell_type'>**primary T lymphocytes**</cons> we show that <cons sem='G#protein_molecule'>**CD28**</cons> ligation leads to the rapid intracellular formation of <cons sem='G#inorganic'>**reactive oxygen intermediates**</cons> (<cons sem='G#inorganic'>**ROIs**</cons>) which are required for <cons sem='G#other_name'><cons sem='G#protein_molecule'>**CD28**</cons>**-mediated activation**</cons> of the <cons sem='G#protein_ molecule'>**NF-kappa B**</cons>/<cons sem='G#protein_complex'> <cons sem='G#protein_molecule'>**CD28**</cons>-responsive complex </cons> and <cons sem='G#other_name'><cons sem='G#protein_molecule'>**IL-2**</cons> **expression**</cons>.

Figure 2.2. Example of domain specific named entities in the Genia corpus

To generalise, *NER is the task of identifying the* **mentions** (or **names**) *of* **entities** *in the text and assign* **semantic categories** *to them.*

- **Entities** – refer to real world objects that are individually distinctive and identifiable by unique identifiers, such as 'George Walker Bush, born July 6, 1946, an American politician who served as the 43rd President of the United States, from 2001 to 2009';
- **Mentions/names** – these are lexical realisations of entities, such as 'George Bush', 'Mr. President', and 'President Bush' that can be used to refer to the same entity above. Other terms such as 'proper names' and 'surface forms' are used interchangeably for the same purpose;
- **Semantic categories** – these are semantic classes used to label same kinds of entities, such as 'person' and 'location'. They are also referred to as **types**, **classes**, or **labels**.

The NER task naturally translates into two sub-tasks: **name detection** or **identification** (the **bold** text in Figure 2.1 and Figure 2.2) that finds the boundaries of entity names; and **semantic classification** (the tags in '< >') that assigns the most appropriate semantic category. Due to the polysemy of human language, a name can be *ambiguous* since it may refer to multiple entities. For example, in Figure 2.1, 'Dooner' may refer to any person with the same surname. The process of resolving these ambiguities can be considered as a process of *recognising the unique identities – or the true referent entities –* that each name refers to, which enables truly 'recognition' of named entities. This is crucial to the ultimate understanding of the text. However, traditionally it is not the goal of NER, but rather to be dealt with by the task of sense **disambiguation**, or Named Entity Disambiguation in this context.

2.2 Applications

NER is an enabling technology to many applications. It is often used in a pre-processing step to many complex IE and IR tasks. This section briefly summaries some of these tasks.

Relation Extraction – Relation Extraction is the task of recognising semantic relations expressed between entities and concepts (Giuliano et al., 2006; Giuliano et al., 2007). Examples of relations include *Person-Affiliation* (Larry Page, Google Inc.), *Located-In* (University of Sheffield, Sheffield), *Born-In* (Albert Einstein, Ulm) etc. Since relations are often found between entities and concepts, recognising named entities and/or concepts is often the essential first step.

Event Extraction – Event Extraction involves detecting multiple entities and relations between them often according to a pre-defined template. For example *seminars* are usually made up of several parts: *speaker*, *topic*, *location*, *start time* and *end time*. Extracting events requires the ability to recognise named entities that form integral parts of the event.

Knowledge base generation and population – A knowledge base refers to a resource of certain types of knowledge units – usually entities and concepts – organised structurally by certain types (e.g., semantic) of relations. Examples of frequently used knowledge bases include taxonomies, ontologies, thesauri, etc. Knowledge bases are often used for automated reasoning, an important capability for enabling the Semantic Web (Berners-Lee et al., 2001). Building knowledge bases involves extracting concepts and entities from texts and learning semantic relations between them, and therefore, requires support from NER and Relation Extraction. Additionally, the process typically requires disambiguation (Dredze et al., 2010) to resolve ambiguities and integrate information.

Question Answering (QA) – Question answering is the task of automatically finding answers to a question expressed in natural language. A core component in many QA systems is NER, which is used to recognise named entities in both the questions and potential answer texts. It is found that often a very large proportion of questions are formed around named entities (Nadeau, 2007a), and entity names are useful for locating supporting information and facts (Lee et al., 2007; Srihari and Peterson, 2008) in texts. QA systems can often benefit from a sense disambiguation processor, which can help better understand the question as well as locating accurate answers (Huang et al., 2005).

Semantic Search – as opposed to the traditional free text search that returns a list of documents matching a query expressed as a set of keywords, semantic search aims to better understand users' intentions and find the information and knowledge that directly answers the query. For example a keyword based search for 'object oriented programming languages' may return a list of documents containing either some or all of the keywords; semantic search may return a list of instances such as 'Java', 'C#', 'Python' etc. Similar to QA, enabling semantic search usually requires recognition of named entities and concepts from documents. For example, Pasca (2004) cited two variants of semantic search: one returns a list of entities of a semantic category; and the other returns a list of siblings of an entity. In both cases, NEs must be identified in the text to support the task.

2.3 Techniques for NER

Techniques for NER are most often divided into two main streams: **handcrafted rules** and **learning based approaches** (Sarawagi, 2007).

2.3.1 Rule-based Approaches

Methods based on handcrafted rules involve designing and implementing lexical-syntactic extraction patterns and using existing information lists such as dictionaries that can frequently identify candidate named entities. An example of such rules can be '*a street name is a multi-word phrase ends with the word 'X' and proceeded by the preposition word 'Y'*", where 'X' and 'Y' are lists of common words that are suitable for this purpose. For example, X could be 'Street' and Y could be 'in', thus the rule can recognise names of streets from texts such as 'The Apple store in *Oxford* *Street* in London'.

Some well-known rule-based systems include FASTUS (Appelt et al., 1995), which essentially employs carefully handcrafted regular expressions to extract names of entities; LaSIE (Kaufmann et al., 1995) and LaSIE II (Humphreys et al., 1998), which made use of an extensive amount of lookup lists of reference entity names and grammar rules such as indicative words to identify candidate entities. Early entity recognition systems primarily adopted rule-based approaches, as noted by (Nadeau, 2007a). They are efficient for domains where there is certain formalism in the construction of terminology. A typical example is the biology domain, where certain types of entities can be extracted by domain-specific rules with sufficient accuracy. Relevant work includes (Seki and Mostafa, 2003; Lin et al., 2004; Roberts et al., 2008). Also it has been successfully applied in open information extraction (Cafarella et al., 2005), where information redundancy is available for relatively simple types of entities.

However, the major limitation of these systems is that they require significant expertise from the human developers, in terms of the knowledge about the language, domain as well as programming skills (Sarawagi, 2007). These knowledge and resources are often expensive to build and maintain and are not transferrable across domains. Consequently these approaches suffer from limited or no portability. As a result, the focus of research has shifted towards more robust learning based approaches since they have been introduced.

2.3.2 Learning-based Approaches

Machine learning is a way to automatically learn to recognise complex patterns or sequence labelling algorithms and make intelligent decisions based on data. Central to the machine learning paradigm is the idea of providing positive and negative **training examples** for the task; modelling distinctive **features** associated with examples; and design **algorithms** that consume these features to automatically distinguish positive from negative examples and to recognise similar information from unseen data.

Training examples or training data are usually an essential input to learning based methods. They often take the form of **annotations** that are labelled instances of named entities, created by domain experts in a **document annotation** process. For example, the annotated entities (text in bold) in Figure 2.1 and Figure 2.2 can be used as training data for building an extraction model of relevant entity types (defined by the tags). In machine learning, such annotated data are often called **labelled data**, which are often used to train an extraction model; on the other hand, the data without annotations are called **test data**. In many unsupervised learning methods (Section 2.3.2.3) that do not require annotations, a set of ‘seed data’ is often needed to support the learning. Seed data are typically lists of example entities of a particular type. Essentially they can be considered as training data in a rather different form.

Features are characteristics of text objects to be studied in a computational linguistic problem. In NER, the target text objects are tokens (e.g., words) or sequences of tokens (e.g., phrases, n-grams) for identification and classification. Features are used to create a multi-dimensional representation of the text objects, which can then be used by learning algorithms for generalisation in order to derive patterns that can extract similar data and distinguish positive from negative examples.

A wide range of features have been introduced for the NER task. Details of these can be found in Nadeau (2007a). The author describes features in three categories: **word-level features, list look-up features,** and **document and corpus features**. Examples of word-level features include word case, morphology, stem, lemma, part-of-speech, and word patterns. List look-up features are usually based on gazetteer, lexicon or dictionary, which can contain a list of reference terms that are likely to be (part of) an entity of interest. Such features are known to be very effective in some NER tasks (Roberts et al., 2008; Sasaki et al., 2008). Document and corpus features are defined by both document content and structure. Examples include co-occurrences with other words or entities and position in the document, particularly in structural elements such as titles, lists and tables.

The effectiveness of features is often dependent on several factors, such as language, domain, qualitative and quantitative characteristics of training data. As a result, the choice of features is usually task-specific, and feature selection can often lead to different performance of NER systems.

Learning algorithms are methods able to consume features of training data to automatically induce patterns for recognising similar information from unseen data. Learning algorithms can be generally classified into three types: **supervised learning, semi-supervised learning** and **unsupervised learning**. Supervised learning utilises only the labelled data to generate a model. Semi-supervised learning aims to combine both the labelled data as well as useful evidence from the unlabelled data in learning. Unsupervised learning is designed to be able to learn without or with very few labelled data. These are discussed separately in the following sections.

2.3.2.1 Supervised learning

In a supervised learning setting an NER system takes training data and their features as input to induce an extraction model, which is then used to recognise similar objects in new data. Supervised learning has been the most frequently used and still the dominant approach in the NER community (Nadeau, 2007a). There are several extensively used machine learning techniques for this task. Support Vector Machines (SVM) builds a model that draws a hyperplane that best separates positive and negative examples in the labelled data. The model represents the examples as points in space, mapped so that the positive and negative examples are divided by a clear gap that is as wide as possible. At application time, new instances are mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. It is used by, for example, Isozaki and Kazawa (2002) and Ekbal and Bandyopadhyay (2010). Hidden Markov Model (HMM) is a statistical Markov model in which the sequence of states is not directly visible (hidden) but can be predicted from a sequence of observations formulated as a probabilistic function of the states. In the context of NER, the learning process infers an HMM based on the observed features of the sequences of tokens and the associated visible states (i.e., tags) in the training data. The model creates a mapping such that a sequence of observations can predict a sequence of states with certain probability. At application time, the observations based on the new data are generated. The inferred model is applied to the observations to calculate the most likely sequence of hidden states. Example studies of NER using HMM include Zhou and Su (2004) and Ponomareva et al. (2007). Conditional Random Fields (CRF) is also a probabilistic model that is similar to HMM but relaxes certain assumptions about the input and output sequence distributions of HMM. Kazama and Torisawa (2007a) and Arnold et al. (2008) employed a CRF-based entity recogniser. Other widely used machine learning techniques such as the Perceptron algorithms (Kazama and Torisawa, 2007b), Naïve Bayes (Mohit and Hwa, 2005) Expectation Maximisation (Pandian et al., 2007), Decision Trees (Finkel and Manning, 2009), and Maximum Entropy model (Chieu and Ng, 2003) have also been applied to NER.

It has been shown that some machine learning techniques can outperform others on certain data (Farkas et al., 2006; Krishnarao et al., 2009). However, exploration of various techniques for the NER task is not the focus of this thesis. This has been partly discussed in (Olsson, 2008).

As mentioned before, compared to other methods, the major limitation of supervised learning methods is its dependence on large amount of training data, which has to be created by the manual document annotation process. This is usually a difficult task that can require substantial investment in terms of both finance and personnel.

2.3.2.2 Semi-supervised Learning

Compared to supervised learning, the major difference in semi-supervised learning is that it makes use of both labelled data and unlabelled data. Many researchers have found that it is possible to combine largely available unlabelled data with small amount of labelled data in the learning process to reduce the system's dependence on training data, yet achieving competitive learning accuracy. A popular form of semi-supervised learning in NER is bootstrapping or self-training, in which a system firstly trained on an initial small set of examples are used to tag unlabelled data. The resulting annotations are then selected to augment the initial training dataset, which is then used to re-train the system. The process repeats for several iterations to progressively refine the learning decisions and annotate the documents. This type of method has gained significant popularity and a large amount of semi-supervised NER methods are based on **bootstrapping** approaches, such as (Vlachos and Gasperin, 2006; Olsson, 2008; Knopp, 2011).

One highly influential work of this type is Riloff and Jones (1999). The method starts with a handful of seed entity names of given types and an unlabelled corpus. The seed entity names are located in the corpus and their contexts are pruned to generalise extraction patterns. The patterns are then ranked based on a confidence score, and the top ranked patterns are selected to be used to discover new examples. The process is repeated in an iterative manner and eventually the corpus is annotated automatically. Thelan and Riloff (2002) extended this idea by incorporating collective evidence from a large set of extraction patterns, which proved to be more effective than the earlier approach. Liao and Veeramachaneni (2009) build a semi-supervised NER system that starts with training a supervised learner using a small amount of labelled data. The trained learner is then applied to unlabelled data to generate new annotations of entities. The newly annotated corpus is merged with previously labelled data to form a new training corpus, which is then used to train a new classifier. The process is repeated in several iterations, ensuring that in each turn, only the accurately tagged (measured by confidence) non-redundant examples are added to the pool of labelled examples to form new training data for the next iteration.

The major limitations of this class of approaches are 'error propagation', that the performance rapidly declines as noisy patterns or entities are introduced in the bootstrapping process (Riloff and Jones, 1999; Ando, 2004). Also, low frequency classes of entities can be problematic since there may be insufficient contextual information for pattern generalisation.

Another semi-supervised approach is **co-training** (Blum and Mitchell, 1998). In co-training, two learning models are trained using the same training data, but each with a disjoint set of features and sometimes with different machine learning algorithms. Each model creates a different view of the data and outputs from each model are aggregated. One of the earlier studies of this branch is Collins and Singer (1999). The authors proposed to build two separate classifiers, one employs the 'spelling' rules of words and the other utilises the 'contextual' rules. Examples of spelling rules can be a look-up for the exact string, its prefix and suffix; while contextual rules consider words surrounding the string in the sentence it appears in. Learning begins by firstly labelling the data with a small set of spelling rules. The annotations are then used to infer contextual rules, which are then scored and selected to re-annotate the same data. This generates new annotations, from which new spelling rules can be derived. This process repeats iteratively until an arbitrary number of rules are reached. Niu et al. (2003) firstly label a corpus with concept-based seeds, such as 'he', 'she', 'man' and 'woman' for the Person class. The motivation is that concept-based seeds share the same grammatical structures as their corresponding instance entities and they occur more frequently in a corpus. Then a decision tree based approach is applied to learn the parsing-based rules from this labelled corpus. Finally, the inferred model is applied to an unlabelled corpus, using which an HMM NER classifier is trained. Other examples of co-training based NER studies include Steven (2002), Chung et al. (2003), Kozareva et al. (2005) and Ma (2009).

Similar to the bootstrapping approach, co-training generally depends on information redundancy (Collins and Singer, 1999), which can make the approach ineffective to low-frequency named entity classes. Also, errors in the annotations created by one classifier may be propagated when the annotations are used for training the other.

2.3.2.3 Unsupervised Learning

Unsupervised learning methods make decisions based on unlabelled data. In NER, most unsupervised learning methods make use of clustering techniques, distribution statistics and similarity based functions.

Evans (2003) studied the problem of NER in the open domain, which is concerned with recognition of any types of entities that may be useful to IE. The method firstly extracts sequences of capitalised words that are likely to be entity names, and then composes search queries using these word sequences together with Hearst patterns (Hearst, 1992). For example, if a capitalised word sequence is '*Microsoft Inc.*', the phrase '*? such as Microsoft Inc.*' is created as a search query. The query is then sent to search engines to retrieve a list of documents, which are further processed to find the hypernyms of the word sequences. These are simply the word or phrase filling the position of '?' in the returned document snippets. The extracted hypernyms are then clustered, looked up in WordNet and labelled by top level concepts in WordNet.

Da Silva et al. (2004) hypothesized that named entities are often lexicalised as Multi-Word Units (MWUs), the components of which occur more often together than separately. They proposed to use mutual information measures and the frequency of words to identify n-grams (where n>1) from corpus that are potential entities. Next, they used a clustering algorithm to group similar named entities together. Later Downey et al. (2007) extended this idea and applied similar method using the Web data. However, these methods do not attempt to classify named entities to pre-defined categories.

Cimiano and Völker (2005) used a vector similarity based model which labels candidate entity names based on its similarity with candidate types. Essentially, candidate entity names and types or classes are modelled as vectors based on certain features. A candidate name string is assigned the type whose feature vector is most similar to its own. Kliegr et al. (2008) followed a similar approach that they call Semantic Concept Mapping. Given a list of candidate entity names and a pool of labels, both names and labels are looked up in WordNet and represented as WordNet synsets. Next, the matching type for an entity name is the one that maximises the similarity between two WordNet synsets using Lin's similarity function (Lin, 1998b).

2.4 Evaluation of NER

Evaluation of NER systems is typically based on the comparison of the output of an NER system with that of human annotators on the same dataset. In this case, the output of an NER system is often called '**predictions'** and the human annotations are called '**gold standard**'. The standard measures for evaluating the comparison are **precision, recall** and **F-measure**.

2.4.1 Precision, Recall and F-measure

The calculations of precision and recall are based on the numbers of **true positives, false positives,** and **false negatives**. Given a list of entity annotations of a particular type predicted by an NER system and the gold standard annotations for the same dataset, true positives are the instances correctly labelled according to the gold standard, false positives are the instances incorrectly labelled, and false negatives are the instances that should be labelled but were missed by the system. Using these numbers, precision and recall are calculated using the formulas as below:

$$Precision = \frac{|True\ Positives|}{|True\ Positives| + |False\ Positives|}$$ **Equation 2.1**

$$Recall = \frac{|True\ Positives|}{|True\ Positives| + |False\ Negatives|}$$ **Equation 2.2**

In simple words, precision measures the ability of an NER system in predicting named entities correctly, whereas recall measures the ability of the system in discovering named entities from text completely. Depending on the purpose of an NER task, it is often desirable to trade off certain precision to obtain higher recall or vice versa (Minkov et al., 2005). However, in most cases, one may want to balance both factors in the evaluation. The standard approach is using the F-measure, which is a harmonic mean of precision and recall, calculated as below:

$$F_{\beta} = \frac{(1+\beta^2)\ \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$$ **Equation 2.3**

where β is the relative importance of precision versus recall. The most often used form of F-measure adopts β*=1,* and is therefore, usually referred to the **F1 measure**.

2.4.2 Matching Predictions against Gold Standard

As described in Section 2.1, NER contains two sub-tasks: entity name detection that identifies the boundaries of a candidate named entity, and semantic classification that assigns semantic category to it. For this reason, qualifying true positives also involves satisfying two criteria: finding the correct boundaries and assigning the correct label.

Several early research conferences and workshops have proposed varying standards for matching true positives. The simplest approach is '**exact match**', under which a predicted named entity mention qualifies a true positive if and only if both the boundaries and the label are exactly the same as in the gold standard. Therefore this is also the strictest matching method, which has been used in the shared NER task in CoNLL2003 (Sang and Meulder, 2003).

The MUC events (Grishman and Sundheim, 1996; Chinchor, 1998) defined a more relaxed scheme, which rewards systems that either predicts correct labels, regardless of whether the boundaries are correct but as long as there is a text overlap; or systems that predict correct boundaries even if the label assigned is incorrect. The overall performance measure takes into account of both types of matches.

ACE (Doddington et al., 2004) defined the most complex form of evaluation in an attempt to incorporate issues such as partial matches, wrong type, and the newly proposed 'subtype' and 'class' scheme. Each NE type is assigned a weight parameter and contributes up to a maximum proportion of the final score. The ACE standard can be problematic due to its sophisticated nature that can complicate the error analysis.

Freitag (1998) and De Sitter and Daelemans (2003) raised the question of what is really needed by an NER task when counting true positives. If the extracted data were to be used for populating a database, high accuracy is necessary and thus exact match is needed. On the other hand, if the purpose was to help a human locate useful information in a document, it may be sufficient to just have a system that predicts overlaps with desired information. For this reason, they proposed and used three different ways of matching: the exact match as introduced above; containment, as whether a predicted instance contains an actual instance, allowing a maximum of *n* neighbour tokens; and overlapping, as whether a predicted instance overlaps with an actual instance, allowing a maximum of *n* neighbour tokens and *m* missing tokens. Similarly, Tsai et al. (2006) suggested several practical matching schemes for evaluating NER in biomedicine that relax the penalties for boundary mis-matching: matching only the left or right boundary, approximate match, core-term match and so on.

2.4.3 Macro- and Micro-averaged F-measure

Since an NER task often involves multiple types of entities, it is often required to obtain an assessment of the overall performance of the system for all named entity types. This is often done in two ways: ***macro*-averaged F-measure** and ***micro*-averaged F-measure**. Macro-average F-measure is the mean of F-measures of all the entity types in the corpus. Micro-average F-measure is obtained by adding together the labelled instances of all entity types and then computing precision, recall, and F-measure. The difference is that the micro-averaged measure can be dominated by the larger classes in the corpus such that the performance of the system on smaller classes is counted much less. However, micro-averaged measure is more frequently used in commonly used evaluation tools such as the MUC scorer (Chinchor, 1998).

2.4.4 Cross Validation

Cross validation is a technique used for balanced evaluation of a system. It is a common technique used to evaluate supervised learning methods. The core idea is to partition the labelled data into complementary (usually equal) *k* subsets, usually performing training on *k – 1* subsets (training data) and test the learned model on the other one subset (*testing data*). The process is usually repeated for *k* iterations and is called *k*-fold cross validation. In each turn, a different subset is used for testing and the final performance is the average of the performance figures obtained in all iterations. Cross validation is a standard approach for evaluating NER and widely in the field.

2.5 Summary

This chapter has presented an overview of NER to provide a basic understanding of the task. Traditionally, NER is often divided into two sub-tasks, named entity detection or identification from text, which finds the boundaries of named entity mentions; named entity classification, which assigns semantic categories to identified entities. Extracted named entities can be ambiguous; however, traditionally disambiguation is not part of the NER process. NER is an important technique to many research fields, and has a wide range of applications. Methods for NER are generally classified into handcrafted rule based methods and machine learning based methods, which are further divided into supervised, semi-supervised and unsupervised methods depending on their requirements for training data. Although the dependence on training data in supervised learning methods may limit its application to some extent, it remains the dominant choice for NER tasks.

There is well-established practice for evaluating NER methods. The standard evaluation measures are precision, recall and F-measure.

The discussion in this chapter has focused on the traditional sense of NER, and aimed at providing an overall background of the field. The next chapter of this thesis discusses three major research questions concerning NER and addressed by this work. The literature reviews concerning each research question will also be presented in details in the subsequent parts of this thesis.

3 Research Questions

PREFACE

This chapter discusses several research questions related to Named Entity Recognition. It is divided into five sections. Section 1 gives an overview of the research questions and challenges concerning NER. Then each of the following three sections (2, 3, 4) discusses in details one specific research question that this thesis aims to address. The last section (5) of this chapter summarises the discussion.

3.1 Overview

NER is a very challenging task that has seen decades of research focusing on different research questions. One of the most extensively studied concerns training data, the essential input to most NER methods. Most research in this direction has focused on semi- and unsupervised learning methods that minimise the use of training data, while few have addressed the actual annotation process. Another major challenge concerns adapting an existing NER model built on certain training data to new datasets. The new data may differ in terms of the feature space whether or not they belong to the same domain of the training data. It is found that often, porting an existing model to new data results in damaged learning accuracy (Jiang and Zhai, 2007; Blitzer, 2008). Thus research has looked for methods of domain adaptation and transfer learning that are able to fit existing NER models to new data without re-training (Jiang and Zhai, 2007; Blitzer, 2008; Pan and Yang, 2010). Another frequently studied question concerns automatically generating background knowledge to support NER (Toral and Munoz, 2006; Smith and Osborne, 2006). Such background knowledge, typically in the form of gazetteers, is found to be very effective in improving NER learning accuracies. However, they are also difficult and costly to build and maintain. A closely related problem that is typically ignored by traditional NER is resolving ambiguities in NER output. As discussed before, entity names can be ambiguous and must be further processed to support machine interpretation or other applications. The process of resolving ambiguities can be considered as an additional ‘entity recognition’ step, in which the unique identity or entity referenced by an entity name is to be recognised. This thesis views Named Entity Disambiguation as NER in a different form.

As discussed before in Chapter 1, this thesis will focus on three research questions concerning training data annotation, gazetteer generation, and sense disambiguation.

3.2 Training Data Annotation

The need for training data – As discussed before, training data are the essential input to supervised learning methods. Although semi-supervised and unsupervised approaches have been introduced to cope with lack of training data in NER, supervised learning methods remain the primary choice in research and applications. In particular, supervised learning methods still dominate in adapting NER to new languages and domains. For example, supervised learning methods remain the primary approach for the Chinese (Duan and Zheng, 2011) and Arabic languages (AbdelRahman et al., 2010), and all participating systems in the IJCNLP 2008 Workshop on NER for South and South East Asian Languages are based on supervised methods (Singh, 2011). NER in technical domains such as history (Byrne, 2007), aerospace (Iria, 2009a) and biomedicine (Ju et al., 2011) has mostly adopted supervised learning methods using domain specific annotations. Particularly in the biomedical domain, there are continuous efforts and studies for creating training data (Ohta et al., 2002; Usami et al., 2011) for NER; as well as public evaluation tracks (Kim et al., 2004) to promote the usage of these data in NER tasks. The performance of semi-supervised approaches can also be controversial. Some studies of semi-supervised learning methods have shown compromised accuracy when compared against the best results reported for the supervised learning competitors. For example Gu et al. (2007) showed that their semi-supervised approach achieved an accuracy of 46.15 points in F-measure on a biomedical dataset when using only 50% of the training data. However, it is 10 points below a supervised model and nearly 26 points lower than the best performing supervised model on the same dataset.

Availability of training data – Thanks to decades of research in NER, several large datasets (Ohta et al., 2002; Doddington et al., 2004; Sang and Meulder, 2003) have been created and constantly maintained by vigorous communities. These are predominantly limited to the newswire and biomedical domains, which has seen considerable research effort over the past years. Similar data for other domains are extremely scarce. In many cases, annotated data cannot be made available for various reasons, which prevents reusability. For example, in commercial environments, documents can contain proprietary information and must not be released to the public (Iria, 2009a). In the clinic domain, due to the concerns of privacy, access to public clinical data has been very limited (Sasaki et al., 2007; Uzuner, 2008). In fact, the first public fully annotated and anonymised clinical corpus was only made available in 2007 in a shared task on clinical text classification (Pestian et al., 2007). This means that introducing supervised NER to new languages and domains often requires creating new annotated training data.

The challenge of document annotation – Creating high quality training data for supervised NER remains a major challenge in this field. It is a process that often requires substantial investment in terms of both personnel and finance. On the one hand, typical annotation procedures adopted for the creation of most public datasets require months and, in rare cases, years of effort from domain experts, linguists and even programmers

(Brants, 2000; Ferro et al., 2000). The high cost would make it inapplicable in many practical situations such as industries, due to resource limitations. On the other hand, despite the use of guidelines and common practices, the quality of annotations can still be unsatisfactory due to the intrinsic difference of human annotators' experience and knowledge. It has been shown that ensuring inter-annotator consistency has a major impact on the quality of training data and therefore, the ability of an NER system to learn. However, this is often difficult to achieve and the inter-annotator consistency reported in many datasets are very low (Saracevic, 1991; Colosimo et al., 2005; Murphy et al., 2006; Wilbur et al., 2006).

The research of **active learning** has been introduced to the field of NER (Shen et al., 2004; Laws and Sch ätze, 2008; Olsson, 2008), aimed at addressing the issue from a different perspective. Active learning aims to reduce the effort of annotation by involving both the annotators and the learning system in a series of annotation-learning cycles, in which both parties provide feedback to one another. The theory is that in each turn the learning system finds the candidates that it is most uncertain with and asks the annotators to annotate them. In doing so, the annotators avoid redundant annotations from which the learning system benefits little, but concentrate on the most informative examples that are most useful to learning. The outcome is reduced overall quantity of annotations but improved quality. However, a new challenge that comes with active learning is selecting the most appropriate examples for annotation, which can involve complex modelling and computation (Shen et al., 2004; Laws and Sch ätze, 2008). Furthermore, some comparative evaluation of machine learning based NER methods has shown that active learning does not always return its benefits (Ireson et al., 2005).

For these reasons, there is still the pressing need for better methods of document annotation to support training data creation for supervised learning methods. Ideally, the method should be easy to implement, and both effective and efficient. Solving this challenge will enable supervised NER to be built at lower cost, and also to be ported to new domains more easily.

3.3 Gazetteers as Background Knowledge

Background knowledge – It is well-known that the lexical-level features play a central role in NER (Li et al., 2009). Such features are usually gathered solely from the training data based on the annotated NEs and their contexts (Smith and Osborne, 2006). It has been argued that this type of features alone is often insufficient. It can be ineffective when contextual evidence is insufficient (Carvalho et al., 2008; Ganti et al., 2008), for highly ambiguous terms (Ratinov and Roth, 2009), and for out-of-vocabulary entity names which tend to cause ‘extreme sparseness in feature space’ (Li et al., 2009).

For these cases, incorporating ‘background’ or ‘external’ (Smith and Osborne, 2006) knowledge can lead to a better representation and eventually improve NER. Despite the lack of a formal definition of **background knowledge**, it is generally agreed that it refers to additional learning evidence that is unavailable from the lexical-level features in the training data.

Gazetteers as background knowledge – In NER, the most often used type of background knowledge is a gazetteer. Generally, gazetteers are a way to group related terms and map them to certain types or categories, such that the same types of named entities tend to be consistently associated to the same gazetteers. As discussed before, this thesis adopts two views of gazetteers: type-oriented or typed gazetteers, and alternative or untyped gazetteers. Typed gazetteers refer to the most commonly adopted sense of gazetteers, which usually contain reference named entities labelled by pre-defined types that are relevant to the task. For example, a person gazetteer may be used as background knowledge to recognise person entities. Alternative gazetteers refer to a more general sense. From the learning point of view, a gazetteer will be useful as long as it returns consistent labels even if these are not the desired named entity types, since the correspondence between the labels and the entity types can be learnt automatically (Kazama and Torisawa, 2008). For example, knowing that ‘Microsoft’ is a company and the fact that it often appears in the same clusters with ‘AT&T’, one can infer that the latter is also a company. In this case, the semantic category represented by the clusters is unknown a-priori; however, it provides additional learning evidence equivalent to a gazetteer. From this broader perspective, gazetteers can include automatically induced clusters of terms that group distributionally similar terms (Freitag, 2004; Miller et al., 2004; Jiang et al., 2006; Kazama and Torisawa, 2008; Saha et al., 2009; Finkel and Manning, 2009; Chrupała and Klakow, 2010), or automatically extracted hypernyms that group semantically similar sub-class concepts under the same super-class concept (Kazama and Torisawa, 2007a; Kazama and Torisawa, 2008).

Gazetteers are found to be particularly effective in improving the performance of NER systems when combined with other lexical-level features (Friedrich et al., 2006; Wang, 2006; Roberts et al., 2008; Saha et al., 2009). For example, in Mikheev et al. (1999), the use of gazetteers improved the accuracy of a supervised NER tagger by 39% in precision and 31% in recall. Particularly in technical domains, gazetteers or technical dictionaries are the major resource for resolving the complexity of domain-specific named entities.

Gazetteer generation – Similar to training data, gazetteers are language and domain specific. They are often unavailable, and creating and maintaining such resources require significant effort (Toral and Munoz, 2006; Kazama and Torisawa, 2008). Due to the evolutionary nature of human knowledge, vocabularies are constantly changing and meanings of particular terms evolve over time. The growth of vocabularies can be too fast to manage easily. For example, the Unified Medical Language System – UMLS (Bodenreider, 2004) is the largest knowledge base containing a large amount of controlled vocabularies in the biomedical domain. It contains a 'Meta Thesaurus', which is a repository of biomedical terminology and their relationships. Woods et al. (2006) reported that in the Meta Thesaurus, the number of new concepts introduced between 1998 and 2002 was over 300,000, while the increase between 2002 and 2003 was nearly 100,000, and today it stands at over two million. Manually creating and maintaining such resources requires significant investment.

As a result, another major challenge concerning NER is how to automatically build gazetteers – either as typed or the alternative form – to support NER. The availability of such methods can enable access to valuable background knowledge for NER, which in turn helps to improve its performance.

3.4 Resolving Ambiguities

Ambiguous entity names – As discussed earlier in Section 2.1, due to the polysemy of human language, the lexical realisation of an entity – the name or mention – can be ambiguous since it may be used to refer to multiple entities. For example 'Bush' in the sentence 'President Bush attended the opening ceremony of the Olympic Games in Beijing' is an ambiguous name that can refer to 50 different identities according to Wikipedia[3]. Combining the contextual word 'President' with additional further context may reveal it to be the 43rd president, George. W. Bush, and not the 41st president, George H.W. Bush or other persons. This type of ambiguities is found to be very common in NER datasets.

[3] http://en.wikipedia.org/wiki/Bush_(surname), last retrieved on 30 Nov 2011. All examples based on Wikipedia in this thesis are last checked up to this date.

To illustrate, the named entities extracted from the dataset used by the CoNLL2003 shared task (Sang and Meulder, 2003) are searched on a local Wikipedia copy (dated 5 Apr 2011) to retrieve the corresponding articles. Among the total of 34,870 (10,347 unique) named entities of all types, 30,377 (7,257 unique) have entries in Wikipedia. Among these, 17,742 – 58.4% (or 2,243 unique – 30.9%) have used an ambiguous name that can refer to multiple entities (because they can retrieve multiple articles).

However, traditionally resolving ambiguous names is not the goal of NER, which only deals with assigning high-level semantic categories rather than distinguishing instances. Instead, this is dealt with by the task of Named Entity Disambiguation (NED), a field closely related to Word Sense Disambiguation (WSD), where meanings of ambiguous words are resolved based their context, usually according to a sense inventory such as a dictionary that lists all possible word senses (Navigli, 2009). NED on the other hand, deals with ambiguous entity names.

Need for disambiguation – While most of the time ambiguity is not a concern for humans, for machines' interpretation of human language it is necessary to resolve the ambiguities and recognise the true entity that the name refers to. Many applications that build on named entities (e.g., see Section 2.2) either require a compulsory sense disambiguation procedure or can benefit from such a process. For example in relation extraction, given the sentence '<PERSON>Bowen</PERSON> published his work <MISC>'Two Intermezzi, Op. 141'</MISC> in 1951' and the knowledge that 'Bowen' in this context refers to the musician 'Edwin York Bowen (22 February 1884 – 23 November 1961)' rather than the novelist 'John Griffin Bowen (born November 5, 1924)' one can infer a more specific relation 'composer-of' between the PERSON and the MISC entity. In knowledge base generation and population, ambiguous entity names must be resolved to individual instances before they can be integrated into the knowledge base. For example, it is necessary to know if 'Manchester' refers to the city in the UK or the town in the USA when populating a knowledge base of world cities. Similarly, for question answering and semantic search, resolving ambiguities is an important step to understanding the users' intentions. For example to answer the question 'what is the last train from Sheffield to Manchester', the system should resolve both ambiguous location names 'Sheffield' and 'Manchester' (cities and towns of UK/USA) and group answers based on their referred instances.

The need to combine NER with a process of disambiguation has also been widely recognised. The TAC is currently organising annually competition events in which participants are invited to solve a task that extracts named entity mentions from a query and linking the mentions to unique entities in an existing entity knowledge base. Essentially the task requires both NER and NED.

Recognition or Disambiguation – From a theoretical point of view, NER and NED are two closely related, complementary tasks. In a broader sense, the two tasks can be considered equivalent. NER identifies named entity mentions in texts, and classifies them into semantic categories. On the one hand, this can be considered as 'recognising' the existence of named entities and their semantic categories; on the other hand, the classification process can be considered as a coarse-grained disambiguation process, which identifies entity name boundaries and resolves entity names to the closest semantic categories. In fact, the pioneer study in NED by Wacholder et al. (1997) particularly addressed NER by resolving two levels of ambiguities: resolving ambiguous boundaries between entity names (e.g., whether to split the phrase into two named entities by 'in' in 'The White House in Washington DC'), and resolving ambiguous names to the most suitable semantic categories (e.g., whether 'Washington' is a person or location). The first corresponds to the name or mention detection subtask in NER, while the latter corresponds to the semantic classification subtask. NED on the other hand, resolves entity names to unique real world entities. This is a fine-grained disambiguation process, which can also be considered as a process of 'recognising' the unique identities of named entities and therefore, a further 'recognition' step. Therefore, NER and NED essentially serve similar goals, but at different, complementary levels. Also due to the similar nature of the two tasks, a large number of empirical methods of NER and NED are also built on certain common grounds.

For these reasons, this thesis views sense disambiguation as the third challenge closely related to NER. Resolving ambiguities in NER output can enhance the 'recognition' of named entities from texts and enables the output of NER to be used by a wide range of applications.

Although an extensive amount of methods have been proposed for WSD and NED in the past, they are limited in different ways and many have adopted supervised methods (Navigli, 2009). This thesis will explore unsupervised methods that do not require training data for NED.

3.5 Summary

This chapter has introduced several research questions to be addressed by this thesis. The first concerns how to create training data in an effective and efficient way. The second concerns how to automatically acquire background knowledge in the form of gazetteers. The third concerns how to perform unsupervised sense disambiguation for named entities extracted by NER. These research questions are core to NER and interrelated. Addressing the training data annotation issue can lower the barrier of porting supervised NER methods across domain boundaries, potentially enabling NER in a wider range of contexts. Addressing automatic gazetteer generation is one of the critical strategies for improving NER learning accuracy, particularly in specialised domains. Resolving ambiguities in NER output and assigning unique identities essentially addresses 'recognition' at a further level, and ensures the output of NER to be ultimately useful to a wide range of tasks. Each research question will be addressed separately in the following chapters of this thesis.

4 A Uniform NER Model for this Thesis

PREFACE

This chapter presents a basic supervised learning model for Named Entity Recognition that lays the common ground to the studies in later chapters. The supervised learner is based on a Support Vector Machine classifier, which will be introduced in Section 1 of this chapter. A number of features to be used by this model will be described in Section 2.

4.1 An SVM Model for NER

SVM is a widely used supervised machine learning algorithm for NER in a wide range of domains (Isozaki and Kazawa, 2002; Mayfield et al., 2003; Li et al., 2005; Ekbal and Bandyopadhyay, 2010). An SVM learner aims to learn a classification model for a problem from a set of training data $\mathcal{D}$:

$$\mathcal{D} = \{(x_i, y_i) \mid x_i \in \Re^p, y_i \in \{-1,1\}\}_{i=1}^{n} \qquad \textbf{Equation 4.1}$$

$\mathcal{D}$ is the training dataset containing a set of pairs (x_i, y_i), where x_i is a single training data instance represented by a p-dimensional real vector, and y_i is either -1 or 1, indicating whether the instance x_i belongs to a particular class. The p-dimensional vector is created based on the features associated with the training data instances. The number of dimensions of the vector is defined by the total number of unique features that represent all training data instances. Given a training dataset, the SVM algorithm tries to find a *maximum-margin hyperplane* in this p-dimensional space that divides the instances having y_i=1 from those having y_i= -1. The intuition is that this hyperplane has the largest distance to the nearest training data points of any class, which effectively reduces the generalisation error for the classifier. This can be illustrated using Figure 4.1.

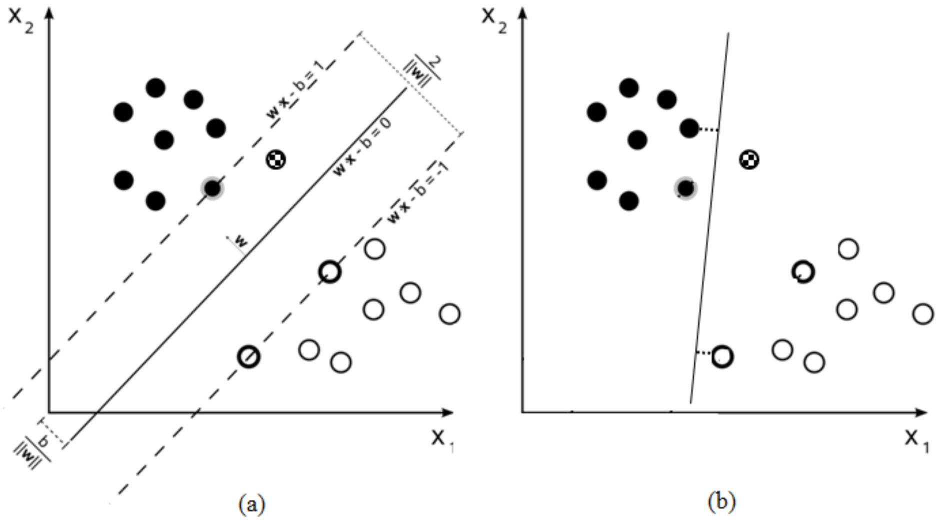


Figure 4.1. Two hyperplanes (solid line) that can be learnt using a sample of training data. Adapted from Wikipedia (Wikipedia, 2011)

In Figure 4.1, dark and white spots represent training data instances of two different classes. The dotted spot in (b) represents a new instance of the dark class to be classified that is unseen at training. Given the training data instances as shown, multiple hyperplanes can be plotted to separate instances of one class from another. However, the hyperplane in (a) has the maximum margin to both classes and therefore, minimises generalisation error. This can be illustrated by introducing a new instance of the dark class for classification as shown by the dotted spot in both (a) and (b). The hyperplane in (b) will make an incorrect prediction, while the maximum-margin hyperplane in (a) still succeeds. This hyperplane can be written as a set of points **x** satisfying:

$$\mathbf{w} \cdot \mathbf{x} - \mathbf{b} = 0 \qquad \textbf{Equation 4.2}$$

where **w** is the normal vector to the hyperplane and · denotes the dot product. The maximum-margin given at this hyperplane is the distance between two parallel hyperplanes that are as far as possible while still separating the data. The two hyperplanes can be described by the following equations:

$$\mathbf{w} \cdot \mathbf{x} - \mathbf{b} = 1 \text{ , a hyperplane closest to x such that y=1} \qquad \textbf{Equation 4.3}$$

$$\mathbf{w} \cdot \mathbf{x} - \mathbf{b} = -1 \textbf{ ,} \text{ a hyperplane closest to x such that y=-1} \qquad \textbf{Equation 4.4}$$

Using geometry the distance between the two hyperplanes is calculated as 2/||**w**||. As a result, the SVM learning problem is transformed into an optimisation problem that looks for a solution to minimise ||**w**|| while satisfying the condition $y_i(w \cdot x_i - b) \geq 1$. Details of how to solve this optimisation problem is beyond the scope of thesis. Readers may refer to Burges (1998) for a full explanation.

After training, the maximum-margin hyperplane is obtained. At application time, previously unseen data are mapped onto the same *p*-dimensional space and the predictions are based on what side of the hyperplane the new instances fall on.

The SVM model as discussed so far works well for data which is linearly separable as in the case of the example in Figure 4.1. However, in some cases, the data are not separable linearly. To fit such data usually a kernel trick is used (Boser et al., 1992) to replace every dot product by a non-linear kernel function. This allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. The transformation may be nonlinear; thus although the classifier is a hyperplane in the higher-dimensional feature space, it may be nonlinear in the original input space.

SVM v.s. other models – Most statistical machine learning algorithms are often classified into discriminative and generative models. A discriminative model learns the conditional probability distribution *P(y|x)* directly from data. A generative model learns the joint probability distribution *P(x, y)* then infers *P(y|x)* by applying the transformation:

$$P(y \mid x) = \frac{P(x,y)}{P(x)}$$ **Equation 4.5**

Classification tasks do not necessarily require the joint distribution since the goal is to predict *y* given *x*. Therefore, intuitively discriminative models take a direct strategy to the task. On the other hand, generative models are more flexible in encoding dependencies in complex learning tasks. It is also known that discriminative models perform better when plenty of training data are available but can become less effective than generative models in semi-supervised learning tasks (Bishop and Lasserre, 2007).

In the NER task, examples of commonly used discriminative models include SVMs, Perceptrons, Maximum Entropy (ME) models and CRFs; while examples of commonly used generative models include HMM and Naïve Bayes. It is well-known that SVM is closely related to the Perceptron learning algorithm, both of which are designed with the goal to find hyperplanes that separate two classes of data. In comparison, SVM is known to be more robust due to margin maximisation, while Perceptron can often overfit the training data. The ME model is also widely used in NER, however, many (Kazama et al., 2002; Wu et al., 2006b) have shown that SVM generally achieves better results than ME. CRF is essentially a kind of Maximum Entropy Markov Model (MEMM), which combines features of HMM and ME. Its strength stems from its ability to encode interdependencies between labels. The hypothesis is that certain NE types can be inter-related, which for example, can be indicated by higher probability of co-occurrence. The SVM algorithm cannot incorporate such information naturally. This must be encoded as separate features. Although some (Li et al., 2008) have reported better results using CRF, others (Tsochantaridis et al., 2005) have shown that SVM was better in NER and Keerthi and Sundararajan (2007) demonstrated that the two were quite close in performance when identical features are used.

SVM for NER – This thesis adopts a linear SVM model for NER for two main reasons: 1) it is state-of-the-art and empirically achieves competitive performance compared against other models (Takeuchi and Collier, 2002; Wu et al., 2006b; Lam, 2010); 2) this research deals with supervised learning tasks in which a discriminative model learns well.

The SVM model adopted in this research is based on that by Finn (2006), which represents the standard approach in SVM-based NER (Lee et al., 2004; Li et al., 2005; Iria et al., 2006). Learning is performed at the token-level, whereby each token is a classification instance and represented as a high-dimensional feature vector. Learning is divided into two stages: **classification** and **decoding**. In the **classification** stage, for each NE type, two binary classifiers are learnt to recognise the *start* and *end* tokens of NEs of that type. All tokens that begin a labelled NE are positive instances for the *start* classifier, while all the other tokens become negative instances (*not-start*) for this classifier. Similarly, the positive examples for the *end* classifier are the last tokens of each labelled NE, and the other instances are negative examples (*not-end*). For example, to build a classifier for recognising Person names, one binary classifier is trained to recognise the beginning of a Person NE; and the other binary classifier is trained to recognise the end of a Person NE. An illustration is shown in Figure 4.2. Thus for an NER learning task that concerns *n* types of NEs, *2n* binary classifiers are trained and applied independently to recognise the start and end boundaries of NEs.

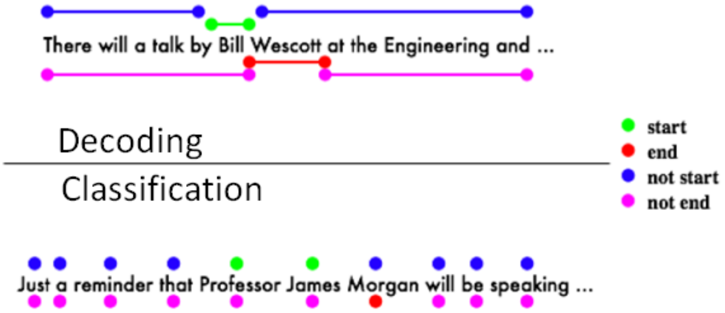

There will a talk by Bill Wescott at the Engineering and ...
Decoding
Classification
start
end
not start
not end
Just a reminder that Professor James Morgan will be speaking ...

Figure 4.2. Illustration of the SVM-based NER model, taken from Finn (2006).

The output from the classification stage is a collection of start and end tags for different NE types. Next, the **decoding** stage scans the output and matches start tags with end tags to create the final NE labels. In Finn (2006), a histogram is firstly generated based on the number of tokens between each start and end tag in the training data. To match predictions, the probability of a start tag being paired with an end tag is estimated as the proportion with which a field of that length occurred in the training data. Furthermore, Iria (2009b) extended this by introducing a number of heuristics such as favouring shorter NEs (NEs with fewer component tokens) rather than longer NEs (using arbitrary threshold of length). Iria's method is followed in this study. Predictions with multiple NE types (i.e., two annotations that are labelled as different NE types and have overlapping tokens) are also resolved following this procedure.

This thesis does not re-implement the model described above but adopts an existing implementation that matches the described model. This is the SVM-based NER tagger in T-Rex (Iria, 2009b), which uses the SVM-light package (Joachims, 1999). All default parameters have been used in this work.

4.2 Features

As described in the previous section, to apply the SVM-based NER, each token must be transformed to a *p*-dimensional real vector, which is to be classified into the start, end boundaries of NEs, or nothing. The vector encodes various features related to the token, and the dimension depends on the total number of unique features that represent all training data instances. A survey on a wide range of features used for NER can be found in Nadeau (2007a). This section briefly introduces some state-of-the-art features that are used in the experiments to be reported in the remaining sections of this thesis.

Token features map the presence of specific words or symbols to specific named entity types. For instance, the presence of the token 'Smith' can be highly indicative that the word represents (part of) a person entity name.

Stem and **lemma** are used to refer to two types of base forms of word. For grammatical reasons, documents use different inflectional forms of a word, such as 'street', 'Streets', 'going' and 'go'. Additionally, there are families of derivationally related words with similar meanings, such as 'realise', 'realistic', and 'realisation'. Stemming and lemmatisation are two techniques to reduce inflectional and derivationally related forms of a word to a common base form. Stemming usually adopts crude heuristics to return a word to a 'pseudo' base form that never changes with inflection or derivation. Thus it often collapses derivationally related words. Lemmatisation uses vocabulary resources and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word known as the lemma. Given the token 'produced', stemming might return 'produc' because there are words such as 'production', whereas lemmatisation would return 'produce'. Therefore, stem or lemma features map the presence of certain base forms of words to specific named entities. In this work, the stemming algorithm is described in Porter (1980) and the lemmatisation is processed using the Dragon Toolkit (Zhou et al., 2007).

Orthography of a token gives information about a word's capitalisation, use of upper case letters, digits and other word formation information such as hyphens ('Stoke-upon-Trent') and punctuations ('Mr.').

Part-of-Speech (POS) tags describe linguistic categories of words (or more precisely lexical items). These are generally defined by the syntactic or morphological behaviour of the lexical item in question. Common categories include noun, verb, adjective and adverb among others. The standard set of POS tags used in the literature as well as in this thesis can be found at the Penn Treebank POS website (Penn Treebank Project, 1998).

Gazetteer in general is a way to group related terms and map them to certain types or categories, such that the same types of named entities tend to be consistently associated to the same gazetteers. In the most commonly adopted sense, a gazetteer contains reference entity names of one pre-defined entity type that are relevant to the task. For example, a person gazetteer may be used as features to recognise person entities. Alternatively, in a more general sense, gazetteers can be any ways of grouping related terms even if they are untyped. For example, they can include automatically induced clusters of terms that group distributionally similar terms. Typically in NER, the gazetteer feature for a token is encoded as a binary value depending on whether or not the token is included as (part of) a term in the gazetteer (e.g., if 'Bob' is included by a person-first-name gazetteer it receives 1 for this gazetteer feature and 0 otherwise). When multiple gazetteers are used, the token receives a binary feature for each gazetteer.

Context Window is a method to include the features of the surrounding tokens of a named entity mention in text. The motivation is that similar types of named entities will be described by similar words, which are often around the named entity mentions. For example, the word 'Mr.' often indicates that the next token followed is often a person name. The context window is usually set to a fixed size *n*, which means to include *n* tokens before and *n* tokens after the current token, in the construction of the feature vector for the token.

Each token in the text is then encoded into a binary vector for classification based on the features described above. Thus given the example annotated text shown previously in Figure 2.1 in Section 2.1, Figure 4.3 below shows the transformed feature representation vectors of the token 'Dooner' using the features *token (t)*, *token Part-of-Speech (pos)*, *context window=1*. ‘prev1’ represents the *previous one* token to the current token; ‘0’ indicates the feature is inapplicable; ‘1’ indicates the feature is applicable.

Features	McCann
t=Dooner	0
t=McCann	1
t_pos=NNP	1
t_prev1=Mr.	0
t_prev1=about	1
t_prev1_pos=RP	0
t_prev1_pos=IN	1
t_next1=met	0
t_next1=acquiring	1
t_next1_pos=VBD	0
t_next1_pos=VBG	1

Figure 4.3. Examples of feature representation for NER

4.3 Summary

This section has introduced the statistical SVM model for classification tasks and how it can be tailored for NER. Given a set of training data containing instances of two different classes that are represented as high-dimensional vectors, SVM learns a maximum-margin hyperplane that separates the data. At application, new instances are projected as the same high-dimensional vectors, and the predictions are based on what side of the hyperplane the new instances fall on. To tailor SVM to NER, each type of named entities is treated separately and a boundary detection model that learns to recognise the start and end boundaries of named entity mentions is used. The output is then combined using heuristics. A list of features to be used with this model has also been introduced.

This supervised learning model makes the fundamental NER system to later chapters in this thesis and is used in corresponding experiments involving different domains and datasets. The evaluation of this model will be presented later in corresponding chapters under specific tasks.

Part II – Training Data Annotation

This part addresses the first research question concerning training data annotation for NER. This is presented in Chapter 5.

5 Training Data Annotation

PREFACE

This chapter addresses the first research question: effectively and efficiently creating training data for supervised Named Entity Recognition. Training data are created by document annotation, a process in which the NER task is undertaken by human domain experts, who tag instances of named entities within a collection of documents manually. This chapter is divided into six sections. Section 1 gives an introduction to the problem. Section 2 presents related work. Section 3 discusses the limitations of existing approaches to document annotation, which motivates the research for an alternative approach. Section 4 introduces the new approach to document annotation using a case study. Section 5 discusses the results and the final section (6) concludes this chapter.

5.1 Introduction

Training data play a central role in supervised learning methods. For many NLP and IE tasks such as Part-of-Speech tagging, chunking, sense disambiguation, NER and relation extraction, the training data are created by manual document annotation. It is a process that requires humans (usually domain experts) to undertake the same tasks to create annotated data which will be used as examples to train a learning model. For NER, this involves tagging named entity instances by pre-defined semantic categories in a collection of documents.

It has been recognised that document annotation is the major bottleneck to the development and adaptation of supervised learning tasks (Nadeau, 2007b; Howlett and Curran, 2008). The process is not only laborious and costly, but also difficult. Crucial to the annotation process is resolving annotator discrepancies and achieving reasonable inter-annotator agreement, the problem stemming from annotators behaving differently and inconsistently for the same annotation task. This is due to the differences in their skills, knowledge and experiences, and issues such as workload and tiredness. The problem can affect the quality of annotation and therefore, the learning accuracy of a system (Brants, 2000; Ferro et al., 2000).

To address this issue, the typical annotation process requires a number of domain experts to work in an iterative and collaborative manner in order to discover and resolve discrepancies progressively. Usually in each iteration, a set of documents are duplicated across all annotators, who are required to annotate the same documents for the same types of information (e.g., for NER this could be the same semantic categories such as person and location names) independently. Outputs from different annotators are then cross-checked, discussed, validated, consolidated and a sophisticated annotation guideline is documented, followed, and refined (Brants, 2000; Ohta et al., 2002; Kim et al., 2008) in following iterations. The process is repeated as much as possible until the level of discrepancies is reduced to a satisfactory level. Such a repetitive process in some cases can require months and in rare cases, even up to years of work of experienced annotators (Brants, 2000; Wilbur et al., 2006). However, discrepancies can never be eliminated (Hripcsak and Wilcox, 2002); they can remain high in some cases (Saracevic, 1991); and the resulting annotations and guidelines are often application-specific and non-generalisable. The high cost of hiring annotation experts means that it is inapplicable in many practical situations such as industries, due to resource limitations (Iria, 2009a). A more effective and efficient approach is required.

The remainder of this chapter analyses the problem in details and proposes a different solution. It is structured as follows: Section 5.2 presents a literature review to better understand the problem and limitations of the commonly adopted approach to document annotation. Section 5.3 describes a different viewpoint at the problem and introduces the fundamental hypothesis of the novel approach. Section 5.4 presents the approach using a real NER annotation case study, detailing the design of the experiments and key findings. Section 5.5 presents the final results and further discussions. Section 5.6 concludes this chapter.

5.2 Related Work

Previous studies on document annotation practices have focused on tackling annotator discrepancy – particularly inter-annotator discrepancy, which is often considered the main contributing factor to annotation quality and an important determinant for annotation effort.

5.2.1 Annotator Discrepancy

Research has shown that human annotators can never agree completely with each other on what and how to annotate (Hripcsak and Wilcox, 2002), and they even tend to disagree with themselves in some situations (Cucchiarini and Strik, 2003). The first case is often referred to as **inter**-annotator ‘**agreement**’, ‘**consistency**’ or ‘**discrepancy**’. The second case is referred to as **intra**-annotator ‘**agreement**’, ‘**consistency**’ or ‘**discrepancy**’. Inter-annotator discrepancies are often caused by the differences in annotators’ knowledge and experiences, their understanding and reasoning of the corpora (Kim et al., 2008). Intra-annotator discrepancies exist because annotators’ level of interest and motivation may drop and level of fatigue rises as the annotation process continues (Gut and Bayerl, 2004); as a result, annotators make mistakes. Most studies have focused on inter-annotator discrepancy, possibly because it is naturally much easier to solve intra-annotator discrepancy (i.e., to agree with yourself) than inter-annotator discrepancy (i.e., to get others to agree with you). One of the most often used metrics for evaluating IAA is the ƙ-statistic (Carletta, 1996).

Inter-annotator discrepancy is a prevailing issue in many research fields. Depending on the specific task, the inter-annotator agreement can vary significantly. For example, Saracevic (1991) indicated that the agreement between human annotators varied between 40% and 75% for different tasks. Most reports of inter-annotator discrepancy are found in the field of NER. Research by Fort et al. (2009) has shown that in these tasks, discrepancies typically arise due to three types of difficulties in annotating entities. Firstly, it is difficult to determine the right category and what they encompass (e.g., ‘Kofi Annan’ can be ‘Person’, but what about ‘Kennedys’, ‘the Conservatives’); secondly, it is difficult to select the candidate texts and delimitation boundaries (e.g., should annotators only tag proper nouns, or also pronouns and definitional descriptions); thirdly, how to annotate homonyms, e.g., ‘England’ may refer to a location or a football team. These problems become even harder to resolve within specialised domains such as biomedicine and engineering, due to the intrinsic complexity of terms in these domains including multi-word expressions, complex noun phrase compositions, acronyms, ambiguities and so on (Tanabe et al., 2005). Typically, the inter-annotator agreement in NER found in these domains is between 60% and 80% (Colosimo et al., 2005; Murphy et al., 2006; Wilbur et al., 2006), often measured by the ĸ-statistic.

From all these studies, it is evident that perfect agreement between annotators is difficult to reach, and it is also difficult to obtain a high level of inter-annotator consistency, especially in specialised domains. However, researchers argue that consistency highly increases the usefulness of annotations for training or evaluation purposes, and it is crucial to the success of machine learning algorithms (Brants, 2000; Ferro et al., 2000). Therefore, research has been conducted to study scientific methods for creating high quality annotations and addressing inter-annotator consistency.

5.2.2 The State-of-the-Art Approach

As introduced before, the typical process of annotating a corpus often involves a group consisting of a number of domain experts and ideally also linguists working on a same range of annotation tasks in an iterative and collaborative approach aimed at resolving discrepancies. The entire process and decision making logic is documented to form a guideline for the annotation task, which is to be followed in future exercises. Due to the nature of the work, it is always a lengthy and costly process. The guidelines are often subject to the specific domain and not generalisable to other problems.

For example, Brants (2000) reported their work on creating syntactic annotations (part-of-speech and structural information) on a German newspaper corpus. The activity involved trained annotators performing the annotation tasks at sentence level independently, then cross-checking and discussing together to resolve discrepancies. They reported that a trained annotator needs on average 50 seconds per sentence, with an average of 17.5 tokens; however, the total annotation effort including the consolidation activity increases to 10 minutes per sentence. Pyysalo et al. (2007) annotated a corpus of 1,100 sentences from abstracts of biomedical research articles for biomedical named entities, relationships between named entities and syntactic dependencies. They also adopted a repetitive process, which took 15 person-months of effort. Wilbur et al. (2006) conducted experiments to investigate inter-annotator agreement in a text annotation task in biomedical domain and identify factors that can help improve consensus. Their experiment involved twelve annotators annotating a same set of 101 sentences. Multiple iterations were conducted in a period of over one year, during which they developed and refined a guideline considered applicable for similar annotation problems. The resulting inter-annotator agreement remained between 70% and 80%. They concluded that annotators must have a good understanding of the language and experience in reading scientific literature, and must be properly trained in order to deliver high quality annotations. Also, they indicated the availability of clear, well developed annotation guidelines as critical.

Other researchers have also recognised the necessity for clear annotation guidelines. Kim et al. (2008) showed by experiments that high level of discrepancy will form without annotation guidelines even if the task is carried out by well-educated domain experts. Their studies on event annotation on the Genia corpus took 1.5 years of effort of five graduate students and two coordinators. Whenever new annotators joined the project, they had to be trained using previously annotated examples and follow the guideline. Colosimo et al. (2005) and Tanabe et al. (2005) also conducted corpus annotation in the biology domain and concluded that clear annotation guidelines are important, and the annotations should be validated by proper inter-annotator-agreement experiments.

Even if well-prepared guidelines are available for annotation problems, they are not the ultimate answer to the problem. Firstly, most guidelines are lengthy documents and are difficult to read. For example, Ferro et al. (2000) designed guidelines for annotating temporal information, which has 57 pages. The entity recognition task defined by ACE (Doddington et al., 2004) is accompanied with guidelines of over 70 pages for annotating only five classes of entities. Secondly, interpretation of the guideline documents differs from annotator to annotator; as a result, some annotation criteria remain problematic and can cause discrepancies (Fort et al., 2009). For example, the event annotation on the Genia corpus by Kim et al. (2008) only achieved 56% inter-annotator agreement with strict match (Morante et al., 2009) even though all annotators were trained and educated using example annotations and guidelines.

To summarise, the standard approach adopted by the literature requires substantial investment, including clear definition of annotation guidelines to be created and followed; well-educated domain experts with proper training in document annotation, careful study of inter-annotator agreement and iterative attempts to address the issues revealed by the study and to resolve discrepancies. Many scientific research tracks such as MUC present a scenario in which the cost of such effort is not considered important (Ciravegna et al., 2000). However, the scenario breaks as the technology is to be adopted by various specialised domains, in which the cost is a serious issue (Nadeau, 2007b). Industries and businesses are not willing to invest resources (personnel, finance and time) into lengthy document annotation exercises (Iria, 2009a); annotators feel overwhelmed by the scale of monotonous annotation tasks expressing a strong reluctance to doing them.

One exception to this is the domain of biomedicine, where well-curated resources are richly available and users are more familiar with the benefits that can follow from annotation. Unfortunately, these resources are hardly re-usable across domains because they address specific issues in bio-informatics; and demands for similar resources in other specialised domains such as aerospace engineering, astronomy and arts and humanity are equally high, these however are scarcely addressed (Murphy et al., 2006; Jeffrey et al., 2009; Iria, 2009a).

Given the complexity of these problems, the crucial status of supervised methods in the NER field and their dependence on training data, there is a strong demand for efficient and practical approaches to manual document annotation.

5.3 Hypothesis

Essentially, the majority of discrepancies among annotators are caused by the differences in their knowledge and experiences (Hripcsak and Wilcox, 2002). The traditional annotation practice aims to identify these differences and minimise them by collaborative and iterative exercises, eventually producing an output that best matches the subtly varying viewpoints across a community. This thesis takes a different viewpoint at the problem that is based on the following hypothesis:

> **H1. Training data annotation: the discrepancies among annotators, caused by the difference in their knowledge and experiences, indicate different levels of annotator's *suitability* for an annotation task. It is possible to assess such suitability and define suitability-based tasks so as to ensure annotations to be generated in a more effective and efficient way.**

This is inspired by the real life experiences that people typically specialise in one or several areas and no one can be perfect for any tasks. For example, computational linguists often have expertise in specific subjects such as sense tagging, NER, sentiment analysis etc. Even experts of the same subject can have varying levels of expertise in terms of the methods, domains, and other factors. In practice, when allocating tasks it is natural to match a candidate's knowledge and experiences against the requirements of the task – an act of assessing suitability.

Similarly, a document annotation task can also be further divided into sub-tasks, each requiring different types of knowledge. Consider a named entity annotation task that requires tagging persons by different occupations, e.g., politician, musician, sports person. Normally, a person who likes sports may be able to do a better (faster and more accurate) job on tagging sports person while a person who is a music fan may be quicker at annotating musicians, because they both possess certain specialist background knowledge that can support their understanding of the content. They both, however, may find it more difficult to annotate politicians and will more likely make mistakes. For technical, specialist domains the problem can be even more acute. For example, it may be difficult to distinguish RNAs, DNAs, protein names, cell types and cell lines when annotating biomedical named entities even for domain experts, due to the complexity and the evolutionary nature of vocabulary in this domain. In these cases, special knowledge about one of these entity types clearly makes a person a better candidate for the annotation of that particular type. On the contrary, inclusion of annotators that lack such knowledge is more likely to cause inconsistencies and errors.

Therefore, the key to improving annotation quality is identifying annotators' suitability and suitability-based task allocation. The annotation task should be sub-divided into smaller components, based on which annotator's suitability can be assessed and assigned to different sub-tasks. Inconsistent annotators unsuitable for a sub-task should be identified and isolated such that the annotation quality is not compromised.

5.4 Suitability-based Document Annotation

This section proposes a new approach towards manual document annotation. The approach will be presented using a case study of named entity annotation in a specialised domain – archaeology. This was part of the effort in a real-life project of enabling e-archaeology (Archaeotools, 2007), which had a substantial focus on NER for the archaeology domain. The annotated data are later adapted and used by further studies of this thesis.

The proposed approach is based on the hypothesis that the different levels of knowledge and experiences of annotators lead to different levels of suitability for an annotation sub-task. This is reflected by inconsistent levels of discrepancies they demonstrate in annotating each type of named entity. Therefore, a named entity annotation task can be further divided into sub-tasks based on the named entity types. Annotator discrepancies and suitability must be studied on a per-entity-type basis, and only the most suitable annotators should be selected for annotating specific named entity types other than all types.

The approach contains four phases, which are illustrated in Figure 5.1. The first phase follows the traditional approach of manual document annotation to study the level of discrepancy of the task using a sample of data. The size of this sample is controlled to ensure that the efforts required from annotators are minimised to an acceptable level, also that the annotations created are adequate for studying the inter-annotator agreement. In the second phase, a set of experiments are carried out to evaluate machine learning accuracy using these annotations. In the third phase, the results from the previous two analyses are used to evaluate annotators' suitability of annotating a particular type of named entity. From this a mapping between named entity types and their **best-fit-annotators** can be created – specific annotators are chosen to annotate the types of named entities for which they are most suitable based on these mappings. In the final phase, the final set of documents to be annotated is firstly selected. Then for each named entity type, the documents are split proportionally between each member of the best-fit-annotators for that type. This ensures all documents are annotated by the most consistent annotators for all named entity types, while no annotators perform redundant work. Compared to the traditional approach, this is a desirable feature since the distributional nature of work in the final phase allows workload to be reduced and total output to be increased. A set of experiments are then carried out to evaluate the machine learning accuracy obtainable on the final annotations.

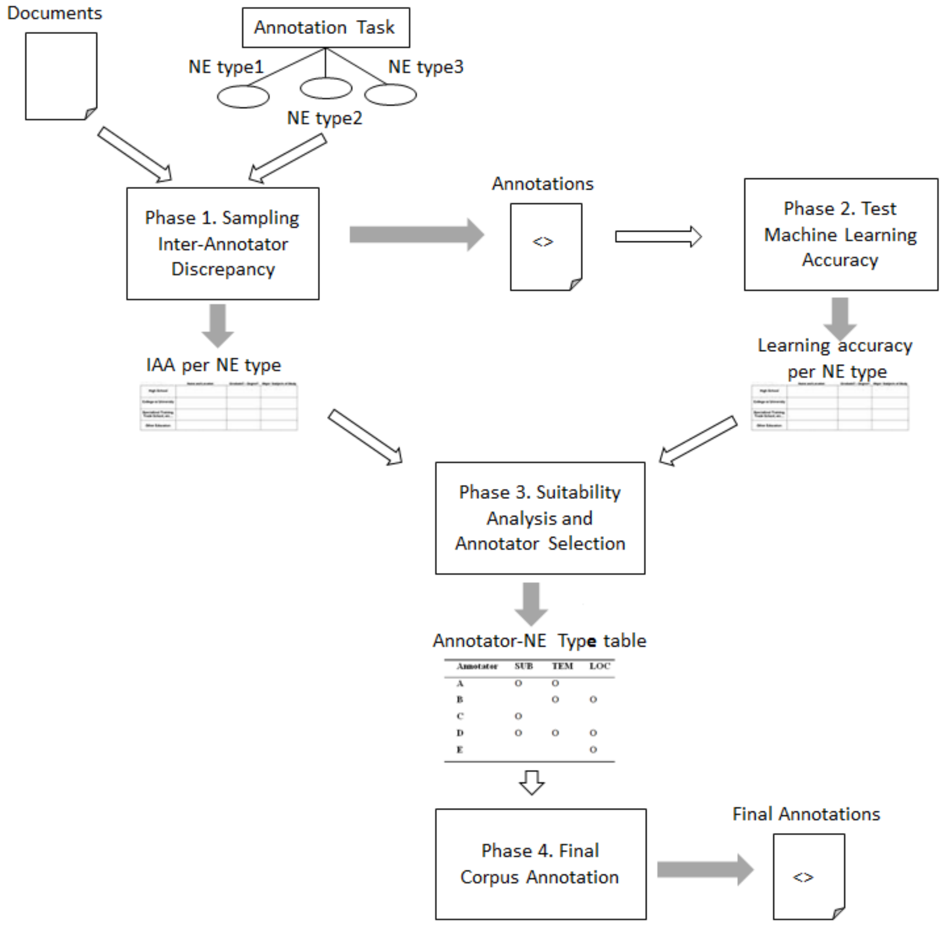

Documents
Annotation Task
NE type1
NE type2
NE type3
Phase 1. Sampling Inter-Annotator Discrepancy
Annotations
<>
Phase 2. Test Machine Learning Accuracy
IAA per NE type
Learning accuracy per NE type
Phase 3. Suitability Analysis and Annotator Selection
Annotator-NE Type table
Annotator SUB TEM LOC
A O O
B O O
C O
D O O O
E O
Phase 4. Final Corpus Annotation
Final Annotations
<>

Figure 5.1. The overall workflow of suitability based document annotation process

5.4.1 Case Study: the Archaeology Domain

The domain of modern archaeology is a good representation of the document annotation problem for two reasons. Firstly, compared to other domains such as newswire and biomedicine, it is rarely addressed in NER but has a pressing need for automatic knowledge acquisition technologies (Jeffrey et al., 2009). It is a discipline that has a long history of active fieldwork and a significant amount of legacy data dating back to the nineteenth century and earlier. Despite fast-growing large corpora existence, little has been done to develop high quality metadata for efficient access to the contained information in these datasets. This is because annotating archaeological documents is a challenging task due to the complexity of language characterised by ambiguities, uncertainties, long and composite terms, changing language use over the extended timeframe of the corpora, acronyms and so on. As a result, low inter-annotator agreement has been noted in related work (Byrne, 2007).

The documents to be used for annotation are a typical representation of the unstructured legacy data in the domain. The collection consists of full-length archaeological reports archived by the Arts and Humanities Data Service in the UK (AHDS, 1995). The reports vary from five to over a hundred pages. According to Jeffrey et al. (2009), important facts in archaeology data can often be summarised by three types of information: *what, where* and *when*. They correspond to three types of named entities:

- **Subject (SUB)** – concerns the 'what' information. These are often the objects that a report refers to, such as findings of artefacts and monuments. This is the most ambiguous type because it covers various specialised domains such as warfare, architecture, agriculture, and machinery. Also, it includes a wide range of general concepts rather than 'named' entities. Examples of such include 'Roman pottery', 'spearhead', 'shard', 'chapel', 'arrowhead' and 'courtyard'.
- **Temporal terms (TEM)** – concerns the 'when' information. These are often mentions of archaeological dates related to findings or events. They are written in a number of ways, such as numerical expressions '1066 - 1211', 'circa 800AD'; centuries 'C11', 'the 1st century'; and concepts 'Bronze Age', 'Medieval'; and acronyms such as 'BA' (Bronze Age), 'MED' (Medieval).
- **Location (LOC)** – concerns the 'where' information. These are typically place names related to findings or events, such as names of cities, streets, place of interests and excavation sites (e.g., Sheffield, City of York, York Minster, the Tower Bridge, A61, M62).

5.4.2 Phase 1 – Sampling Annotator Discrepancy

Purpose – The purpose of this step is to sample the inter-annotator discrepancy in the named entity annotation task for archaeology. Each type of named entity is treated separately and the inter-annotator discrepancies for each type are analysed. The hypothesis is that annotators may have different levels of knowledge and understandings of different concepts such that their suitability for annotating specific named entity types will differ.

The sample corpus – In this phase, five documents were randomly selected from the AHDS archive. Each document varied from five to thirty pages. These documents are much larger than the standard datasets used in the newswire and biomedical domains, where typically short articles and abstracts of reports are used. Meanwhile, the selection criteria ensured that there were sufficient contents for annotation (as indicated by the tag density and number of annotations revealed in the post-annotation statistical analysis). The total number of words in this corpus was 47,101. On average, the total number of annotations created by each annotator was approximately 2,100, with 58% for SUB, 19% for LOC and 23% for TEM. This corpus is referred to as '*sample corpus*'. It was then to be annotated by five archaeology researchers in four iterations following the traditional document annotation approach.

The state-of-the-art process – Throughout phase one, two annotators were constantly involved in all meetings with knowledge acquisition (KA) experts to provide feedback from all annotators and design simple annotation guidelines and ensure they are followed. The annotation process consisted of four mini-iterations, as shown in Figure 5.2.

Iteration 1	A trial attempt to identify major discrepancies and create initial guidelines (2 annotators, 2 documents)
Iteration 2	A further test round to refine the guidelines (5 annotator, 1 ~ 2 documents each)
Iteration 3	Re-annotate the whole sample corpus (5 annotator, all documents)
Iteration 4	Validation against the guideline (1 annotator, all documents)

Figure 5.2. Four mini-iterations adopted in Phase 1 document annotation

In the first iteration, two annotators made trial attempts at annotating two medium sized documents from the sample corpus. Discrepancies were identified at this early stage and were discussed and resolved in the meeting with the KA experts. The outputs of this process were some guidelines for annotation, which were then provided to all five annotators in the second iteration, during which each annotated between 1 and 2 documents. The purpose of this exercise is again to identify as many discrepancies as possible. By studying these annotations, the guidelines for annotation were further refined and enriched. In the third iteration, all five annotators were required to follow the guideline to re-annotate the trial corpus independently and fully in a series of intensive workshops. In the final iteration, one annotator undertook final validation by checking 10% of all annotations to correct obvious mistakes that violated the guidelines. These corpora are used to study inter-annotator consistency and machine learning accuracy.

Cost of the Process – due to the sampling technique, according to the annotators’ estimation, the first iteration of Phase one took 2 person-days of work; the second iteration took 5 person-days of work; the third iteration took 5 person-days of work; and the final iteration took 2 person-days of work. The total estimated cost in terms of person-days of work is 14.

Inter-Annotator Agreement – As mentioned before, the most popular approach for measuring IAA is the ƙ-statistic. However, it is argued that ƙ-statistic is not a very suitable measure when evaluating inter-annotator agreement in NER tasks (Pyysalo et al., 2007). Instead, the F-measure proposed by Hripcsak and Wilcox (2005) is used for this purpose. This measure allows computing pair-wise inter-annotator agreement using the standard precision, recall and the harmonic F-measure by treating one annotator as gold standard and the other as predictions. Table 5.1 shows the pair-wise agreement in F1 for each named entity type. A, B, C, D, E are identifiers of different annotators.

LOC							**TEM**						
	A	*B*	*C*	*D*	*E*	Avg		*A*	*B*	*C*	*D*	*E*	Avg
A	1	0.8	0.69	0.77	0.66	0.73	*A*	1	0.83	0.77	0.79	0.77	0.79
B	0.8	1	0.72	0.75	0.75	0.76	*B*	0.83	1	0.67	0.77	0.83	0.78
C	0.69	0.72	1	0.69	0.7	0.7	*C*	0.77	0.67	1	0.78	0.71	0.73
D	0.77	0.75	0.69	1	0.69	0.73	*D*	0.79	0.77	0.78	1	0.77	0.78
E	0.66	0.75	0.7	0.69	1	0.7	*E*	0.77	0.83	0.71	0.77	1	0.77

SUB						
	A	*B*	*C*	*D*	*E*	Avg
A	1	0.55	0.65	0.63	0.62	0.61
B	0.55	1	0.51	0.53	0.49	0.52
C	0.65	0.51	1	0.51	0.51	0.55
D	0.63	0.53	0.51	1	0.5	0.54
E	0.62	0.49	0.51	0.5	1	0.53

Table 5.1. Pair-wise inter-annotator-agreement in F1.

Comparing the figures, it is evident that even with reasonable effort from well-trained and skilled archaeology professionals devoted to developing annotation guidelines and resolving discrepancies in several iterations, the task of annotating named entities remained difficult and the level of discrepancy remained high. Annotating SUB is a much harder task than the other two types of named entity. This is expected because SUB spans across multiple specialised domains and terms are characterised by a high level of ambiguity and heterogeneity. Most discrepancies were due to identifying the boundaries of composite noun phrase entities, acronyms and identifiers (e.g., object codes or ID’s). Also for every type of named entities, there are always sub-groups of annotators that are more mutually consistent than with other annotators. This raised the issue of annotator suitability and the question that it is beneficial to eliminate inconsistent annotators from an annotation task to reduce discrepancies.

5.4.3 Phase 2 – Evaluating Machine Learning Accuracy

In order to gain a different view of the quality of the annotations produced in Phase one, two sets of experiments were conducted to evaluate how well an NER system can learn from these annotations. The first set of experiments used a corpus including annotations from all annotators to reflect the high level of discrepancy in the annotations. To do so, annotations produced by the five annotators were selected randomly, whilst ensuring the five documents are covered in full and roughly equal numbers of annotations were selected from each annotator. This corpus is referred to as **consolidated-sample-corpus**. The second set of experiments contained five sub-experiments, each using the annotated data created by an individual annotator. Thus there were five corpora for testing and they are referred as **individual-sample-corpus**. Theoretically, these corpora are free from inter-annotator discrepancy (but can still be limited by certain levels of intra-annotator agreement). On each of these six corpora, the SVM-based named entity tagger previously introduced in Chapter 4 was trained and evaluated in a *five-fold cross validation* experiment. All experiments have been carried out under consistent settings in order to fairly compare the effect of corpus quality. The following set of features are selected for this study as they are the most widely used features in the majority of NER research and prove to be effective in different domains (Byrne, 2007; Collier et al., 2000; Iria, 2009a; Nadeau, 2007a):

- The exact token string
- Orthographic type of the token
- Morphological root of a token (i.e., lemma)
- A context window of 5
- Domain specific gazetteers, including the MIDAS English archaeology period terms as the gazetteer for TEM, the Thesaurus of Monuments Types from English Heritage and the Thesaurus of Archaeology Objects from the STAR (STAR, 2007) project as gazetteers for SUB, and the UK Government list of administrative areas as the gazetteer for LOC.

Results of these experiments are shown in F1 in Table 5.2.

Corpus	SUB	TEM	LOC
Corpus annotated by A	0.73	0.78	**0.62**
Corpus annotated by B	**0.66**	0.78	0.65
Corpus annotated by C	0.76	0.74	0.69
Corpus annotated by D	0.78	0.84	0.7
Corpus annotated by E	0.79	**0.67**	0.75
Consolidated-sample-corpus	0.53	0.68	0.64

Table 5.2. F1 of NER learning accuracy obtained on the sample corpus used in phase 1.

Results of these experiments have shown interesting findings. Given no inter-annotator issues in each individually annotated corpus, one would expect higher levels of consistency and better annotation quality than the consolidated annotations, which translate to better machine learning accuracy. This was mostly true compared to results obtained on the consolidated-sample-corpus. However, exceptions were noticed for annotator A on LOC (0.02 lower), and E on TEM (0.01 lower). Also, comparing the figures across different named entity types, the named entity tagger had the lowest performance on LOC among four annotators (A, B, C, D), possibly indicating the lower quality of annotations and that it was the hardest task among all the three types. For the annotations created by each annotator, for person E the learning algorithm performed badly for TEM. The result in F-measure was even lower than that obtained from the consolidated-sample-corpus. However, on this named entity type other annotators produced fairly good annotations, as indicated by higher learning accuracies. This possibly indicates that E may find it more difficult at annotating TEM entities. Similar patterns were found for person B on SUB, and person A on LOC (figures in **bold**).

The results so far have revealed several conclusions that are useful for document annotation. Firstly, inter-annotator discrepancy has a major impact on the training data and therefore, machine learning accuracy. High level of discrepancy harms the quality of annotations, and decreases obtainable machine learning accuracy on a corpus. On the other hand, given uniform settings for a learning algorithm, different accuracies obtained from the same set of documents may indicate different levels of quality of the annotations. Secondly, annotators may have different skill levels in annotating different named entity types, possibly due to the difference in the focus of their knowledge. This has caused varying levels of inconsistencies in an annotator's annotations, depending on the specific named entity type. Therefore, there is the need for considering annotator's suitability for a task and isolating inconsistent annotators from a task. In line with the results from Table 5.1, these fostered the motivation of splitting an annotation task to sub-tasks depending on specific named entity types and selecting the most suitable annotators – as being mutually consistent – for each named entity type annotation sub-task.

5.4.4 Phase 3 – Annotator Selection

Annotator selection – In this stage, the document annotation task is split into three sub-tasks, each addressing the SUB, TEM, and LOC entities respectively. For each named entity type, the most suitable *three* annotators – best-fit-annotators – are identified and selected based on the analyses above. However, depending on the availability of annotators, the workload and inter-annotator consistency analysis, more or fewer annotators may be selected.

In the simplistic form, best-fit-annotators can be selected as those with the highest average inter-annotator agreement for each named entity type. These figures are shown in Table 5.3. However, as concluded from Table 5.2, certain annotators had high levels of inconsistency in annotating a particular type of named entities as shown by low machine learning accuracy (F-measure) tested on their annotations. This possibly suggests lack of knowledge in these annotators and therefore, their annotations can be of lower quality and an inadequate reference to others. As a result, it is important to exclude these annotators and their contributions from the calculation of inter-annotator agreement. Following this, for each named entity type, the annotations on which the learner obtained the lowest F-measure – particularly those below that obtained on the consolidated-sample-corpus – are eliminated. This caused person A eliminated from LOC, person B eliminated from SUB and person E eliminated from TEM. The average agreement is re-calculated and shown as the revised figures (the 'R' columns) in Table 5.3.

Annotator	**SUB**	***SUB-R***	**TEM**	***TEM-R***	**LOC**	***LOC-R***
A	0.61	*0.63*	0.79	*0.8*	0.73	-
B	0.52	-	0.78	*0.76*	0.76	*0.74*
C	0.55	*0.56*	0.73	*0.74*	0.7	*0.7*
D	0.54	*0.55*	0.78	*0.78*	0.73	*0.71*
E	0.53	*0.54*	0.77	-	0.7	*0.713*

Table 5.3. Average inter-annotator agreement in F1 for each NE type

Using the revised IAA figures, for each named entity type three annotators are selected as those with the highest average inter-annotator agreement scores. That is, persons A, C, D for SUB; persons A, D, B for TEM, and persons B, E, D for LOC, as shown in Table 5.4.

Annotator	SUB	TEM	LOC
A	O	O	
B		O	O
C	O		
D	O	O	O
E			O

Table 5.4. Selected best-fit-annotators for NE types

5.4.5 Phase 4 – Final Corpus Annotation

Next, the most suitable annotators identified for each entity type were asked to annotate the final corpus, which contains 25 full-length documents from the AHDS archive. There are two major differences between this annotation activity and that in the first phase:

- Annotators were only required to annotate entity types that they were most suitable for;
- The corpus was distributed evenly among annotators. No duplicate documents were used for further inter-annotator agreement analysis.

For each type of named entity, the documents were split into equal portions among its best-fit-annotators. For example, the 25 documents were split into three sets and each set was given to an annotator (A, C, or D) for annotating SUB. In the end, all annotations were merged into a single collection. This is based on the assumption that mutually consistent annotators will continue annotating consistently for the same annotation problem and the same type of corpus even without the process of consolidation and discrepancy resolution. Therefore, the workload can be distributed among the annotators for each particular entity type, while reasonable level of consistency can be expected.

5.5 Final Results and Discussion

The final annotation process (Phase 4) took roughly 10 to 15 person-days of work, although in practice it was spread across a couple of weeks to minimise tiredness and tedium to ensure annotators have the highest level of concentration during the work. To verify the quality of the annotations created in such a way, the final annotated corpus (**final-corpus**) was also used for a 5-fold cross validation experiment using the same settings as Phase 2. The results in F1 are shown in Table 5.5.

	SUB	TEM	LOC
Final-corpus	*0.68*	*0.83*	*0.71*
Consolidated-sample-corpus	0.53	0.68	0.64
Best result on individual-sample-corpus	0.79	0.84	0.75

Table 5.5. Learning accuracies on the final annotated corpus

Compared against results obtained on the consolidated-sample-corpus, the NER tagger obtained much better results on the final-corpus, which can be attributed to lower level of discrepancy in the annotations and therefore high quality of the annotations. Compared against the best results obtained on the individual-sample-corpora, which we consider the upper bound learning accuracy under no inter-annotator discrepancy, the NER tagger achieved very good results. The relatively smaller improvement on SUB is believed to be due to the heterogeneity of information included by the named entity type, which would have increased the difficulty of reaching agreement, as shown by the inter-annotator agreement studies before.

In terms of the effort spent on the annotation process, the method has significantly shortened the process required in the traditional document annotation approach. Phase 1 annotation process that follows the traditional approach was estimated to cost 14 person-days to annotate 5 documents; whereas, the Phase 4 annotation process following the annotator suitability theory was estimated to cost only 10-15 person-days to annotate 25 documents. In total, the annotation exercise undertook less than 1 person month, yet produced high quality annotations for machine learning purposes. These results are encouraging evidence of the applicability and technical soundness of the suitability-based annotator selection and document annotation approach, which can produce high quality annotations in a much more effective and efficient way.

Although the method was applied to the named entity annotation task, in theory, it can be generalised and applied to other document annotation tasks. Essentially, the key is to divide an annotation task into smaller components such that inter-annotator discrepancy can be sampled and each sub-task addressed separately with suitability analysis. For example, to adapt the method to event annotation for the Genia corpus, one can divide the task based on different event types, such as ‘binding’, ‘localisation’, and ‘positive regulation’. Given the high level of expertise required due to the complexity of the domain, it is likely that annotators may possess differing levels of knowledge about different event types. This can be revealed by sampling inter-annotator agreement at per-event-type basis, and identifying inconsistent learning accuracies obtained on the annotations created by individual annotators. The suitability of annotators can then be defined based on these findings and used to support further annotation activities. In document classification, the problem may be analysed based on the genre of documents (e.g., science, entertainment) since some annotators may be more familiar with certain kinds of topics than others, especially when they have different academic backgrounds. Similarly in sense disambiguation, the analysis may also be performed from the angle of the genre of documents (e.g., financial news report, sports news report), since different people may have different level of knowledge of certain areas, which will affect their ability to understand the content.

5.6 Conclusion

This chapter has addressed training data annotation for supervised NER. Training data are crucial resources to enable supervised machine learning methods. However, creating high-quality annotations is a difficult task due to inter-annotator discrepancies caused by differences in annotators' knowledge and experiences. Consequently, the annotation process typically requires significant amount of effort and time from multiple domain experts to work iteratively and collaboratively to identify and resolve discrepancies. The process is often expensive and time-consuming, creating a barrier for porting supervised learning methods to new tasks, especially in commercial and industrial environments.

To address this issue, an alternative approach to document annotation has been introduced. It is based on the idea of dividing an annotation task to smaller components, assessing annotator suitability, and annotator selection for sub-tasks. Illustrated using a real named entity annotation task, the method starts by dividing the annotation task by different entity types and then sampling the annotator discrepancy problem using the traditional document annotation process on a small corpus; the annotations are then used to evaluate machine learning accuracy to gain an insight to the annotator discrepancies in the task. Results of these experiments have shown that even with reasonable effort following the traditional annotation approach, high-level discrepancy may still remain, and can lead to low machine learning accuracy. Further analysis revealed that annotators may have different skill levels for annotating different types of named entities, suggesting the need for considering annotators' suitability in specialised annotation tasks. Using this information, the annotation sub-tasks are treated separately where only the most suitable candidates are required to tag the documents for specific named entity types. Furthermore, by matching best-fit-annotators to named entity types the workload can be distributed among the annotators since the intuition is that the best-fit-annotators are mutually consistent, and therefore, discrepancy can be irrelevant. This effectively reduces the workload per annotator, but increases the potential amount of annotations that can be produced whilst retaining high quality of annotations. Shown by the final results, the approach produced a final annotated corpus of five times of the size of the corpus created using the traditional approach (Phase one). The machine learning accuracy obtained on these annotations is better than that obtained from the annotations created in the traditional way, and is very close to the best result obtained under zero inter-annotator discrepancy using the individual-sample-corpora.

Several inadequacies will be further investigated in future research. First, intra-annotator agreement has been isolated from this study. It can be argued that the machine learning accuracy obtained on individual-sample-corpora in Phase two is partially attributed by intra-annotator discrepancy. This was not studied in this work. Studying intra-annotator agreement can reveal further useful details of annotators' capability in an annotation task, and evidence should be aggregated to make stronger support for annotator selection. Second, ideally the assessment of suitability should be and parameterised and the selection of suitable annotations should be formalised. This will be explored in the future. Furthermore, crowdsourcing (Wang et al., 2010) has become an interesting solution for document annotation in NLP research in recent years and it is known that quality control is a difficult issue because, for example, developing and enforcing annotation guidelines can be difficult. It is expected that the method introduced in this study can be adapted to help quality assurance in crowdsourcing-based annotation tasks. This will also be explored in the future work.

Part III – Gazetteer Generation

This part addresses the second research question concerning automatically building gazetteers for NER. Chapter 6 discusses automatic expansion of typed gazetteers; Chapter 7 discusses automatic generation of alternative, untyped gazetteers without the need of predefined seed gazetteers.

6 Typed Gazetteer Expansion

PREFACE

This chapter discusses automatically expanding existing gazetteers of pre-defined types. It is divided into six sections. Section 1 gives an introduction to the problem. Section 2 discusses related work. Section 3 discusses the hypothesis behind this work and Section 4 introduces a novel approach that automatically expands existing gazetteers based on knowledge in Wikipedia. Section 5 presents experiments, results and discussion. Section 6 concludes this chapter.

6.1 Introduction

Gazetteers, in the context of NER, can be either typed or untyped. Typed gazetteers contain reference entity names of pre-defined semantic types that are relevant to the task. For example, a person gazetteer may be used as background knowledge to support recognising person entities. This type of gazetteers is more frequently used in NER. Untyped gazetteers provide a simple way of grouping related terms without explicitly defining the type or categories of the groups, such as word clusters. This chapter discusses typed gazetteers, while the next chapter (Chapter 7) discusses alternative untyped gazetteers for NER. For brevity, the term 'gazetteer' refers to the 'typed' sense in the remainder of this chapter unless otherwise stated.

Gazetteers (both typed and untyped), are found to be particularly effective in improving the performance of NER systems when combined with other lexical-level features (Friedrich et al., 2006; Wang, 2006; Roberts et al., 2008; Saha et al., 2009). For example, in Mikheev et al. (1998), the use of gazetteers improved the accuracy of a supervised NER tagger by 39% in precision and 31% in recall. Particularly in technical domains, gazetteers or technical dictionaries are the major resource for resolving the complexity of domain-specific named entities (Roberts et al., 2008; Sasaki et al., 2008). Unfortunately, gazetteers are not always available or are often found to be incomplete, especially in technical domains. Even if gazetteers are already available, due to the evolutionary nature of human knowledge, terminologies and vocabularies are constantly changing, which requires frequent maintenance and update. Such task, if done manually, can be a laborious process and potentially very expensive (Kazama and Torisawa, 2008).

For these reasons, research has been carried out to develop methods of automatically generating or expanding typed gazetteers. In theory, the task of gazetteer generation or expansion shares the similar goal as NER, i.e., to recognise named entities of pre-defined types. While NER focuses on recognising and annotating each instance of named entities in texts, gazetteer construction ignores individual occurrences in the source text but focusing on creating lexical resources.

Traditional methods for gazetteer construction exploit lexical and syntactic patterns to extract named entities from unstructured corpora (Riloff and Jones, 1999; Thelan and Riloff, 2002). With increasing availability of semi-structured documents from the Web, methods have been proposed to harvest named entities from webpages by exploiting the structures in such documents (Ciravegna et al., 2004; Blanco et al., 2010). Recently, a new type of web-based resource has gained significant attention in NLP research. This is Wikipedia, a free online encyclopeadia that is created and maintained by collaborative effort. It is widely recognised that Wikipedia is a massive knowledge resource of named entities (Bunescu and Pasca, 2006). Its semi-structured nature enables easy access to vast amount of knowledge of named entities. It has been employed in a wide range of NLP tasks, such as document classification (Gabrilovich and Markovitch, 2006), Named Entity Disambiguation (Bunescu and Pasca, 2006), and semantic relatedness (Gabrilovich and Markovitch, 2007).

Using Wikipedia for NER or gazetteer construction is rarely studied. Its potential for such tasks has been unleashed in a number of studies, such as Toral and Munoz (2006), and Kazama and Torisawa (2007a). These methods are still limited in several ways. First, none have exploited the full content and structure of Wikipedia articles, but only focused on the article's first sentence. However, the full content and structure of Wikipedia carry rich information that can be potentially useful. Second, evaluation has been focused on the newswire domain and the four classic entity types defined in MUC6, i.e., location (LOC), person (PER), organisation (ORG) and miscellaneous (MISC). The usefulness of Wikipedia for technical domains has not been addressed. NER in technical domains is often much harder due to complexity of domain languages, density of information and specificity of classes (Nobata et al., 2000; Murphy et al., 2006; Byrne, 2007). As a result, gazetteers can play a more important role in domain specific NER.

This study proposes a new approach to automatically expand existing typed gazetteers using Wikipedia as an external knowledge resource. Unlike previous work, the method exploits various kinds of content and structural elements of Wikipedia, and does not rely on domain-specific knowledge. Briefly, given an existing seed gazetteer containing named entities that are described by Wikipedia articles, it firstly extracts hypernyms of the entities in the seed gazetteer using their Wikipedia article contents and structures. Next, related entities are identified as the links on these articles. If a related entity shares hypernyms with the seed gazetteer, they are added to the expanded set. The method is empirically tested in the Archaeology domain, where three existing gazetteers are automatically expanded following the proposed method. The resultant gazetteers are then used in an NER task, where the results have shown that they have contributed to further improvement in NER learning accuracy.

The rest of this chapter is structured as follows. Section 6.2 describes related work on automatic gazetteer generation; Section 6.3 discusses the hypothesis behind this work; Section 6.4 introduces the proposed methodology; Section 6.5 describes the experiment and evaluation, followed by the conclusion in Section 6.6.

6.2 Related Work

Methods for automatically generating or expanding typed gazetteers can be divided into three categories: **pattern driven approaches** that use unstructured corpora; **wrapper based approaches** that use structures of webpages; and **knowledge resource approaches** that use external knowledge resources, usually well-structured.

Pattern driven approaches uses lexical and syntactic patterns to extract entity names from unlabelled corpora. Such patterns are often domain- and language-specific patterns. A highly influential work of this type is Riloff and Jones (1999), which aims to build dictionaries of named entities using seed entities and unlabelled corpora. The method adopts an iterative learning strategy bootstrapped with a small amount of examples. Starting with a handful of seed entity names of a pre-defined type and an unlabelled corpus, the seed entity names are firstly located in the corpus. Then, lexical patterns are extracted for each occurrence based on its context to obtain a collection of patterns that can extract entities of the same type. Each pattern is then scored to promote patterns that extract a larger number of named entities and that often correlates to one particular semantic type. The pattern with the highest score is selected to be used for a new iteration of learning. Names that are extracted by the selected pattern are also submitted to a scoring function, which promotes names that are extracted by multiple patterns belonging to the same semantic type, and by patterns that have high scores. Finally, the five highest scored extractions are added to the seed gazetteer, and a new iteration of learning is repeated following the same pattern-extraction, name-extraction workflow. This gradually grows the seed gazetteer until certain arbitrary threshold is reached, for example, a given number of iterations. This method is later extended in Thelan and Riloff (2002), which permitted more extraction patterns to be learnt in each iteration and introduced a different scoring function for candidate entity names. This proved to be more effective than the earlier approach.

Talukdar et al. (2006) followed a similar approach to create gazetteers using seed entities and unlabelled corpora, but used different pattern induction and scoring methods. Given seed entities, they are searched and labelled in texts. The contexts of each occurrence are gathered for pattern induction. The pattern induction process begins with identifying from the contexts the so-called ‘trigger words’ that often indicate the presence of entities belonging to the same type. Trigger words are selected if they are frequently found in the contexts associated with the type of interest. Next, rather than using individual contexts as lexical patterns, an extraction pattern is induced as automata that summarises the most significant regularities of the contexts sharing a given trigger word. A pattern automaton represents the set of contexts that share the same trigger word as transitions that connect contextual words, the trigger word and also the named entity position. The pattern scoring method promotes patterns that extract more entity instances and penalises patterns that extract entities belonging to seed entities of other types (negative entities). Any patterns that extract negative entities, or whose scores are below a certain threshold are discarded. After the pattern filtering, newly extracted entities are scored based on the number of different patterns that extracts them. Eventually, the learning process also adopts an iterative nature, which gradually grows the seed gazetteers.

Pattern based approaches are generally effective and are often preferred in tasks involving large scale of data such as webpages (Etzioni et al., 2004; Freeman et al., 2011; Nakashole et al., 2011). A major limitation of this class of approaches is ‘error propagation’, that the performance rapidly declines as noisy patterns or entities are introduced in the bootstrapping process (Riloff and Jones, 1999; Ando, 2004). Also, low frequency entities can be problematic since there may be insufficient contextual information for pattern generalisation. Furthermore it has been criticised for weak domain adaptability and inadequate extensibility due to the specificity of derived patterns (Toral and Munoz, 2006; Kazama and Torisawa, 2008).

Wrapper based approaches exploit the structure of webpages. They are based on the idea that webpages often present similar information in similar structures. For example, a football league table will list instances of football teams; a yellow page website will list instances of companies. Therefore, if seed entities can be used to locate such webpages and structures, entities of the same type can be harvested by a wrapper program that processes the structured data and extracts information from similar structures. For example, Ciravegna et al. (2004) proposed to harvest person names from webpages based on seed entities. Firstly, webpages are crawled and those that contain mentions of seed entities are kept. All the occurrences of seed names are then annotated on the webpages and only those that contain a reasonable quantity of known names organised in structures such as lists and tables are selected to be further processed. Then, if a list or a table structure (e.g., column, row) contains multiple seed entities (e.g., at least four), a wrapper that can extract data from such structures is automatically induced, and applied to the similar structure to extract new entities. For example, if a list contains four seed entities that are already known person names, a wrapper is induced to extract all elements in the list and labels the extracted elements as person names.

Other wrapper based studies have been carried out (Blanco et al., 2010; Dalvi et al., 2011), generally based on the same principle but are distinguished by focusing on integration of knowledge extracted from different sources, or scoring and selecting induced wrappers in case of noisy annotations.

Knowledge resource approaches rely on the abundant information encoded in external knowledge resources and exploits domain-independent structures in such resources. Magnini et al. (2002) used WordNet as a gazetteer together with rules to extract named entities from texts. WordNet is a lexicalised ontology of words. It encodes word senses as synsets, which are indexed by their word forms and connected to other synsets by lexical and semantic relations. They suggested two ways that WordNet can be used as a gazetteer for this task. First, WordNet defines concepts that are hyponyms of an entity type. For example, the 'person' synset corresponding to the person named entity (as PER in MUC6) contains over six thousand hyponyms, among which words such as 'astronomer', and 'musician' can be used as trigger words to identify presence of the person entities in texts. Second, WordNet also defines instances of concepts, such as 'Galileo' and 'New York', which can be used directly as gazetteers. Based on these observations, they proposed to extract words that have the hyponymy relation with the desired named entity type from WordNet (i.e., by traversing the *IS-A* relation in WordNet), and then used simple heuristics to classify the words into 'trigger' words, and named entity instances (e.g., using capitalised word sequences). These are then used with a rule-based method to extract new named entities from texts. The main limitation of WordNet is lack of domain specific vocabulary, which is critical to domain specific applications.

Research in gazetteer construction and NER has also started to benefit from the successful lessons of using Wikipedia for NLP tasks. Toral and Munoz (2006) proposed to build gazetteers for location, person and organisation using Wikipedia. Given a Wikipedia article, they firstly extracted the noun phrases from the first sentence on the article page. The noun phrases are then mapped to WorldNet synsets. Next, starting from the mapped synset, the hypernymy relation is traversed until a higher level synset that satisfies one of the two conditions is found: 1) it represents the desired named entity type; 2) or it represents a sub-class concept of the desired named entity type (e.g., ‘country’ is considered a sub-class concept of the type location). If such a synset can be found, the title of the Wikipedia article is added as an instance of the named entity gazetteer.

Kazama and Torisawa (2007a) proposed to extract hypernymy labels from Wikipedia and use the labels as features for NER. Firstly, capitalised word sequences are extracted from a corpus, and then looked up in Wikipedia. If a Wikipedia article is found for a candidate, the first sentence of the article is processed to extract the hypernym of the concept or entity described by the article. This is done by extracting the head noun of the first noun phrase after *be* in the first sentence of the article. For example, ‘mammal’ is the extracted hypernym for the word ‘cat’, which has the first sentence as ‘The domestic cat *is* a small, usually furry, domesticated, carnivorous mammal’ in its Wikipedia article. The hypernyms are used as features for the search candidates, which are to be classified into predefined named entities. Empirically in an experiment, they mapped over 39,000 search candidates to approximately 1,200 hypernyms. Essentially, the process is equivalent to generating 1,200 gazetteers (labelled by the hypernyms) that include a total of 39,000 candidate entities. Although the hypernyms can be more specific than the required types of an NER task, the correspondence between them can be automatically learnt by a statistical classifier using training data.

While these earlier methods of gazetteer generation using Wikipedia have shown encouraging results, one major limitation is that they only make use of an article’s first sentence. Other content and structural elements of Wikipedia can also carry rich and potentially useful information, but have been ignored. Meanwhile, it is unclear whether the methods can be extensible to technical domains, where due to the complexity of domain languages and specificity of classes, the suitability of Wikipedia can be questioned.

6.3 Hypothesis

This work proposes an approach that automatically expands typed gazetteers by exploiting various content and structural elements in Wikipedia. It is based on the following hypothesis:

H2.1 Type-oriented gazetteer: Wikipedia can be used as a knowledge base of named entities. An existing gazetteer of predefined types can be automatically expanded using Wikipedia by defining gazetteer hypernyms using the structure

and content of Wikipedia, and extracting similar entities that share similar hypernyms with the seed gazetteer.

The first part of this hypothesis views Wikipedia as a knowledge base of named entities. As discussed before, this has been proposed by earlier studies and justified in various NLP tasks concerning named entities (Bunescu and Pasca, 2006; Gabrilovich and Markovitch, 2006; Toral and Munoz, 2006; Gabrilovich and Markovitch, 2007). Furthermore, a number of studies have been carried out to study Wikipedia's coverage of domain specific vocabulary.

Holloway et al. (2007) showed that by 2005, Wikipedia already contained 1,069 disconnected clusters of categories of articles each denoting a distinctive subject. Milne et al. (2006) studied Wikipedia's coverage of domain specific terminology in the domain of food and agriculture. Firstly, they made a direct comparison between Wikipedia and a manually created domain-specific thesaurus, and showed that approximately 50% of all terms in the thesaurus are included in Wikipedia. Further analysis showed that the majority of missing terms in Wikipedia are generally scientific terms and highly specific multi-word phrases. Next, they investigated how well Wikipedia provides thesaurus support for a domain-specific corpus by studying the coverage of terminology found in the corpus. Interestingly, it was found that many of the missed terms by Wikipedia are rarely used in the corpus; and as a result, the coverage of Wikipedia increased to over 70%. Overall Milne et al. concluded that Wikipedia can be used as a reliable terminology source for the food and agriculture domain.

Halavais (2008) compared the topical coverage distribution of Wikipedia against that of Bowkers Book in Print, which lists nearly all books that are currently available in English and in the United States from major publishers. A sample of 3,000 articles was drawn randomly from a 2006 English Wikipedia dump and articles with less than 30 words of text were discarded. These articles were manually classified by the Library of Congress (LC) category at the broadest level by two coders familiar with both Wikipedia and the LC system. The distribution of topics was then compared against that of the collection by Bowkers Books in Print. They found that the topical coverage of Wikipedia was generally good across all areas, although it seemed to be driven by the interests of its users. In particular, the sciences were well represented. However, it was not universally the case for every sub-area. For example, articles in medicine and law were particularly sparse. Other studies by Altmann (2005) and Clauson et al. (2008) also confirmed that Wikipedia's coverage of biomedical terminology is generally very limited when compared against specialist resources.

It is unsurprising that the usefulness of Wikipedia as a biomedical knowledge base is very limited. The biomedical domain is an area that has seen decades of development of lexical resources and benefited from a vigorous community constantly contributing to such resources. However, well-curated knowledge resources in other domains can be lacking and therefore, Wikipedia can still be a very useful resource for other domains that are not well-represented. On the other hand, to some extent, the exponential growth of Wikipedia may compensate towards its coverage limit.

The second part of the hypothesis states that given an initial gazetteer that contains named entities defined in Wikipedia, additional named entities of the same type can be identified based on the content defined for the entities and structural links with other Wikipedia resources. This requires: 1) that the Wikipedia articles must be linked in certain ways such that additional resources can be collected by following the links; 2) that the hypernyms of named entities can be labelled based on the content and structural elements in their corresponding Wikipedia articles, such that they can be matched. The first condition is easily satisfied since Wikipedia articles are intensively hyperlinked. In addition, a categorisation system is used to group articles under similar topics. The second condition can also be satisfied as it is justified by previous studies (Kazama and Torisawa, 2007a). Nevertheless, a different approach is explored in this work.

6.4 Gazetteer Expansion using Wikipedia Content and Structure

This section introduces the proposed method of automatic gazetteer expansion. Given an existing gazetteer containing named entities of a predefined type, the named entities are searched in Wikipedia and the articles describing the entities are retrieved. Next, the named entities are labelled by hypernymy terms that are extracted from their Wikipedia articles. These hypernymy terms are often more fine-grained class labels than the desired entity type. To contrast, the pre-defined entity type is named **Coarse-Grained Class (CGC)** labels and the extracted hypernyms are named **Fine-Grained-Class (FGC)** labels. Next, candidate named entities are identified as the links found on the articles of the initial seed named entities. Finally, to decide whether a candidate named entity belongs to the pre-defined type (i.e., CGC), it is also labelled by its FGCs. Its FGCs are then compared with the pool of FGCs extracted for the initial gazetteer to decide whether the candidate entity qualifies for the same type. This process is divided into three steps: **the**

matching step, the classification step, and **the expansion step**. The pseudo-algorithm is illustrated in Figure 6.1.

```
Input: initial gazetteer of named entities SE of a predefined type (CGC) C
Output: new entities NE of type C
STEP 1  - matching
  a. Initialise Set A to contain articles for SE;
  b. For each entity e: SE
  c.     Retrieve article a from Wikipedia for e;
  d.     Add a to A;
STEP 2  - classification
  a. Initialise Set L
  b. For each a: A
  c.     Extract fine grained class labels (FGC) l;
  d.     Add l to L;
  e. Filter L
STEP 3 –expansion
  a. Initialise Set HL;
  b. For each a: A
  c.     Add hyperlinks from a to HL;
  d. (optional) recursively crawl extracted hyperlinks and repeat b and c
  e. For each link hl: HL
  f.     Extract fine grained class labels (FGC) l';
  g.     If match_function(l', L) = true
  h.         Add title of hl to NE;
  i.         Add titles of redirect links of hl as entity names to NE;
```

Figure 6.1. Pseudo algorithm for gazetteer expansion using Wikipedia

6.4.1 The Matching Step

In the matching step, a given named entity is searched in Wikipedia to obtain the corresponding article describing the entity. Three types of outcomes can be expected. First, Wikipedia returns a single article page for terms that are unambiguous, or those of which a most commonly used sense is available. For example, the phrase ‘natural language processing’ has a unique article page in Wikipedia; the word ‘cat’ is given the article page that describes the most widely use sense of ‘a kind of domesticated animal’ rather than anything else that can also be referred by the same word. In this case, the single article page is selected for the named entity.

Second, some terms will not point to any articles in Wikipedia. In this case, the ‘leftmost longest match’ rule is applied to fuzzily match the entity to the closest Wikipedia article. For example, for the phrase ‘Stone Age flint arrowhead’, the entire phrase returns no articles. Therefore, it is reduced to ‘Age flint arrowhead’, ‘flint arrowhead’ and ‘arrowhead’ and searched in turn in Wikipedia until a match is found. The intuition is to match the named entity to the closest concept that is likely to be the hypernym of the entity.

Third, for polysemous terms and names that can be used to refer to different concepts and entities, Wikipedia uses 'disambiguation' pages as directory lists for such articles. A disambiguation page lists different meanings with links to corresponding article pages. For example, the search for 'George Bush' returns a disambiguation page that lists all named entities that are referenced by this name. In this case, the named entity is discarded and not used for the following steps.

Using 'Sheffield' as a running example, it is matched to a single article 'http://en.wikipedia.org/wiki/Sheffield' in this step as it is defined as the most commonly used sense by Wikipedia.

6.4.2 The Classification Step

Once Wikipedia articles are retrieved for all seed entities, the entities are labelled by their FGCs based on the article content. There are two types of information from Wikipedia that can be used as reliable labels (Step 2, a – d in Figure 6.1). The first is based on the study by Kazama and Torisawa (2007a), who observed that the first sentence of an article is often a definitive sentence. Specifically, the head noun of the noun phrase just after *be* is most likely the hypernym of the entity of interest.

There are two issues in this approach. First, the head noun may be too generic to represent a domain-specific class. For example, following their approach the FGC extracted for the archaeological term 'Post-Classic Stage' from the sentence 'The Post-Classic Stage is an archaeological term describing a particular developmental level*'* is 'term', which is the head noun of 'archaeological term'. Clearly in such a case the phrase is more domain-specific than the head noun. For this reason, the *first noun phrase* after *be* is used as FGC instead of the head noun. Second, their method ignores a correlative conjunction that often indicates equally useful FGCs. For example, the two noun phrases in italics in the sentence 'Leeds is a *city* and *metropolitan borough* in West Yorkshire, England' are equally useful FGCs for the article 'Leeds'. For this reason, we also extract the noun phrase that is connected by a correlative conjunction as the FGC. For brevity, this method of classification is referred to as **FirstSentenceLabeling**, and the FGCs extracted are referred to as FGC_s.

Therefore, using the previous example, the first sentence ‘Sheffield is a city and metropolitan borough in South Yorkshire, England’ is extracted from the Wikipedia article, and ‘city’ and ‘metropolitan borough’ are extracted as FGCs for ‘Sheffield’ and its corresponding gazetteer.

The second method for extracting FGC is based on the Wikipedia category structure. Wikipedia articles are labelled by one or multiple categories, which are generalised concepts organised in a hierarchical structure, creating a category tree generally resembling the broader and narrower sense of relation between categories. Similar articles are grouped by same category labels. Although the hierarchy does not define strict taxonomic relations between categories, research (Strube and Ponzetto, 2006; Zesch and Gurevych, 2010a) has shown that it can be used as an approximate taxonomy in many tasks. Therefore, category labels of articles are extracted, filtered and selected as FGCs of entities. This approach is named as **CategoryLabeling**, and the extracted FGCs are denoted by FGC_c. Following this approach, the category labels extracted from the ‘Sheffield’ article include: ‘Populated places established in the 1st millennium’, ‘Cities in Yorkshire and the Humber’, ‘Local government districts in South Yorkshire’, ‘Metropolitan boroughs’, ‘Sheffield’, ‘Local government districts of Yorkshire and the Humber’.

There are three situations in which the extracted FGCs must be revised (Step 2, e ‘Filter ***L***’ in Figure 6.1). Firstly, some articles have a category with the same name as the article title. In the above example, the article of ‘Sheffield’ has a category also named as ‘Sheffield’. In this case, the category tree is traversed to the next level up to extract categories of the category ‘Cities in Yorkshire and the Humber’, ‘Metropolitan boroughs’, ‘Local government districts in South Yorkshire’, ‘Districts of England’. Secondly, for management purposes, arbitrary categories have been created by Wikipedia moderators to group and organise articles. Examples include ‘Articles to be Merged since 2008’, ‘Wikipedia Templates’, etc. Such categories do not carry useful semantics, and can introduce noisy labels. Therefore, a stopwords list is manually created to filter out such categories. The full list is shown in Figure 6.2[4]. Any categories that contain a word in Figure 6.2 are ignored. Thirdly, to further filter out noisy labels, only FGCs that are extracted for at least 2 seed entities are kept.

[4] These stopwords are used for a version of Wikipedia dated 6 Feb 2007. It is known that later versions of Wikipedia have introduced more category labels for management purposes and therefore, additional stopwords may be needed.

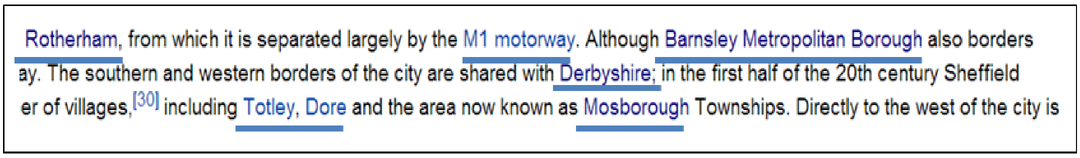
cleanup, articles, pages, disambiguation, infobox, Wikipedia, Wiktionary, Wiki, underpopulated, disputes from, accuracy disputes, categories, classification, uncategorized, wikify

Figure 6.2. A list of stopwords used to filter out noisy category labels

The classification process generates a pool of FGCs which are hypernyms of input named entities and potentially hyponyms of pre-defined CGCs. In the next step, they are used as a control vocabulary to guide the expansion of similar named entities.

6.4.3 The Expansion Step

Next, expanding the gazetteer involves identifying from Wikipedia the candidate entities that are related to the input named entities. This is done by following the hyperlinks from the full content of articles retrieved for the input named entities (Step 3, a – c in Figure 6.1). The hyperlinks connect the main article of an entity (source entity) to other sets of entities (related entities). Therefore, by following these links a large set of related entities to the initial gazetteer can be reached. These are considered candidate entities for selection. For example, Figure 6.3 shows a screenshot of the Wikipedia article for 'Sheffield' with a number of links highlighted (as underline).



Rotherham, from which it is separated largely by the M1 motorway. Although Barnsley Metropolitan Borough also borders
ay. The southern and western borders of the city are shared with Derbyshire; in the first half of the 20th century Sheffield
er of villages,[30] including Totley, Dore and the area now known as Mosborough Townships. Directly to the west of the city is

Figure 6.3. Linked articles that are relevant to 'Sheffield'

Furthermore, the hyperlinks can be recursively followed to retrieve more candidate entities and Wikipedia articles if necessary (Step 3, d in Figure 6.1), e.g., when the initial gazetteer is very small and very few Wikipedia articles can be found for the initial named entities.

Next, the two classification approaches introduced in the previous section are used to identify the FGCs of candidate entities (Step 3, e, f in Figure 6.1). For example, the link 'Barnsley Metropolitan Borough' is followed to retrieve a candidate entity's article. Then, based on *FirstSentenceLabeling* the FGCs are extracted as 'metropolitan borough'; based on *CategoryLabeling* the FGCs are extracted as 'Politics of Barnsley', 'Local government districts in South Yorkshire', 'Metropolitan boroughs', and 'Local government districts of Yorkshire and the Humber'. These extracted FGCs are matched against the pool of FGCs extracted for the initial gazetteer using a match function; if a match condition is satisfied, then the candidate entities – the title of the corresponding Wikipedia article – are accepted to extend the gazetteer (Step 3, g, h in Figure 6.1). In this study, the match function simply checks if the FGCs of a candidate entity are included by those of the initial gazetteer extracted using the same classification method. That is, if the FGCs of the initial gazetteer are built by *FirstSentenceLabeling*, only the candidate entity's FGCs labelled by the *FirstSentenceLabeling* approach are used for matching. Thus following the previous example, the FGCs extracted by *FirstSentenceLabeling* for 'Barnsley Metropolitan Borough' – 'metropolitan borough' – is checked against those extracted for 'Sheffield' by the same classification method, and is found to be shared by the two entities. As a result, it is considered a valid match and 'Barnsley Metropolitan Borough' is accepted into the gazetteer. The same procedure applies to *CategoryLabeling* for gazetteer expansion. The intuition is that if a candidate entity shares a hypernym with a source entity, then any higher level hypernyms – and eventually the desired entity type – of the source entity should also apply to the candidate.

In addition, for each qualifying Wikipedia article accepted into the gazetteer, the associated 'redirection' titles are also selected as entity names (Step 3, i in Figure 6.1). Redirection titles for a Wikipedia article are usually name aliases for the same entity or concept. In Wikipedia, all redirection titles point to the same article page. To further eliminate potentially ambiguous entities, for each extended gazetteer, we exclude entities that are found in domain-independent gazetteers. For example, we use a generic person name gazetteer to exclude ambiguous names from the extended gazetteers for LOC (Location).

After applying these processes, the initial gazetteer is expanded by entities with which they share the same FGCs. The method can be repeated for a number of iterations, in which the newly added entities serve as seed entities and go through the three stages again. Depending on the size of seed entities and the desired scale of the output, one can customise the number of runs to build various sizes of gazetteers.

6.5 Evaluation and Discussion

The proposed method is evaluated in an NER task in the archaeology domain. As discussed before, gazetteer generation is rarely addressed in domain-specific contexts, particularly scientific domains. Existing studies have predominantly evaluated their methods in the newswire domain.

In brief, the proposed method is firstly applied to extend three domain-specific gazetteers, which are referred to as initial gazetteers. Then, both the initial gazetteers and expanded gazetteers are used in an NER experiment on archaeology data. The results are compared. The corpus described previously in Chapter 5 is used in this experiment. The dataset is then split into five equal parts for five-fold cross-validation experiments. The SVM-based NER tagger used in Chapter 4 is reused with the same setting. The baseline features for the classifier are:

- The exact token string
- Orthographic type of the token
- Morphological root of a token (i.e., lemma)
- A context window of 5 (i.e., five tokens before and five tokens after the current token)

To access Wikipedia content, the JWPL (Java-based Wikipedia Library) library (Zesch et al., 2008a) is used. A version of Wikipedia dated 6 Feb 2007 is parsed by this library and accessed locally.

The accuracies (Precision – P, Recall – R, F1) obtained with the baseline are shown in Table 6.1.

Next, the same domain specific gazetteers previously used in Chapter 5 are used as additional features to the baseline. To re-cap, these include the MIDAS English archaeology period terms as the gazetteer for TEM, the Thesaurus of Monuments Types from English Heritage and the Thesaurus of Archaeology Objects from the STAR (STAR, 2007) project as gazetteers for SUB, and the UK Government list of administrative areas as the gazetteer for LOC. These will be referred to as *GAZ$_{init}$*. The learning accuracies with the added gazetteer features are shown in Table 6.2.

	LOC			**SUB**			**TEM**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline (B)	*69.4*	*67.4*	*68.4*	*69.6*	*62.3*	*65.7*	*82.3*	*81.4*	*81.8*

Table 6.1. Baseline learning accuracy

	LOC			**SUB**			**TEM**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
B+GAZ$_{init}$	*69.0*	*72.1*	*70.5*	*69.7*	*65.4*	*67.5*	*82.3*	*82.7*	*82.5*

Table 6.2. Learning accuracies by the baseline with the initial gazetteers

The initial gazetteers are then expanded using the proposed method. Since two separate methods are introduced for the classification stage and used separately for labelling candidate entities, they are applied separately and compared. Specifically for each entity type, *GAZ*$_{exp\_firstsent}$ denotes an expanded gazetteer built using *FirstSentenceLabeling* for classifying initial gazetteer entities and candidate entities; *GAZ*$_{exp\_category}$ refers to an expanded gazetteer built with *CategoryLabeling*. Table 6.3 shows statistics of the gazetteer expansion results. Table 6.4 shows the most frequently extracted FGCs for each gazetteer by each classification method.

	Number of unique entries in gazetteers		
	LOC	**SUB**	**TEM**
GAZ$_{\text{init}}$	11,786, 8,228 found in Wikipedia	5,725, 4,320 found in Wikipedia	61, 43 found in Wikipedia
GAZ$_{\text{exp\_firstsent}}$	19,385, 7,599 new to GAZ$_{\text{init}}$	11,182, 5,457 new to GAZ$_{\text{init}}$	163, 102 new to GAZ$_{\text{init}}$
GAZ$_{\text{exp\_category}}$	18,861, 7,075 new to GAZ$_{\text{init}}$	13,480, 7,745 new to GAZ$_{\text{init}}$	305, 245 new to GAZ$_{\text{init}}$

Table 6.3. Number of unique entities in each gazetteer

The expanded gazetteers then replace the initial gazetteers, and are used for NER. Results are shown in Table 6.5.

The results so far have shown that, despite the large sizes of the initial gazetteers, they are still incomplete and can be further expanded. The expansion process significantly increased the amount of domain-specific entities as indicated by the numbers in Table 6.3. Careful analyses have shown that there are gaps between the annotations and initial gazetteers. For the LOC gazetteer, many street names (‘Blue Stone Heath Road’, ‘A61’), place of interests (‘Royal Armory Museum’, ‘Abbey Village Reservoir’) and alternative names are used in the corpus; however, these are largely missing in the initial LOC gazetteer, which only contains UK administrative areas. Similarly for TEM, many alternative and new names are found in annotations but not included in the gazetteer. Examples include ‘renaissance’, ‘Roman Republic’, ‘Byzantine Empire’. The problem is even more acute for SUB due to the heterogeneity of information in this class. The initial gazetteers were initially divided into 44 sub-topics, which is equivalent to an average of roughly 130 entities per topic. The gazetteer expansion process successfully doubled the size of SUB gazetteers. The quality of the generated gazetteers is considered to be good since they improved the performance of the baseline with the initial gazetteers by 1 – 3 points in F1.

LOC	
FirstSentenceLabeling	***CategoryLabeling***
village,	villages in north Yorkshire,
small village,	north Yorkshire geography stubs,
place,	villages in Norfolk,
town,	villages in Somerset,
civil parish	English market towns
SUB	
FirstSentenceLabeling	***CategoryLabeling***
facility,	ship types,
building,	monument types,
ship,	gardening,
tool,	fortification,
device	architecture stubs
TEM	
FirstSentenceLabeling	***CategoryLabeling***
period,	Periods and stages in archaeology,
archaeological period,	Bronze age,
era,	middle ages,
century,	historical eras,
historical era	centuries

Table 6.4. Top 5 most frequently extracted FGCs by each classification method

	LOC			**SUB**			**TEM**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
B+ GAZ$_{exp\_firstsent}$	69.9	76.7	73.1	70.0	68.3	69.1	82.6	84.6	83.6
B+ GAZ$_{exp\_category}$	69.1	75.1	72.0	68.8	67.0	67.9	82.0	83.7	82.8

Table 6.5. Learning accuracies with the expanded gazetteers.

Furthermore, the effects of combining the two classification methods for initial and related candidate entities are studied. Two additional sets of gazetteers were created and tested with the NER tagger. Firslty, *GAZ$_{exp\_union}$* merges gazetteers built using two different approaches; secondly, *GAZ$_{exp\_intersect}$* takes the intersection of *GAZ$_{exp\_firstsent}$* and *GAZ$_{exp\_category}$* i.e., only entities that are generated by both approaches. The sizes of the two new gazetteers are shown in Table 6.6. The NER performance using these gazetteers is shown in Table 6.7.

	LOC	**SUB**	**TEM**
GAZ$_{exp\_union}$	23,741 11,955 new to GAZ$_{init}$	16,697 10,972 new to GAZ$_{init}$	333, 272 new to GAZ$_{init}$
GAZ$_{exp\_intersect}$	14,022, 2,236 new to GAZ$_{init}$	7,455, 1,730 new to GAZ$_{init}$	133, 72 new to GAZ$_{init}$

Table 6.6. Number of unique entities in each gazetteer built by combining the two approaches

	Location			**Subject**			**Temporal**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
B+GAZ$_{exp\_union}$	68.9	75.0	71.8	69.8	66.5	68.1	82.4	83.4	82.9
B+GAZ$_{exp\_intersect}$	69.3	76.2	72.6	69.7	67.6	68.6	82.6	84.3	83.4

Table 6.7. Learning accuracies with *GAZ$_{exp\_union}$* and *GAZ$_{exp\_intersect.}$*

Results have shown that taking the intersection of gazetteers generated by the two approaches outperformed the union, but figures are still lower than the best results obtained with *GAZ*$_{exp\_firstsent}$ as shown in Table 6.5. Also, learning accuracies obtained with *GAZ*$_{exp\_category}$ are lower than with *GAZ*$_{exp\_firstsent}$. These observations suggest the quality of gazetteers generated using *CategoryLabeling* is lower than those by *FirstSentenceLabeling*, therefore, merging the gazetteers included noisy entities from the low-quality gazetteer, while intersecting the gazetteers excluded valid entities from the high-quality gazetteer. Analysing examples of the FGCs extracted by the two methods showed that this could be due to two reasons. First, the loose structure of the Wikipedia category graph does not always follow the *IS-A* relationship. Although several heuristics have been introduced to reduce noise, the FGCs extracted by this method are still noisier than those built by *FirstSentenceLabeling*. Such examples include 'Bronze' for TEM, and 'Units of force' for LOC. These noisy FGCs accepted invalid entries in the gazetteers. On the other hand, compared to Wikipedia categories, the FGCs extracted from the first sentences are sometimes very fine-grained and restrictive. For example, the FGCs extracted for 'Buckinghamshire' from the first sentence are 'ceremonial Home County' and 'Non-metropolitan County', both of which are UK-specific Location concepts. These fine-grained FGCs are believed to help control the gazetteer expansion to focus on the domain of interest. The better performance with *FirstSentenceLabeling* suggests that this has played a positive role in improving the quality of candidate entities.

6.6 Conclusion

This chapter has addressed methods for expanding existing gazetteers of pre-defined types. Gazetteer is a type of background knowledge that is important in NER. However, it is not often available and manually creating gazetteers is a time consuming and costly process. To address this issue, research has been carried out to develop methods for automatically generating or expanding existing gazetteers of pre-defined types. The majority of these methods use lexical and syntactic patterns to identify named entities from unstructured corpus. Many have exploited the regularities of webpages and developed wrapper based methods that extract named entities from webpage structures such as tables and lists. Recently, several studies are made to exploit the structure and content in Wikipedia to create named entity gazetteers for NER. Compared to other types of methods, Wikipedia provides the advantage of easier access to richly structured information, good coverage of named entities and specialised terminology, and reasonable coverage for many technical domains. The exponential growth of Wikipedia knowledge base ensures promising prospects for methods built on top of it. Unfortunately, existing studies are inadequate in several ways. Firstly, they only make use of very limited content and structures of Wikipedia; secondly, they have only addressed the newswire domain. The applicability of Wikipedia in domain specific tasks has not been tested.

In this study, a new method of gazetteer expansion has been proposed to address these issues. Given an initial gazetteer of a pre-defined type, the method automatically expands the gazetteer by exploiting various content and structural elements in Wikipedia. The method is domain-independent, only relying on the generic structures of Wikipedia. Empirically tested in an NER task concerning three domain-specific entity types in the archaeology domain, the method has doubled the sizes of initial gazetteers with additional entities of the same type harvested from Wikipedia. The extended gazetteers have also further improved learning accuracies in an NER task.

Several questions remain to be answered in the future research. Firstly, the method is evaluated indirectly by an NER application. Alternatively, the expanded gazetteers could be manually inspected to assess its quality. This will be carried out in the future work.

Secondly, the method will be revised to improve its scalability. As discussed before, the method is designed to be scalable, in the way that it can be repeated in iterations to generate various sizes of gazetteers. Theoretically, it can also be applied with much smaller initial seed gazetteers. However, these have not been empirically tested. With much smaller seed gazetteers (e.g., gazetteers with less than 50 elements), the classification stage may have to be revised to relax the granularity of the extracted FGCs (e.g., ‘cities’ instead of ‘cities of the Yorkshire county’) in order to bootstrap iterative gazetteer generation. Additionally, certain noise control strategies may be necessary as the number of iterations grows. These will be explored in the future.

Further, the current method explores only the named entities that have a dedicated article page in Wikipedia. However, a vast amount of named entities exist in the articles of Wikipedia but they do not have a dedicated page. For example, ‘Dell Latitude D600’ is a named entity that does not have a dedicated Wikipedia page but is mentioned on the page of ‘Dell Latitude’. It may be beneficial to capture and include these entities in the learning process.

Last but not least, the long term goal will be exploring methods that are based on a combination of online resources, including generally structured webpages, Wikipedia and other web resources, such as the Open Directory Project (ODP[5]), a large directory of named entities. It is impossible to create an ultimate, complete knowledge base; however, different knowledge sources may complement or re-enforce each other. Therefore, gazetteer generation may benefit from collective evidence based on a combination of resources. One interesting research that may help towards this goal is the DBpedia project (Bizer et al., 2009), which interlinks knowledge from different sources and publishes them through a uniform access protocol. It is a free online multi-million triple store that links concepts and entities by semantic relations. For example, as by 29th Feb 2012, DBpedia includes triples that describe 416,000 persons, 526,000 places, 106,000 music albums, 60,000 films, 17,500 video games, 169,000 organisations, 183,000 species and 5,400 diseases. Therefore, DBpedia can be a potentially powerful knowledge resource for gazetteer construction. Methods based on DBpedia will be explored in the future.

[5] ODP (Dmoz), http://www.dmoz.org/, last retrieved on 29 Feb 2012

7 Alternative Gazetteer Generation

PREFACE

This chapter discusses automatically generating alternative gazetteers for NER. Alternative gazetteer is a concept relative to traditional and typed gazetteers discussed in the previous chapter. Alternative gazetteers simply provide a way of grouping related terms without explicitly defining the type or category of the groups. This chapter is divided into seven sections. Section 1 gives an introduction to the problem. Section 2 discusses related work. Section 3 details the hypothesis behind this work and Section 4 introduces an approach that exploits word topicality for alternative gazetteer generation. Section 5 presents experiments and results. Section 6 discusses results, and presents an in-depth analysis of lessons learnt. Section 7 concludes this chapter.

7.1 Introduction

Alternative gazetteers provide a simple way of grouping related terms without explicitly defining the type or categories of the groups as the typed gazetteers do. Thus a typed gazetteer of companies may say that 'Microsoft' and 'Google' are both company entities; an alternative gazetteer simply says that the two terms are always found belonging to the same group, regardless of what the group is called (e.g., 'American companies', 'IT companies', or even 'unknown group A'). They are useful to NER because from the learning point of view, a gazetteer will be useful as long as it returns consistent labels for the same types of named entities, since the correspondence between the labels and the named entity types can be learnt automatically from training data (Kazama and Torisawa, 2008). An advantage over typed gazetteers is that it eliminates the need of labelling gazetteers or their entries, which encourages unsupervised methods to be adopted for gazetteer generation. In fact, the majority of studies that address automatically generating alternative gazetteers have adopted unsupervised approaches that require no manually provided input.

Several forms of alternative gazetteers have been explored for NER. One commonly adopted technique is using word clusters, where words are clustered based on their distributional similarity and the derived clusters are used as a handful of gazetteers for NER (Freitag, 2004; Miller et al., 2004; Jiang et al., 2006; Kazama and Torisawa, 2008; Saha et al., 2009; Finkel and Manning, 2009; Chrupała and Klakow, 2010). Some studies have proposed to automatically extract hypernyms of terms using external resources and group terms by their hypernyms, which effectively creates a set of gazetteers that can be used for NER (Kazama and Torisawa, 2007a; Kazama and Torisawa, 2008).

This chapter explores a new dimension of building alternative gazetteers for NER. It builds on the hypothetical relationship between the **topicality of words** and named entities. Topicality of a word refers to the degree to which a word represents a document's topic. Although a formal definition of 'topic' is not available, it can be considered as several key terms that summarise the 'aboutness' of a document. The relationship between document topics and named entities was initially introduced by a number of studies, for example, Clifton et al. (1999) showed that named entities are highly relevant to the topic of a document. Hassel (2003) argued that named entities are often among the most information dense tokens of the text and largely define the domain of interest. Rennie and Jakkola (2005) and Gupta and Bhattacharyya (2010) proposed to measure topicality of words in terms of '**informativeness'**, which is then transformed to some features for named entity detection. Generally, informativeness is quantified based on a word's distribution over a collection of documents: it is generally agreed that informative words often demonstrate a 'peaked' frequency distribution over a collection of documents, such that the majority of their occurrences in the collection are found in a handful of documents (Church and Gale, 1995b). In practice, most informativeness measures (Spark Jones, 1973; Harter, 1975; Church and Gale, 1995a; Church and Gale, 1995b; Rennie and Jaakkola, 2005) have employed two distributional properties of words: document frequency and term frequency in the corpus.

However, informativeness may not always represent topicality for two reasons. First, document topics can vary largely even if they belong to the same domain. For example, articles in the Genia corpus are from scientific journals, each discussing a finely constrained subject that is related to biomedical science but individually distinctive. This is reflected by largely varying vocabularies as well as varying frequency patterns of words at individual document basis. Informativeness however, studies the *global* distribution of words and ignores such varying patterns specific to documents. As it will be discussed later in this chapter, many informativeness measures are biased by document frequency and can promote words that are irrelevant to topics of individual documents, or miss those that are in fact relevant. Second, informativeness scores are globally uniform and specific to a collection. As a result, ambiguous words that carry different senses in different document contexts can be mis-interpreted. In the biomedical domain, a fair amount of named entities can contain common English words. The word 'bright', 'white' and 'cycle' can be used to refer to protein or gene names in some documents but also widely used as common words carrying no special senses in most documents (Morgan and Hirschman, 2003). A uniform informativeness score cannot distinguish these cases.

For these reasons, topicality should be measured locally and specific to individual documents, taking into account a word's distributional patterns at document levels. Following this hypothesis, this study proposes to measure the topicality of a word in terms of its relevance to a document – a widely adopted notion for Information Retrieval. Next, words can be grouped based on their level of topicality and the intuition is that those falling under the highly topic-oriented groups can be useful features to named entities in that document; while those belonging to non-topic-oriented groups can be negative features. This has motivated the idea of using such document-specific groupings of words as gazetteer (or non-gazetteer) features based on the topicality of words. The proposed method begins with evaluating topicality of words using four simple relevance functions: Term Frequency (TF) in documents, Term Frequency-Inverse Document Frequency (TFIDF), weirdness (WD, Ahmed et al., 1999), and one that combines both TFIDF and WD. Next, words are ranked based on their topicality scores and a simple equal interval binning technique is applied to segment the list into a handful of sections, which effectively creates a handful of document-specific gazetteers. These are then used for a statistical NER model, which, when thoroughly evaluated using five datasets from three domains, consistently improves a baseline by between 0.9 and 3.9 points of F-measure and always outperforms methods based on informativeness. This confirms that locally measured topicality is an effective feature for generating alternative gazetteers for NER and is generalisable across domains.

The remainder of this chapter is organised as the follows. Section 7.2 presents related work. Section 7.3 further discusses the hypothesis. Section 7.4 introduces the method in details. Section 7.5 presents the experiments and results. Section 7.6 analyses the results and discusses the lessons learnt and Section 7.7 concludes this chapter.

7.2 Related Work

Methods using word clusters – A common approach to building alternative gazetteers for NER is using word clusters. Freitag (2004) showed that word clusters derived based on distributional similarity tend to have a useful semantic dimension. For example, clustering words extracted from a sample news corpus yielded two clusters that clearly correspond to first names and last names. To exploit this nature they derived word clusters as gazetteers from an external corpus similar to that used in an NER task. Specifically, words were firstly extracted from a corpus of hundreds of thousands of news articles. These were then clustered based on the similarity of their context into 200 clusters. The clusters were treated as unlabelled gazetteers to support NER from the MUC6 datasets. Miller et al. (2004) applied hierarchical clustering, which generates a binary tree that attaches each word as leaf nodes. As a result, nodes higher in the tree correspond to larger word clusters, while lower nodes correspond to smaller clusters. Each word is then assigned a binary string by following the traversal path from the root to its leaf. The strings, indicating the cluster membership of words, are then used as features for NER. This type of methods has gained substantial popularity and is adopted by a number of later studies. Jiang et al. (2006) applied a similar hierarchical clustering approach to that of Miller et al. (2004) and used derived clusters as gazetteers for Chinese NER. Kazama and Torisawa

(2008) addressed the issue of computational complexity when deriving clusters from very large corpora (i.e., millions of documents) and used word clusters for Japanese NER. They also modelled similarity based on syntactic features rather than contextual features. Finkel and Manning (2009) used word clusters in nested named entity recognition, i.e., identifying named entity mentions that are constituents of longer entity names. Saha et al. (2009) tested different methods of computing similarity and studied their effect on word cluster gazetteers in biomedical NER. Chrupala and Klakow (2010) clustered words based on co-occurrence statistics in a large corpus and used word clusters as gazetteers for German NER.

The major limitation of word cluster based approaches is selecting an appropriate level of granularity. Too many clusters provide insufficient generalisation; while too few clusters provide insufficient discrimination (Miller et al., 2004). An optimum level is often empirically derived depending on the data, which can involve extensive experimentation. For example, the studies described above have all reported different settings of cluster numbers (i.e., number of gazetteers), while Kazama and Torisawa (2008) tested several different settings. Meanwhile, the clustering process can be computationally extensive and adds considerable cost to the NER task. The effectiveness of clustering may also depend on the choice of the similarity function.

Methods using hypernymy/hyponymy relations – A recent study by Kazama and Torisawa (2007a) proposed using automatically learned hypernymy relations as alternative gazetteers for NER. This has been discussed previously in Chapter 6. To re-cap, they firstly extracted candidate phrases containing *n* tokens with at least one capitalised word from documents. These are then looked up on Wikipedia to find matching articles describing a particular concept or entity. Each Wikipedia article is then mapped to its hypernym, which is the first noun phrase after *be* in the first sentence of the article. The process effectively generates gazetteers that group candidate phrases into a smaller set of hypernyms. In their experiment, they mapped over 39,000 search candidates to approximately 1,200 hypernyms. A large amount of these hypernyms are irrelevant to the named entity types required in the final task. However, they acted as alternative gazetteers and provided useful evidence for recognising named entities from the CoNLL2003 dataset. The same method was later used by Kazama and Torisawa (2008) in a Japanese NER task. This type of approach can be limited to the external knowledge resource of choice. It depends on particular content and structure of the knowledge resource, the coverage of which may also limit the capacity of the generated gazetteers.

Topicality and named entities - Clifton et al. (1999) argued that named entities are highly relevant to the topic of a document. In an experiment of topic identification, they showed that document topics represented by named entities are much more accurate and interpretable by humans than keywords. They further demonstrated the document clustering task can benefit from a representation based on named entities. Hassel (2003) argued that named entities are often important cues to the topic of a text. They are 'among the most information dense tokens of the text and largely define the domain of the text'. They showed that a text summarisation system can benefit by combining named entities in generating document summaries. While these argue that named entities can be indicative of document topics, Rennie and Jaakola (2005) suggested that the opposite can be also true, i.e., topic-oriented words can be useful indicators of named entities. They further suggested that topicality of words is equivalent to the sense of 'informativeness', a property which can be evaluated using informativeness measures.

Informativeness measures – Although a formal definition is lacking, it is generally agreed that informative words often demonstrate a 'peaked' distribution over a collection of documents such that the majority of their occurrences are found in a handful of documents (Church and Gale, 1995b). A large number of informativeness measures have been introduced in the past and used in a wide range of applications such as Information Retrieval (Mei et al., 2007), language modelling (Pan and McKeown, 1999), and machine translation (Wong and Kit, 2011). Most measures have employed two distributional properties of words: document frequency and term frequency in the corpus. This section briefly introduces informativeness measures that have been tested in NER or related tasks.

The first group of informativeness measures are solely based on document frequency. Document frequency is the number of documents in which *w* is found, given the entire collection as *D*, $d \in D$. The assumption is that words that are rare and unique to a small set of documents are informative. Inverse Document Frequency (Spark Jones, 1973) is a measure based on this hypothesis. It is calculated as

$$IDF(w) = log\frac{|D|}{|\{d : w \in d\}|}$$ **Equation 7.1**

Later Papineni (2001) showed that IDF is a better indication of the 'weight' of a word rather than its importance. Instead, the author proposed to quantify informativeness as the optimal gain, calculated as:

$$Gain(w) = \frac{|\{d : w \in d\}|}{|D|} \times (\frac{|\{d : w \in d\}|}{|D|} - 1 - log\frac{|\{d : w \in d\}|}{|D|}) \quad \textbf{Equation 7.2}$$

Under this model, extremely rare and extremely common words have low gain and are therefore less informative. Medium-frequency words have higher gain and are therefore, more informative.

The second group of informativeness measures explicitly study the 'peaked' or 'burst' distribution of words. They study the document frequency of words with respect to their overall frequency in the corpus. Bookstein and Swanson (1974) proposed the x^I measure to address this:

$$x^I(w) = tf(w,D) - |\{d : w \in d\}| \qquad \textbf{Equation 7.3}$$

where *tf(w, D)* returns the frequency of w in the entire collection. A similar approach proposed in Church and Gale (1995b) measure 'burstiness' as:

$$burstiness(w) = \frac{tf(w,D)}{|\{d : w \in d\}|} \qquad \textbf{Equation 7.4}$$

Intuitively, for two words with the same frequency in the collection, the one that is more concentrated will have the higher score. However, this score can be biased towards frequent words, which tend to be less informative (Rennie and Jaakkola, 2005).

Some proposed to evaluate informativeness by studying the degree to which the distribution of a word demonstrates the 'peakness' or 'burstiness'. These methods (Harter, 1975; Church and Gale, 1995b; Rennie and Jaakkola, 2005) often employ two kinds of well-known probability distribution models: binomial and Poisson. When applied to model word distributions, both model the correlation between the document frequency of a word and the average number of occurrences of the word per document. They can answer the question that, if empirically a word is found on average *n* times per document (frequency), then for any dataset, what is the likely number of documents in which it is found (i.e., document frequency) given that the frequency is *m*? For an informative word, one would expect high frequency numbers 'clustered' (thus a 'burst') for a small range of relatively low document frequencies, such as that shown in Figure 7.1.

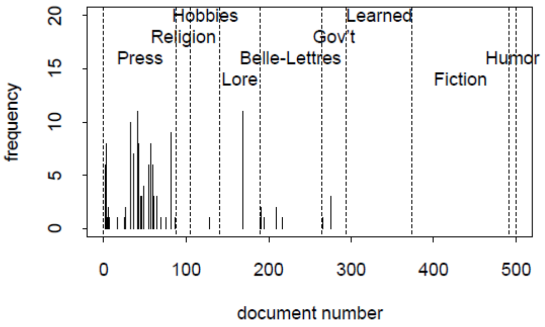


Figure 7.1. The word 'Kennedy' has high frequencies in a small set of documents belonging to the genre 'Press' in the Brown corpus (Church and Gale, 1995b)

Under the binomial model, documents in the collection are assumed to have uniform (or nearly equal) length *N* measured as number of words. It says that over the entire collection, the probability $P_w(k)$ that a word *w* has *k* occurrences (i.e., $freq(w, d) = k$) in a document can be computed as the chance of seeing *k* heads in *N* independent, biased-coin flips where the chance of heads on a single flip is λ. Under this model, λ is the mean probability of seeing *w* in any document. Thus the document frequency for each *k* can be computed as $D{\cdot}P_w(k)$. The Poisson model is a limiting case of the binomial model as *N* becomes unbounded while λ remains constant. Thus under the Poisson model the only factor that determines the frequency distribution of a word is λ.

It has been found that both models fit poorly with informative words (Harter, 1975; Church and Gale, 1995b). The binomial model tends to predict near linear distribution that fails to capture the 'peaked' nature; the Poisson model tends to significantly underestimate word frequencies with respect to document frequencies (Church and Gale, 1995b).

Based on these observations, Harter (1975) proposed a 'Mixture' model that better describes the frequency distributions of informative words. He hypothesized that for each word *w* that is informative in a document collection *D*, *D* can always be divided into two classes such that one is relevant to the subjects that *w* denotes while the other is irrelevant. Under this hypothesis, practically it would be more likely to see *w* in the class one documents but more unlikely to see *w* in the class two documents. The implication of this is that for each word that is informative in this collection, there are two modes of frequency distribution, which if modelled by Poisson, will have different λ values. Thus the frequency distribution of informative words can be modelled by a 2-Poisson model that takes into account both classes of documents. Let $Poi_1(k;\lambda_1)$ denotes the probability that a word *w* has *k* occurrences (i.e., $freq(w, d) = k$) in class one documents with the mean probability of seeing *w* in any member of this sub-class as λ_1, and $Poi_{II}(k;\lambda_2)$ denotes the probabilistic distribution of the word's frequency in the class two documents with the mean probability of seeing *w* in any member of this sub-class as λ_2, the 2-Piosson model returns the revised probability of frequency as:

$$2Poisson(k) = \pi Poi_I(k;\lambda_1) + (1-\pi)Poi_{II}(k;\lambda_2) \qquad \textbf{Equation 7.5}$$

The parameters λ_1, λ_2 and π have to be empirically derived based on data. Then naturally, the degree of informativeness of a word can be determined based on the 'fitness' of its frequency distribution against the prediction made by this model, a task that can be achieved using statistical significance testing metrics such as Chi-square. Further, Harter demonstrated that the informativeness of a word is purely based on the two λ values combined under the *z*-measure (Brookes, 1968):

$$z(w) = \frac{\lambda_1 - \lambda_2}{\sqrt{\lambda_1 + \lambda_2}} \qquad \textbf{Equation 7.6}$$

The *z*-measure denotes a sense of 'overlap'. Intuitively, if λ_1 is close to λ_2 then the mean probabilities of seeing a word in the two classes of documents are nearly the same, indicating that the two-class documents separation does not exist, or the word can simply be modelled by a Poisson distribution and is not informative.

In fact, the later study by Church and Gale (1995b) confirmed the validity of this hypothesis by showing that words tend to have different frequency distributions with respect to the genres of documents in the Brown corpus (Kucera et al., 1967). This is the main reason why a single Poisson model fails since it assumes a single λ for the entire collection, while in fact this can be different depending on the genre of sub-sections of the documents.

Rennie and Jaakkola (2005) also proposed a mixture model where they combined two binomial models. The frequency distribution of a word is firstly matched against a binomial model, and then the mixture model to derive two separate figures as the quantification of the matches. The degree of informativeness is then quantified as the log-odds ratio between the two figures. The intuition is that if the word is highly informative, it will have a bad match in the first case but a better match in the second case.

Church and Gale (1995a) proposed the measure of Residual IDF, which is the deviation of the actual IDF score (Equation 7.1) of a word from its 'expected' IDF score predicted based on the Poisson distribution:

$$RIDF(w) = IDF(w) - I\hat{D}F(w) \qquad \textbf{Equation 7.7}$$

$$I\hat{D}F(w) = log_2\left(\frac{1}{1 - Poisson(0;\lambda)}\right) \qquad \textbf{Equation 7.8}$$

where *1-Poisson(0; λ)* is the probability of a document having at least one occurrence of *w* predicted by the Poisson distribution model with *λ*. The hypothesis is based on the fact that Poisson model fits poorly with informative words. Thus a prediction of IDF based on Poisson can deviate from its actual IDF observed based on a corpus. Empirically, they showed that all words have real IDF scores that deviate from the expected value under a Poisson distribution model. However, informative words tend to have larger deviations than non-informative words.

Word topicality for NER – Very little work has explored the relationship between topicality of words and named entities in NER. The most relevant work includes Rennie and Jaakkola (2005) and Gupta and Bhattacharyya (2010), both of which studied named entity detection rather than classification. Rennie and Jaakkola (2005) argued that the topic-oriented words can be identified using informativeness measures and tested a number of measures, including IDF, RIDF, x^I, Gain, *z*-measure, and a mixture model. Using a corpus of forum posts annotated for restaurant names and gathered from a bulletin board dedicated to restaurant information, they analysed the usefulness of these measures in detecting restaurant names. Based on the data, they showed that IDF, Residual IDF and the mixture model are the best options while others (x^I, Gain, *z*-measure) 'have relatively little to offer in terms of identifying informative words'. To further validate this conclusion, the scores are used as features in a statistical named entity detection model. They also experimented with combinations of scores returned by different measures to obtain the best results. Gupta and Bhattacharyya (2010) proposed to create gazetteers dynamically at both training and testing phase using word informativeness measures. The core of the process is creating a lexicon by selecting the most informative words in a corpus – evaluated by a so-called 'ratio of frequency' measure that is identical to the 'burstiness' measure, and then filtered by an arbitrary threshold. The list is then pruned by two strategies. Firstly, words from the corpus are clustered based on their distributional similarity, and then clusters that contain mostly highly informative words (informative clusters) and that contain mostly non-informative words (non-informative clusters) are identified. The list is then modified by adding words from the informative clusters and discarding words in the non-informative clusters. Secondly, language and domain specific heuristics – e.g., removing stopwords – are used to discard words that are unlikely to be part of entity names. The final lexicon is considered to be words that are commonly used in naming entities and used as gazetteers in a statistical learning model. The method was shown to be effective in named entity detection for Hindi texts. Zhang et al. (2004) and Wan et al. (2011) studied methods for finding the most important named entities from the output of Chinese NER tasks. The named entities identified by an NER tagger were submitted to a further classification process, which aimed at filtering the most important named entities in the document. They showed that the distributional characteristics of named entities such as frequency are strong features for this purpose. These are essentially in line with the informativeness hypothesis; however, they do not deal with NER but a post-processing task.

All these studies have only presented a partial view of the usefulness of word topicality in the NER task. On the one hand, they do not directly address NER but a partial phase (e.g., named entity detection) or a related task. On the other hand, different methods have been evaluated in different languages, for single domains, and mostly single self-created datasets that are unavailable for comparative studies. It is unclear whether the lessons can be generalised across these boundaries to support NER in general.

7.3 Hypothesis

This study exploits the relationship between topicality of words and named entities for automatic generation of alternative gazetteers for NER. It addresses Hypothesis H2.2:

> **H2.2 Alternative gazetteer: named entities are highly to contain topic-oriented words specific to a document. The topicality of words can be evaluated based on the relevance measures widely used for Information Retrieval. It can be used for generating alternative gazetteers for NER.**

It builds on the same ground of the previous work by Rennie and Jaakkola (2005) and Gupta and Bhattacharyya (2010): the topicality of words can be useful for identifying named entities. The major difference is that the topicality of words is defined locally with respect to documents, measured by their relevance to documents. As discussed before, informativeness is measured globally with respect to the entire collection, and it may not always represent topicality.

First, document topics can vary largely even if they belong to the same domain. As mentioned, the Genia corpus widely used for biomedical NER contains thousands of abstracts of medical journal publications. Each of these focuses on rather narrowly defined, specific topics such as specific proteins (e.g., NF-Kappa B, proteasome), DNAs (e.g., AP-1 site, murine IL-2 promoter), cell types (e.g., Th1 cell, T cell), or specific interactions between these objects (e.g., binding, signalling). The CoNLL2003 NER corpus spans across a number of different domains, containing news articles of a wide range of topics, such as economics, politics, sports and entertainment. This has led to largely varying vocabularies as well as varying frequency patterns of words at individual document basis. Informativeness measures focus on the *global* distribution of words and ignores such varying patterns specific to documents. A major component of many informativeness measures is document frequency. The assumption is that informative words tend to be specific to a small set of documents and therefore, have low document frequency. However, this study shows that informativeness biased by document frequency can mis-represent topicality particularly when documents in a corpus are characterised by largely varying vocabularies, an indicator of varying topics.

Second, informativeness scores are globally uniform and specific to a collection. As a result, ambiguous words that carry different senses in different document contexts can be mis-interpreted. In the biomedical domain, a fair amount of entity names can contain common English words. The word ‘bright’, ‘white’ and ‘cycle’ can be used to refer to protein or gene names in some documents but also widely used as common words carrying no special senses in most documents (Morgan and Hirschman, 2003). A uniform informativeness score cannot distinguish these cases.

Based on the Genia corpus as a sample, Table 7.1 and Table 7.2 below show several examples to illustrate the above limitations of informativeness measures. The IDF and RIDF measures were chosen because they were shown to be effective at identifying informative words in a different domain (Rennie and Jaakkola, 2005). Each unique word is scored using the two measures and ranked by the scores. Then, the documents containing the word and the documents containing annotations including the word are counted respectively.

	IDF total different scores =313			
Word	Rank by score.	#Docs in which word found as part of NEs	#Docs containing the word	Error type
2alpha	2	1	2	Ambiguity
TT	2	1	2	Ambiguity
cripple	2	1	2	Ambiguity
formally	1	0	1	Biased by doc freq
disappointing	1	0	1	Biased by doc freq
fifty	2	0	2	Biased by doc freq
get	1	0	1	Biased by doc freq

Table 7.1. Examples of highly informative words (IDF) that can harm learning accuracy

IDF is a measure that is purely based on document frequency. This has led to a major limitation: empirically, 57% of all unique words have received the highest IDF score, because they are found in only one document. This suggests that documents have used largely varying vocabularies, a strong indicator of largely varying topics. However, many of these words are not related to the topics of documents. Some negative examples are shown in Table 7.1. The word 'formally' (document id 99138988, 'formally demonstrates…'), 'disappointing' (document id 99300859, 'The disappointing results of …'), 'fifty' (document id 96071057, 95161757, used to describe experimental data), and 'get' (document id 97210575, 'To get further insights into…') receive the highest IDF score because they are only found in one or a couple of documents. However, these words do not carry useful information with respect to the topics of the document and are not part of any entity names. Furthermore, the globally uniform informativeness score can mislead extraction of entity names containing ambiguous words such as '2alpha', 'TT', and 'cripple'. For example, the word '2alpha' is a highly informative word according to IDF (ranked as the 2nd most informative word). In the document with id 99008517, it is used to refer to a protein in the sentence '… LEF-1 and PEBP 2alpha …' and annotated as part of a protein entity (PEBP 2alpha). While in the document with id 20570933, it is used to refer to a type of natural prostaglandin in the sentence 'Interestingly, addition of PGF (2alpha), which was not known to affect …', which is not related to the core topic and not annotated as named entities. Similarly, 'TT' and 'cripple' denote different meanings in different documents, where in one case they are used as part of named entities that are relevant to the topic while in the other they are used to describe other information that is less relevant.

	RIDF total different scores = 1717			
Word	Rank by score.	#Docs in which word found as part of NEs	#Docs containing the word	Error type
tip	24	1	2	Ambiguity
bright	112	2	3	Ambiguity
interleukin-5	1083	10	10	Biased by doc freq
oncogene	798	45	48	Biased by doc freq
NFAT-1	219	12	12	Biased by doc freq
CD4	405	114	149	Biased by doc freq

Table 7.2. Examples of informative and non-informative words (RIDF) that can both harm learning accuracy

RIDF partially overcomes the limitations of IDF by also taking into account word frequencies in the collection. It promotes words that are found many times, but within a specific set of documents. Empirically, this produced 1,717 unique scores. Manual inspections have shown that the upper sections of the ranked list contain mostly words that are strongly related to topics of documents and that are often part of entity names. However, it still suffers from the same limitations as IDF. As it will be shown in further analyses later in this chapter (Section 7.6), the upper sections of the list represent only a very small proportion of named entities in the dataset and are therefore, not very informative to the NER learner. In contrast, a much larger amount of entity names contain words from the lower sections of the list, which also include the majority of noisy words. In Table 7.2, the examples of ‘interleukin-5’, ‘oncogene’, ‘NFAT-1’ and ‘CD4’ illustrate this problem. For all of the four words, in their occurrences they are primarily used as entity names or as part of the names. However, they spread across a wide range of documents (i.e., high document frequency), which biased their informativeness scores. Some words also suffer from the problem of ambiguity. The word ‘bright’ (ranked at the 112nd) is used as a protein name in the sentence ‘Bright (B cell regulator of IgH transcription) is a B cell-specific, matrix associating region-binding protein that …’. It is highly related to the topic of the document with id 21293104, which discusses ‘transcriptional activation by a matrix associating region-binding protein’. In the different document with id 96178227, it adopts the common sense of ‘shining’ in the sentence ‘Specific bright foci of GATA-1fluorescence were observed in erythroleukaemia cells…’, where it is less relevant to the topic.

Further analyses were carried out to study the problems with IDF and RIDF in a task of NER using a total of 5 datasets. These will be presented later in Section 7.6.

During the manual inspection, it has been found that for many errors by the two informativeness measures, word frequencies observed within the local document context can be an effective indicator of topicality. For example the words ‘formally’, ‘disappointing’ and ‘get’ have only a single occurrence in the documents. While in most cases, ‘interleukin-5’, ‘oncogene’, ‘NFAT-1’ and ‘CD4’ are found multiple times. The ambiguous words ‘2alpha’ and ‘bright’ are also found only once in the documents where they denote an irrelevant sense to the topics, but many times when they are more topic-oriented. This motivated the consideration of distributional characteristics observed at document level in measuring word topicality.

In the area of Information Retrieval, relevance measures have been used to assess the importance of a word to a document. They represent a sense of topical relevance (Mizzaro, 1997) and often employ word frequencies in document contexts as an important type of feature. Therefore, in this study, relevance measures are proposed as a proxy for topicality. Furthermore, it is expected that highly topic-oriented words are rare, since topics are often composed of a confined small set of keywords. However, they may be found in a large proportion of named entity mentions in a document, since named entities that are highly relevant to the topic of a document are also likely to be repeated frequently. Based on these hypotheses, for each document, words are grouped based on their level of topicality and it is hypothesized that those falling under the highly topic-oriented groups are indicative of named entities. This has led to the creation of document-specific, untyped gazetteers. Details of the method are presented in the next section.

7.4 Alternative Gazetteers based on Topicality

The method of topicality-based alternative gazetteer generation consists of two parts: measures of topicality of words (Section 7.4.1); feature extraction method that generates gazetteers based on topicality (Section 7.4.2).

7.4.1 Measuring Topicality

Let *topcat* denote a function that measures the topicality of words (*w*) in a document (*d*) as positive real numbered values:

$$topcat(w,d) \in \Re^{+} \qquad \textbf{Equation 7.9}$$

The first and the simplest relevance measure that can be used to evaluate topicality is Term Frequency, denoted by *TF(w, d)*, which is simply the count of occurrence of *w* in *d*.

Intuitively, words associated with the topics of a document are likely to be repeated throughout the text and therefore have higher frequency since they are the focus of the content. This simple technique was used in earlier studies and found effective in extracting important terms from documents (Dagan and Church, 1994). However, it is also well-known for its bias towards highly frequent but non-content bearing words (e.g., functional English words), and inability to identify less frequent yet equally important words, such as those that are unique to a smaller set of documents in the whole collection. Thus the second measure used for evaluating topicality is Term Frequency - Inverse Document Frequency (Spark Jones, 1973) which is the classic measure used in IR for evaluating the word-document relevance:

$$TFIDF(w,d) = TF_{norm}(w,d) \times IDF(w) \qquad \textbf{Equation 7.10}$$

where $TF_{norm}(w,d)$ is the normalised frequency of w in d, calculated as:

$$TF_{norm}(w,d) = \frac{TF(w,d)}{\sum_{w' \in d} TF(w',d)} \qquad \textbf{Equation 7.11}$$

and $IDF(w)$ is the inverse document frequency of w in the entire collection D $(d \in D)$, calculated using Equation 7.1. The intuition is that word associated with the topic of a document should be frequently used and also unique to that document. The latter can be measured by document frequency – a word that is only found in a handful of documents is likely to be specific to those documents and bear specific meanings; in contrast, words that are frequently found in any document are likely to be functional words or less important. The major difference from IDF and other informativeness measures that employ document frequency is that it balances IDF with word frequency in the local document context. This returns a score that is specific to individual documents, and balances out the effect of over-promoting low document frequency words in a collection.

The third measure is the Weirdness function (WD) introduced by Ahmed et al. (1999) for document indexing and retrieval. It is also based on word frequency in the local document context, and captures the sense of ‘uniqueness’ by comparing the distribution of a word in the target document against its distribution in a reference corpus:

$$WD(w,d) = \frac{TF_{norm}(w,d)}{TF_{norm}(w,C)}$$ **Equation 7.12**

where $TF_{norm}(w, C)$ is the normalised frequency of w in a reference corpus (C), indicating the probability of encountering w in other context. The intuition is that words that are more likely to be found in a document than a reference collection are ‘special’ to that document and therefore, more relevant to its topic. It is similar to TFIDF in the way that both normalise $TF_{norm}(w,d)$ by a different factor. Empirically, they have led to different results.

Additionally, this study also proposes a function that combines both TFIDF and WD with equal weights. The motivation is to balance the different views of ‘uniqueness’. Given W_d the set of words found in d, their TFIDF and WD scores are firstly calculated using the above equations and then ranked by the scores to obtain two ordered list R_d^{TFIDF} and R_d^{WD}. The final combined score for a word w, to be called as **Combined Inverse Rank (CIR),** is calculated as:

$$CIR(w,d)=\frac{1}{R_d^{TFIDF}(w,d)}\times 0.5+\frac{1}{R_d^{WD}(w,d)}\times 0.5$$ **Equation 7.13**

Furthermore, for all of the four measures, a list of stopwords is filtered out prior to the calculation of topicality. Stopwords will always have a topicality score of 0.

7.4.2 Gazetteer Generation

The scores returned by the *topcat* measures are real numbers indicating how topic-oriented a word is to a document. As discussed before, they are document-specific and unbounded in range, and cannot be used directly by statistical learning models since the scores are non-comparable across documents and therefore, non-generalisable.

To transform the scores to useful features, this work proposes to create untyped, document specific gazetteers based on the strength of topicality of words. Generally, it is expected that highly topic-oriented words are rare but can be found in a large proportion of named entity mentions in a document. As the scores drop, their usefulness (i.e., being indicative of named entities) drops disproportionally faster. Thus a handful of highly topic-oriented words are most useful indicators of named entities in a document. Theoretically, if it is possible to determine a cut-off threshold for each document such that it correctly splits the highly topic-oriented words for the document from the others, one can use these words as gazetteers for NER in that document. However in practice, this may be infeasible since the number of documents can be large and the thresholds can vary. Instead, this work proposes to use a simple binning method to split the ordered list of words from a document into *k* equal sized intervals, and construct a binary feature for each interval.

Effectively, this transforms the non-comparable topicality scores for each document into a uniform set of *k* groups for each document, while the correspondence between the group numbers (between *1* and *k*) and the classification decision can be automatically learned from training data. Formally, this can be represented as a feature function f_{bin} that returns a vector of binning values for every word *w* in *d*:

$$f^{i}_{bin}(w) = \begin{cases} 1, & \text{if } \frac{|W_d|i}{k} < rnk(w,d) \leq \frac{|W_d|(i+1)}{k} \\ 0, & \text{otherwise} \end{cases}$$ **Equation 7.14**

where $i \in \{1, 2, 3, \ldots k\}$, W_d is the collection of all words in *d*, and *rnk(w, d)* returns the position of *w* in W_d ranked by their topicality scores in descending order. Essentially, the features are used in the same way as gazetteers. Each group can be considered a gazetteer (or non-gazetteer) that may have either positive or negative examples. This is different from the traditional notion of gazetteer that usually contains only positive examples. Another key difference is that the gazetteers created in such a way are document-specific, while conventionally, gazetteers are not bounded to document context. This is because, as argued before, the topicality of words is document-specific. Therefore, gazetteers generated based on topic-oriented words may only be applicable within the specific document context.

7.5 Evaluation

This section presents the experiments designed to evaluate the effectiveness of topicality-based gazetteers for NER. The proposed method is compared against two other methods based on the similar ground (Rennie and Jaakkola, 2005; Gupta and Bhattacharyya, 2010) on five widely used datasets covering three domains. To focus on the effects of the topicality based features, a uniform learning model is created and used as a baseline; then, the topicality based features generated by each method are added to the baseline and the changes in learning accuracies are compared.

7.5.1 Datasets

The experiments contain five corpora selected from the newswire, biomedical and archaeology domains, as shown in Table 7.3.

Corpus	Domain	Docs	Avg. len. by tokens
Archaeo	Archaeology	38	6862
Bio1	Biomedical	100	226
CoNLL03	Newswire	1,000	237
Genia BioNLP04	Biomedical	2,400	227
Yapex	Biomedical	200	225

Table 7.3. Datasets used for experiments

The Genia BioNLP04 (Genia in the following) dataset is obtained from the Bio-Entity Recognition Task at BioNLP/NLPBA 2004 (Kim et al., 2004). It contains 2,400 MEDLINE abstracts with about 56.5k annotations of protein (60%), cell type (15%), cell line (7%), DNA (16%) and RNA (2%) entities. The Bio1 (Collier et al., 2000) dataset contains 100 MEDLINE abstracts with about 3.2k annotations[6] of protein names (63%), DNA (11%), and SOURCE (26%), which include 7 sub-types such as 'cell type' and 'cell line' etc. In this study these are treated as a single entity type. The Yapex (The Yapex Corpus, 2005) dataset contains 200 MEDLINE abstracts with about 3.7k annotations of protein names. The CoNLL03 dataset contains 70% of the original English dataset used by the CoNLL2003 shared NER task (Sang and Meulder, 2003). Documents that contain only tables or lists of entities such as game score boards or phone directories, are discarded. This is because they do not carry a focused topic and are not suitable testbed for the hypothesis. The dataset contains newswire articles with about 22k annotations of person names (PER, 28%), locations (LOC, 32%), organisations (ORG, 21%) and miscellaneous (MISC, 19%) such as events, languages and nationalities.

While these datasets contain documents of similar size, to gain a comprehensive understanding of the effectiveness of different methods the Archaeo corpus is included. The dataset is based on the annotations created in Chapter 5 and used in Chapter 6. Additional 8 documents were added, creating a total of 38 articles averaging about 7,000 words. There are about 17.5k annotations of three types: archaeological temporal terms (TEM, 23%), such as 'Bronze Age', and '1089AD'; location (LOC, 14%), which is UK-specific and often refers to place names of findings and events; subject of interest (SUB, 63%),

[6] 'RNA' annotations are ignored as there are less than 40 instances only.

which is a highly heterogeneous category containing terms from various domains, such as architecture, warfare, maritime, and education. Each dataset is then split into five equal parts for five-fold cross-validation experiments.

7.5.2 Baseline

The baseline NER learner used for this experiment is the same SVM-based NE tagger previously introduced in Chapter 4, and used in Chapter 5 and Chapter 6. The following basic features are used:

- The exact token string, and its stem
- The orthographic type of the token (e.g., alpha-numeric, digits only, capitalised)
- Context window of 3 – the above features are applied to the previous and the following three tokens of the current token

The baseline setting is denoted as ***B*** in the following. No domain and language specific features or data specific optimisations are used. Although it is known that domain specific features and optimisations can be particularly effective, the goal of this study is to thoroughly test the effectiveness of different methods of using topic-oriented words for NER and their generality across domains, rather than maximising learning accuracies.

7.5.3 Methods for Comparison

The proposed method of document specific gazetteer generation based on word topicality (denoted as *topcat*) is compared against the studies by Rennie and Jaakola (2005) and Gupta and Bhattacharyya (2010). Both proposed to use topic-oriented words for named entity detection but calculated topicality using informativeness measures. In this work, their methods are adapted and tested for NER.

Topcat – For each of the *topcat* measures introduced before, the binning method is applied to generate document specific gazetteers which are then used as additional features to the baseline. In the experiment, *k* is initially set to *20,* and the British National Corpus (The British National Corpus, 2007) is used as the reference corpus for WD. Thus there are four settings: *B, TF_{bin}; B, $TFIDF_{bin}$; B, WD_{bin} and B, CIR_{bin}*. Several *topcat* measures have been implemented based on the JATE toolkit[7].

Rennie and Jaakkola (2005, **RJ05**) – In **RJ05**, several different informativeness measures are tested for named entity detection. Informativeness scores are computed for each word using one of the measures, and the scores are then normalised against the mean or median score to obtain relative scores, which are used as features in a statistical named entity detection model. It has been found that the IDF and RIDF are two of the most effective measures for identifying informative words and also contributed most to the named entity detection task. Additionally, as discussed before (Section 7.3), both measures can be very sensitive to document frequency and lead to inaccurate predictions of informative words. Therefore, a third informativeness measure – the 'burstiness' measure (Equation 7.4) – is also chosen to be tested following the RJ05 approach. Therefore, four different settings are created using the IDF, RIDF, and burstiness measures:

1) *B, median-normalised relative IDF score (IDF_{med})*
2) *B, mean-normalised (i.e., average score normalised) relative RIDF score ($RIDF_{avg}$)*
3) *B, median-normalised relative burstiness score (BUR_{med})*
4) *B, mean-normalised relative burstiness score (BUR_{avg})*

IDF scores are normalised against the *median* of all scores; RIDF scores are normalised against the *mean* of all scores. These settings are chosen because they contributed to the best results in the work by Rennie and Jaakkola. For the burstiness measure, both normalisation methods are tested. Furthermore, the feature generation is adapted in the following ways. First, scores are rounded to two decimal places because empirically this led to better results. Second, the same set of stopwords used for *topcat* based methods is also used to filter out noisy words. Empirically, the adaptations reduced the sparseness of feature space and led to better learning accuracy. These are referred to as RJ05 based methods in the following.

Gupta and Bhattacharyya (2010, **GB10**) – The method introduced in **GB10** employs fairly complex processing involving computing distributional similarity, clustering, and filtering by language and domain specific heuristics. However, the core lesson is creating a

[7] JATE – Java Automatic Term Extraction toolkit, last retrieved on 14 Mar 2012 from http://code.google.com/p/jatetoolkit/

lexicon to be used as a gazetteer automatically from informativeness scores. The informativeness measure used in this study is the burstiness measure. To compare this against *topcat* based methods, the informativeness score is calculated for each unique word extracted from a corpus using this measure. Then words whose scores exceed an arbitrary threshold are selected for gazetteers. Since it is impractical to manually inspect the lists and select a threshold for each dataset and each cross-fold experiment, the top *n%* of the entire list (each from the training and testing parts separately) is chosen. Empirically *n* is set to *20*, which has produced the best results on some sample datasets. Additionally, the four *topcat* measures are also adapted to calculate a *global* score for each word based on their distributional statistics in a corpus. Specifically, the TF measure is adapted to use word frequency in a corpus; the TFIDF and the WD measures use normalised word frequency in a corpus; and the CIR measure combines the adapted TFIDF and WD. Thus for GB10 five settings are created: B, Gaz_{TF}; B, Gaz_{TFIDF}; B, Gaz_{WD}; B, Gaz_{CIR}; B, Gaz_{Bur}. Stopwords are also filtered out. These are referred to as GB10 based methods in the following.

7.5.4 Results

Firstly, the baseline system is evaluated on all five datasets. The baseline has obtained comparable results to state-of-the-art. The results are as shown in F1 in Table 7.4.

Corpus	**Entity Types**					**Micro-average**
Archaeo[8]	LOC	SUB	TEM			
	60.1	66.14	80.1			68.43
Bio1	Protein	DNA	Source			
	68.56	50.36	64.3			65.74
CoNLL03	LOC	MISC	ORG	PER		
	82.77	79.43	66.03	81.74		78.22
Genia	Protein	DNA	RNA	Cell type	Cell line	
	64.12	58.04	64.99	64.17	56.15	62.49
Yapex	Protein					
	55.49					55.49

Table 7.4. Baseline accuracy in F1

[8] Note that the results on the Archaeo dataset are different from those in Chapter 6. This is because different features have been used in this study in other to keep a uniform experiment setting for all datasets.

Next, the *topcat* based methods are evaluated. The gazetteer features are added to the baseline system, and the best results are shown in Table 7.5.

For ease of comparison, in the following only micro-average F-measure scores are used. The absolute improvement in F-measure (micro-average) over the baseline is shown in Figure 7.2. The RJ05 and GB10 based approaches are also evaluated and the changes to the baseline are shown in Figure 7.3 and Figure 7.4 respectively.

Corpus	**Entity Types**					**Micro-average**	**Topcat measure**
Archaeo	LOC	SUB	TEM				
	62.62	68.4	81.36			70.74	CIR
Bio1	Protein	DNA	Source				
	70.17	51.23	64.79			66.99	CIR
CoNLL03	LOC	MISC	ORG	PER			
	83.13	79.69	66.49	83.1		79.13	WD
Genia	Protein	DNA	RNA	Cell type	Cell line		
	65.38	58.97	66.19	66.76	57.77	64.0	CIR
Yapex	Protein						
	59.36					59.36	CIR

Table 7.5. Best accuracy obtained with topcat based methods

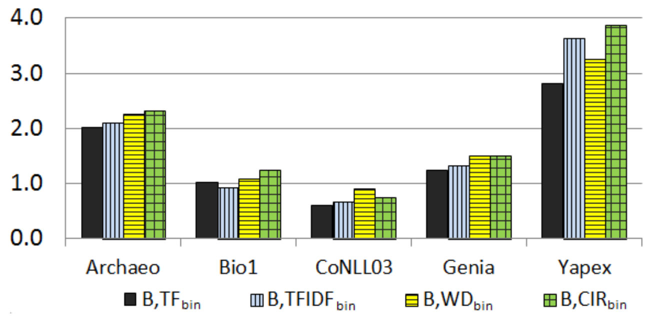


Figure 7.2. Absolute improvement over the baseline obtained by *topcat* based methods

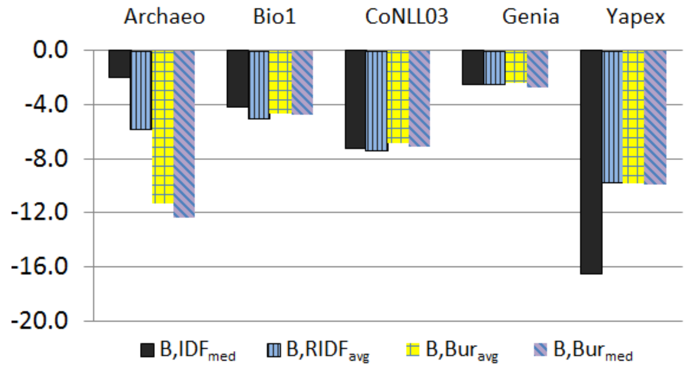


Figure 7.3. Absolute changes to the baseline obtained with the RJ05 based methods

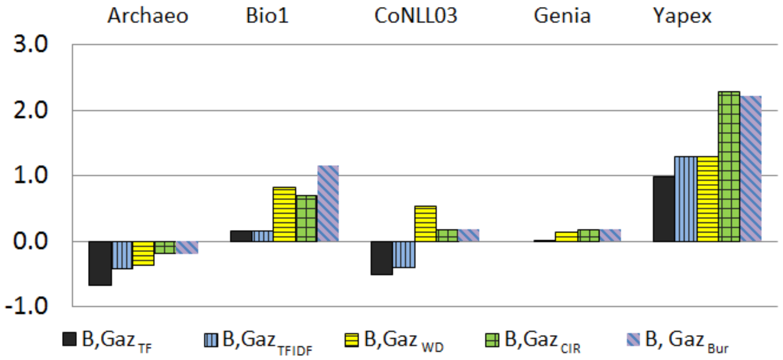


Figure 7.4. Absolute changes to the baseline obtained with the GB10 based methods

7.6 Discussion and Analysis

7.6.1 Overview

Firstly, regardless of the choices of *topcat* measures, the document-specific gazetteers based on the locally (i.e., within specific document context) assessed topicality of words have consistently improved the baseline. The maximum improvements for each dataset are: *2.3 (CIR) for Archaeo; 1.3 (CIR) for Bio1; 0.9 (WD) for CoNLL03; 1.51 (CIR) for Genia; and 3.9 (CIR) for Yapex*. It is particularly effective for specialised domains, in which NER is much harder. Also, considering only the Archaeo dataset, the results on the SUB and LOC entities are very competitive compared to those reported previously in Section 6.5. This is very encouraging as it suggests that the quality of the gazetteers created in such a completely unsupervised way can be as good as those built based on carefully curated resources. Further, taking into account the diversity in the datasets, it is safe to conclude that the proposed method is generalisable and effective for NER.

The different accuracies given by different *topcat* measures are rather insignificant. In most cases (expt. Yapex), this is less than 0.01. TFIDF and WD outperformed TF marginally, and combining both of them (CIR) can lead to further small improvement, except for the CoNLL03 dataset.

In comparison, the results obtained with RJ05 based approaches are contradictory to the findings by Rennie and Jaakkola. The GB10 based methods are also found to be less effective. For RJ05, the methods originally succeeded in finding a single type of entity – restaurant names – in forum posts, where each thread often discusses a restaurant and is treated as a separate document. It is possible that the nature of the texts has made informativeness scores such as IDF and RIDF sensible features for named entities. Forum discussions are generally very focused on narrowly defined distinctive topics, and can involve a more limited set of vocabularies compared to formally written scientific and news articles. However, due to the unavailability of the original data from Rennie and Jaakkola, it is not possible for direct analysis. In the NER experiments using standard evaluation datasets, these methods have caused significant drop in accuracies on all datasets. For GB10, in most cases the gazetteer based features have contributed to small improvements over the baseline. For the Bio1 and CoNLL03 datasets some settings produced comparable results to *topcat* based methods. But the improvements are generally smaller. Also there are several occasions (especially on the Archaeo dataset) where these features have failed and caused decreased accuracy. This is likely to be caused by inappropriate thresholds for gazetteer selection. Overall the results are rather inconsistent and empirically deriving suitable thresholds for each dataset and measure can be difficult. Although the GB10 settings are not an identical replication of the original method, we believe they provide a useful reference. Both observations have confirmed that informativeness is not always a good indicator of named entities and can make spurious features that harm learning accuracy.

7.6.2 Feature Analysis

In order to uncover the contributing factors to the effectiveness of *topcat* based gazetteers and understand the contradictory results of the RJ05 based methods, a series of analyses is conducted to study whether and how the additional features to the baseline can contribute to the learning accuracy. For *topcat* based methods, 20 (k = 20) document specific untyped gazetteers are generated, which is equivalent to 20 unique feature values. All four *topcat* measures are studied. For RJ05 based methods, informativeness scores of words are used as additional features. The number of unique scores – i.e., feature values – depends on the choice of informativeness measures and the data. Features generated by IDF, RIDF, and burstiness with the median score normalisation are studied. Features based on burstiness with the mean score normalisation are excluded from the analysis since there are no significant difference in the learning accuracies given by the two different score-normalisation methods (see Figure 7.3).

Two types of analyses are carried out for each dataset using a sample of positive and negative examples (words that are part of an annotation and words that are not) drawn from the gold standard. The first addresses the *precision* perspective by computing the fraction of positive instances among both positive and negative instances captured by each feature value. Intuitively, the higher the fraction, the more useful the feature can be to boost learning precision. This is named **ratio of positives** and denoted by **POSRatio**. Let *v* be a unique feature value (e.g., for *topcat*, this could be the gazetteer id given by the equal interval binning method; for RJ05, this could be a unique informativeness score), *PosInst(v)* and *NegInst(v)* be the functions that return the set of positive and negative instances in the sample represented by *v* respectively, then *POSRatio(v)* is formally defined as:

$$POSRatio(v) = \frac{|\{w : w \in PosInst(v)\}|}{|\{w : w \in PosInst(v)\}| + |\{w : w \in NegInst(v)\}|} \qquad \textbf{Equation 7.15}$$

The second analysis addresses the *recall* perspective by computing the fraction of all positive instances in the sample captured by each feature value. Intuitively, the higher the fraction, the more useful the feature can be to boost learning recall. This will be referred to as **coverage of positives** and denoted by **POSCov**. Let *V* denote the entire set of unique feature values, *POSCov(v)* is computed as:

$$POSCov(v) = \frac{|\{w : w \in PosInst(v)\}|}{\sum_{v \in V} |\{w : w \in PosInst(v)\}|} \qquad \textbf{Equation 7.16}$$

Additionally, the fractions of the entire sample data represented by each feature value are also calculated. Intuitively, one would like features that are useful to precision or recall to capture a higher volume of data. This is referred to as **volume** and denoted by **VOL**:

$$VOL(v) = \frac{|\{w : w \in PosInst(v)\}| + |\{w : w \in NegInst(v)\}|}{\sum_{v \in V} |\{w : w \in PosInst(v)\}| + |\{w : w \in NegInst(v)\}|} \qquad \textbf{Equation 7.17}$$

The findings: The analyses are performed for every *dataset-topcat measure* pair combinations, as well as every *dataset-informativeness measure* pair combinations for the three informativeness measures used by RJ05 based methods. 50% of each dataset (by the number of documents) are used as sample for analysis.

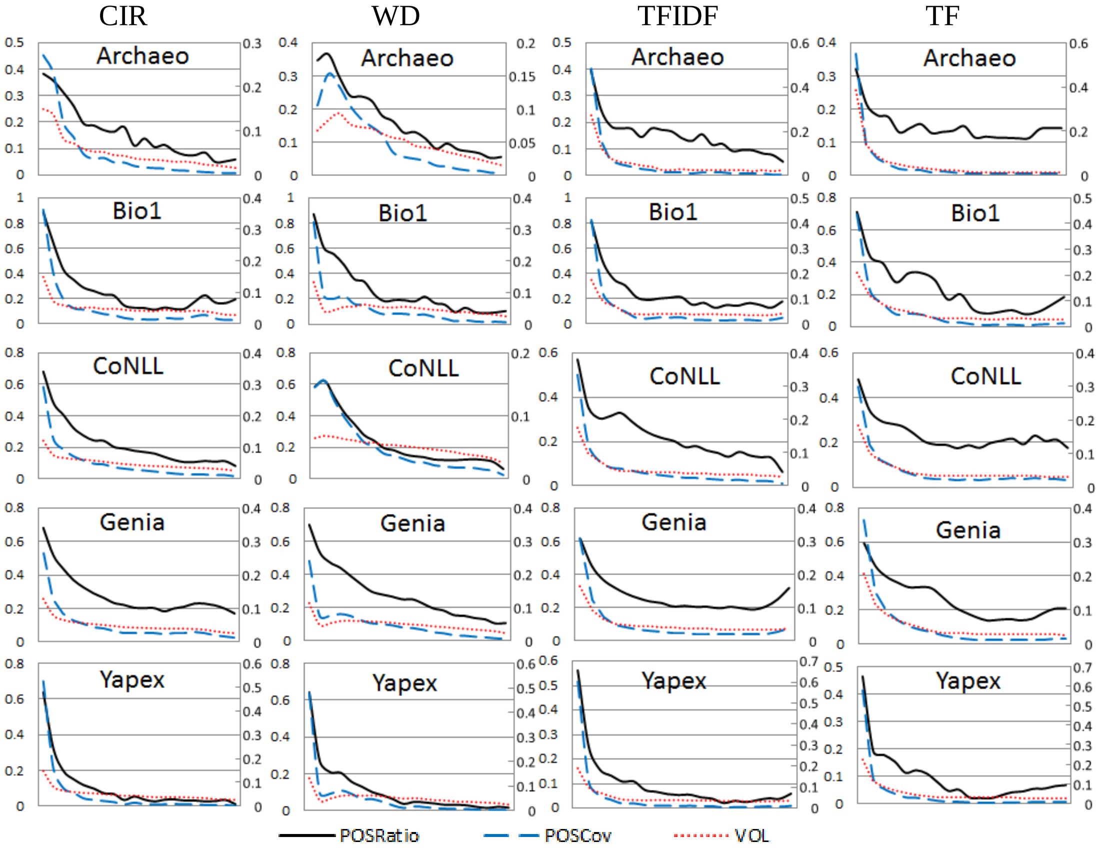


Figure 7.5. Feature analysis for topcat based methods (POSRatio, POSCov, VOL)

For *topcat* based methods, on each dataset, 20 document specific gazetteers are generated based on each *topcat* measure, which creates 20 unique feature values. The results for each dataset-topcat measure pair are shown in Figure 7.5. On each chart, the *x*-axis corresponds to unique feature values, i.e., the 20 gazetteers, organised by the descending order of topicality from left to right. Therefore, the leftmost feature corresponds to the gazetteer that contains the most topic-oriented words. The *y*-axes denote fractions. The POSRatio curves are aligned against the left *y*-axis; the POSCov and VOL curves are aligned against the right *y*-axis.

As shown in Figure 7.5, for all topcat measures, both the POSRatio and POSCov curves show a clear pattern of non-linear and long-tailed distribution on all datasets. This suggests that highly topic-oriented words (in gazetteers of smaller IDs, e.g., 1 or 2) are useful to both the recall and precision of the NER learner and as the topicality scores drop, their usefulness decreases more rapidly. This non-linearity relation between the topicality of words and classification decisions is well reflected by the simple binning based gazetteer generation, which effectively collapses most topic-oriented words for a document into a handful of gazetteers useful for discrimination. The strongest pattern is noted for the

Yapex dataset, which possibly contributed to the greatest improvement among all datasets. For example with the CIR topicality measure, the leftmost feature (corresponding to the gazetteer containing words with the highest topicality scores) has received both a very high POSRatio of over 0.6 and a very high POSCov of over 0.5. This means that more than 50% of positive instances are represented by this feature, and among all instances represented by this feature, over 60% are positive. Furthermore, the VOL curves also show a generally consistent pattern of a fairly uniform distribution that tends to become more non-linear at the left end of the curves (except WD on CoNLL). Accordingly the most useful gazetteers account for a higher proportion of data, justifying why this is a useful feature for the classifier.

Next, the same set of analyses is applied to RJ05 based methods. As described before, three informativeness measures are studied, and the number of unique features depends on the data and the choice of informativeness measures. A summary of these are shown in Table 7.6.

Corpus	IDF	RIDF	burstiness (median)
Archaeo	36	486	611
Bio1	43	195	131
CoNLL03	57	510	242
Genia BioNLP04	74	510	363
Yapex	48	253	168

Table 7.6. Number of unique features generated on each dataset by RJ05 based methods

The results for the IDF based features are shown in Figure 7.6. Same as Figure 7.5, on each chart, the *x*-axis corresponds to unique feature values organised by the descending order of informativeness from left to right. The POSRatio curves are aligned against the left *y*-axis; the POSCov and VOL curves are aligned against the right *y*-axis.

As shown in Figure 7.6, features given by the IDF informativeness measure do not display consistent patterns. POSRatio values appear to be random with respect to the level of informativeness and higher values tend to spread towards both ends of the *x*-axis, possibly suggesting that there is no definitive correlation between the informativeness of words and NER learning precision. For the POSCov curves, non-linear patterns are noted on some datasets (Bio1, CoNLL, Yapex), while random patterns are noted on others. There is no consistency between the POSRatio and POSCov curves. Furthermore, the VOL curves across different datasets also appear to be more random compared against those obtained with topcat based measures. This nature of randomness and inconsistency displayed by the informativeness based features can make it difficult to generalise for a learner, which may explain the damaged learning accuracies.

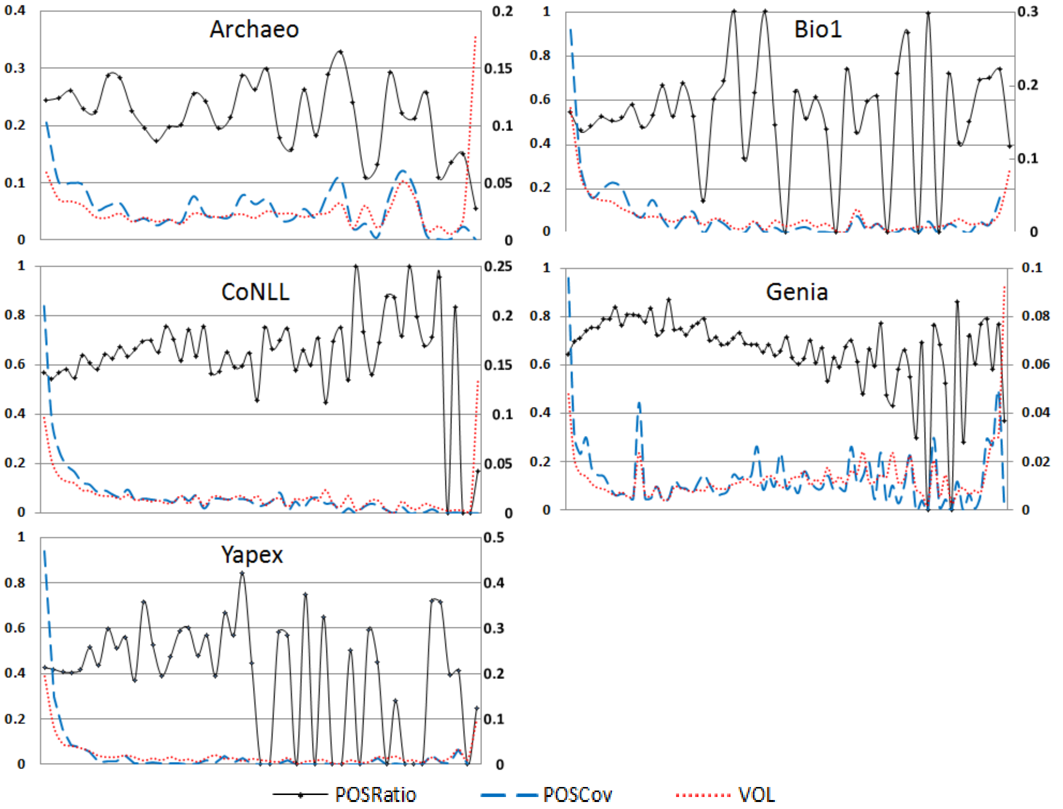


Figure 7.6. Feature analysis for RJ05 based method with the IDF measure (POSRatio, POSCov, VOL)

The results for the RIDF and burstiness based features are shown in Figure 7.7 and Figure 7.8 respectively. Due to the high numbers of unique features given by these measures and for presentation purpose, POSRatio values are shown as points. In general, similar to the features generated by the IDF measure, no consistent patterns can be generalised across datasets or between different POSRatio and POSCov values. POSRatio values appear to be random. Both high and low POSRatio values can be found for any ranges of informativeness scores, suggesting that informativeness scores may be non-discriminative for learning NER. POSCov curves and VOL curves also behave randomly and inconsistently with POSRatio values, making it difficult to generalise. Although on some datasets (e.g., Bio1, CoNLL), highly informative words appear to be useful for learning precision since these features obtained POSRatio values of 1.0 (as indicated by the points that cross the leftmost sections of the topmost gridlines), they account for a very small portion of data (less than 1%, as indicated by the corresponding points on the POSCov and VOL curves) and therefore, may be less informative to the learner. For exa-

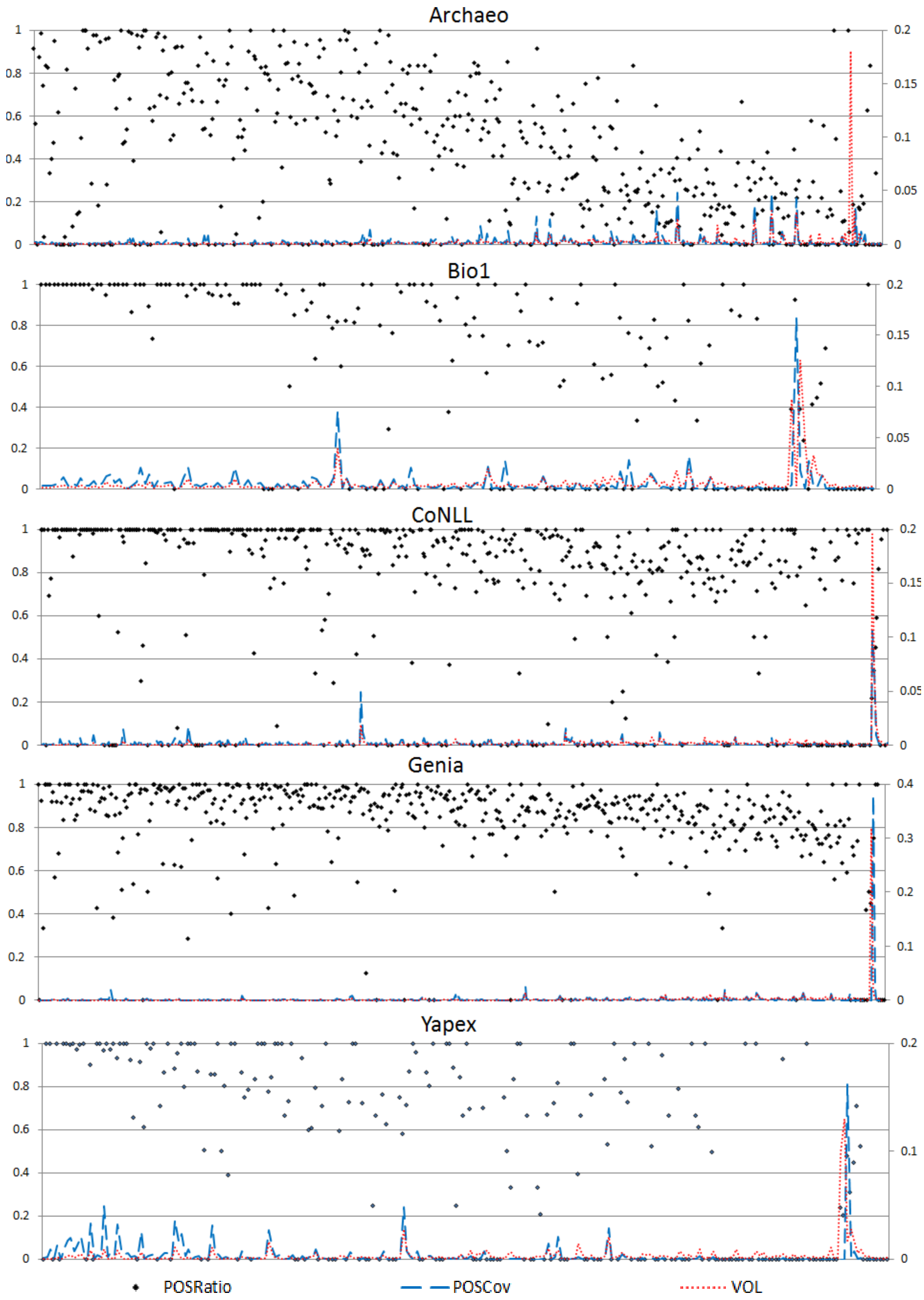


Figure 7.7. Feature analysis for RJ05 based method with the RIDF measure (POSRatio, POSCov, VOL)

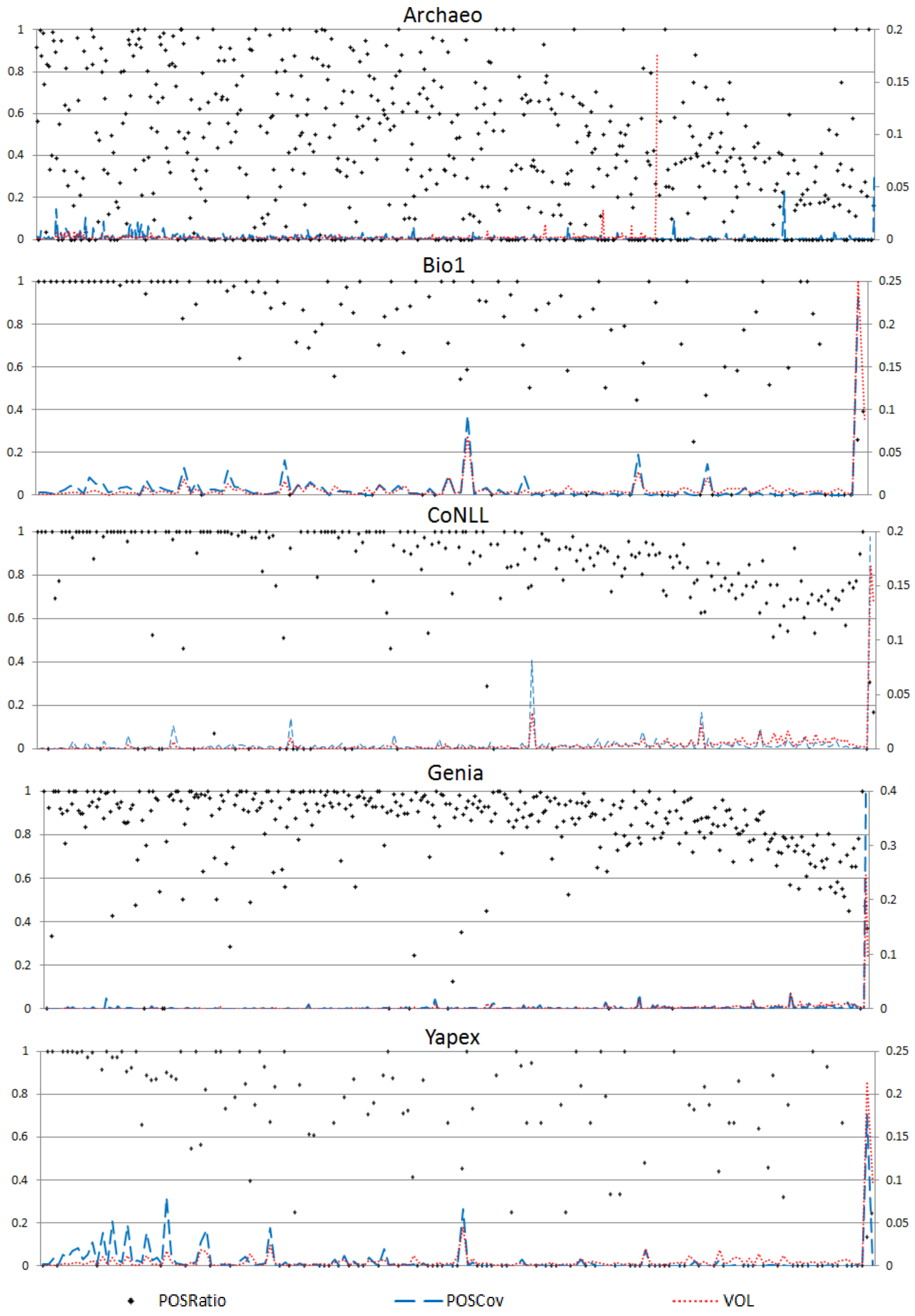


Figure 7.8. Feature analysis for RJ05 based method with the burstiness measure (POSRatio, POSCov, VOL)

mple, on the Bio1 dataset, the top 3 most informative RIDF scores have corresponding POSRatio values of 1.0, suggesting that all the instances represented by these features are positive. However, these instances only accounted for 0.13%, 0.11% and 0.11% of all data (VOL values) respectively, and 0.3%, 0.27% and 0.25% of all positive instances (POSCov) respectively. This makes these features insignificant and less informative to the learner.

The variable *k*, and an alternative feature extraction function – One potential limitation of the *topcat* based methods is the optimisation of *k* for the gazetteer generation process. An alternative method that overcomes this issue would be to normalise document-specific topicality scores by the mean or median scores to a comparable range and use the normalised scores as features – the same approach used by RJ05. To test this, further experiments are performed. Given a document, the topicality scores of a word given by each of the *topcat* measures are normalised against the mean of the scores of all words from that document. The normalised scores are then used as features and added to the baseline. However, as shown in Table 7.7, its effects are rather inconsistent. In most cases it has caused reduction (-4.3 – -0.5) in the overall learning accuracy, which could be due to overfitting training data.

	TF	**TFIDF**	**WD**	**CIR**
Archaeo	-0.5	1.3	1.1	0.9
Bio1	-1.5	-2.0	-2.9	-1.0
CoNLL	-0.5	-0.6	-1.8	-0.2
Genia	-1.3	-0.5	-4.4	-0.9
Yapex	1.4	-0.6	-1.3	1.4

Table 7.7. Absolute changes (F1) to the baseline using normalised topcat scores as features

Further, to understand if and how the value of *k* can affect the learning accuracy, experiments of *topcat* based methods are rerun with different settings of *k* = 5, 10, 15, 25, and 30. In general, it is found that a too small (≤5) or too large (≥25) value of *k* have both harmed the learning accuracy. Intuitively, the first case creates only a few bins of very large size that may not discriminate well. The second case creates many small bins of very small size that may not generalise well. However, the differences given by *k* = 10 or 15 or 20 are insignificant. In most cases, the difference is less than 0.3%. Therefore, it is believed that there is no strong motivation for tuning *k* for particular datasets.

7.7 Conclusion

This chapter has addressed automatically generating an alternative form of untyped gazetteers for NER. Alternative gazetteers group related terms without explicitly defining the type or categories of the groups as the type-oriented gazetteers do. They are useful to NER because the correspondence between the gazetteer labels and the named entity types can be learnt automatically based on training data. The majority of methods of alternative gazetteer generation are unsupervised, and often based on clustering approaches. This chapter explored a different direction that is rarely studied – gazetteer generation based on the hypothetical relation between topic-oriented words and named entities.

While the hypothesis that topic-oriented words can be used to predict named entities is not new, existing methods have proposed to assess topicality of words by informativeness measures. Informativeness is typically evaluated based on word's distributional patterns in a corpus and is therefore defined with respect to a collection of documents. As shown in this study, they can mis-represent topicality and harm NER learning accuracy. Instead, this study argued that topicality should be defined with respect to specific documents, and proposed to use the relevance measures widely used in IR tasks as a proxy for measuring topicality. Furthermore, to exploit word topicality in NER, a simple equal interval based binning approach is applied to group words based on their level of topicality, which effectively creates document-specific gazetteers based on the topicality of words. These can then be used for learning statistical NER models.

The proposed method was submitted to a comprehensive and comparative evaluation against several methods based on informativeness measures using five datasets covering three domains. The results have shown that it consistently improves the baseline by between 0.9 and 3.9 points of F-measure on all datasets. It is particularly useful to specialised domains such as archaeology and biomedicine, where NER is much harder and often requires domain specific external resources that can be expensive to build. On the contrary, methods based on informativeness have shown unstable performance with often damaged accuracies. This has confirmed that topicality measured locally specific to documents is a more effective and generalisable feature to NER. Further analyses have shown a highly non-linear long-tailed relation between topicality and NER classification decisions, which is well captured by the topicality measures and reflected by the simple binning based approach to gazetteer generation.

A number of directions will be further explored in the future. Firstly, other topicality measures will be explored. Topic-oriented words can be also related to keywords. Therefore, keyword extraction methods such as Hulth (2003) and Mihalcea and Tarau (2004) may be adapted to this task. The focus of these comparative studies will be identifying the strength and weakness of different methods and eventually proposing novel measures of topicality. Secondly, other methods of gazetteer generation will be explored. Given the non-linear long-tailed distributional patterns of topic-oriented words over named entities, a simple approach would be to apply exponential or logarithmic functions to normalise topicality scores to comparable ranges. These as well as novel methods will be explored in the future.

Part IV – Resolving Ambiguities

This part addresses the third research question concerning automatically resolving ambiguous entity names. Chapter 8 discusses measures of lexical semantic relatedness, which are the enabling technique to the disambiguation method proposed in this thesis; Chapter 9 discusses Named Entity Disambiguation based on measures of semantic relatedness.

8 Lexical Semantic Relatedness

PREFACE

This chapter presents a study on lexical semantic relatedness methods, which is an essential technique to enable the Named Entity Disambiguation approach to be introduced in the next chapter. Lexical semantic relatedness describes the strength of the semantic association between two terms or concepts. It is an enabling technique to many complex Natural Language Processing tasks. This chapter is divided into six sections. Section 1 gives an introduction to the research field. Section 2 presents a comprehensive literature review with remarks regarding this research. The goal of this review is to bridge and connect studies carried out in different sub-areas and create a first-point reference to researchers and practitioners in this field. Section 3 discusses the hypothesis behind this work and Section 4 introduces a novel approach that exploits multiple knowledge resources for measuring lexical semantic relatedness. Section 5 presents experiments and discussions. The final Section (6) concludes this chapter.

8.1 Introduction

Lexical semantic relatedness describes the strength of the semantic association between two terms or concepts. This strength of association is typically evaluated based on certain background information of terms or concepts. It is often a pre-processing step to many NLP applications such as Word Sense Disambiguation (Leacock and Chodorow, 1998; Han and Zhao, 2010), Named Entity Recognition (Kliegr et al., 2008), sense clustering (Matsuo et al., 2006; Bollegala et al., 2007), and Information Retrieval (Finkelstein et al., 2002). For this reason, a significant number of methods have been introduced in the past years.

It has been noted that, despite the availability of such abundant literature, a comprehensive review of the studies and their connections is lacking. First, although most of the pre-2006 studies based on WordNet as the background information resource have been thoroughly discussed in Budanitsky and Hirst (2006), a large number of new methods – particularly those based on collaborative resources such as Wikipedia and Wiktionary – have been proposed but their connections with previous research are rarely discussed. Second, a great number of semantic relatedness methods have been introduced in the biomedical domain, an area where semantic relatedness is considered an important technique for knowledge discovery (Pesquita et al., 2009). However, the work in the general domain and the biomedical domain is rarely communicated. Third, a comparative analysis of background information resources for the task is unavailable. These issues have caused obstacles in the research and application of lexical semantic relatedness. Firstly, it has been noted that near-identical methods have been introduced in different contexts, costing expensive research effort. Secondly, it has been difficult to compare and select the most appropriate methods for NLP applications; this study shows that from a practical point of view, the choice of lexical semantic relatedness methods can depend on many factors. However, this has been hardly discussed in the literature.

For these reasons, the first goal of this study is to present a comprehensive literature review to fill these gaps. First, background information resources widely used for lexical semantic relatedness are introduced and their characteristics are analysed. Then, different methods covering both the general and biomedical domains are discussed from a generic viewpoint, focusing on the rationales and the connections among different methods. Finally, the strengths and limitations of different types of methods are analysed and compared, which leads to a conclusive remark regarding the research and application of lexical semantic relatedness.

Furthermore, the literature review reveals that the vast majority of lexical semantic relatedness methods have employed a single source of background information of terms or concepts, while the literature has generally preferred a number of resources, among which, the most frequently used are WordNet, Wikipedia, and Wiktionary. It has been noted that, due to the different purposes of such resources, they encode different kinds of information of terms and concepts, and have different focuses. On the other hand, they often cover the same sections of vocabularies and concepts. This motivates the hypothesis that different background information resources can complement each other, which leads to the idea of combining them in a uniform framework for measuring lexical semantic relatedness.

Therefore, a new method is proposed. This method exploits three different background information resources – WordNet, Wikipedia, and Wiktionary – in a uniform framework for measuring lexical semantic relatedness. It is based on the idea of creating a joint feature representation of terms or concepts using the background information encoded in the three different resources, which ultimately improves the feature quality and outperforms a representation that is based on any one of the three resources. The method is thoroughly evaluated on 9 benchmarking datasets, including three datasets from the biomedical domain and four from the general domain. It has significantly outperformed the baselines that use each single resource, and also achieved higher accuracies on certain datasets when compared against state-of-the-art.

The remainder of this chapter is outlined as the following: Section 8.2 presents the literature of lexical semantic relatedness; Section 8.3 further discusses the hypothesis behind the proposed method; Section 8.4 introduces the new method of semantic relatedness; Section 8.5 presents experiments and discussion; Section 8.6 concludes this chapter.

8.2 Lexical Semantic Relatedness – A Survey

8.2.1 Terminology and Notions

In the literature on lexical semantic relatedness, the term '**semantic relatedness'** is often confused with three different but relevant terms: **semantic similarity**, **semantic distance**, and **distributional similarity**. S**emantic relatedness** essentially describes the strength of the semantic association between two concepts, or their lexical realisations. It encompasses a variety of relations between terms and their underlying concepts, including the classical relations such as hypernymy, hyponymy, meronymy, antonymy, synonymy; and any other 'non-classical relations' (Morris and Hirst, 2004) and 'implicit connections' (Zesch and Gurevych, 2010a). Semantic similarity is a specific case of relatedness, where the sense of relatedness is dependent on the 'degree of synonymy' (Weeds, 2003), which is usually accounted by classical relations. Terms or concepts that are semantically related are not necessarily similar, such as 'car' and 'fuel'. Another example is that antonyms are considered to be semantically related, such as 'beautiful' and 'ugly'; however, they are dissimilar. Computational applications typically require relatedness rather than similarity (Budanitsky and Hirst, 2006). For example, for sense disambiguation of the names in the sentence 'President Bush attended the opening ceremony of the Olympic Games in Beijing', it is more useful to know whether the referent entities of 'Bush', 'Olympic Games' and 'Beijing' are related rather than similar. The term semantic distance has been used in the literature to refer to the inverse of semantic relatedness or similarity. Concepts that are semantically similar or related are considered to be semantically close to each other, thus denoting a sense of distance. It is also worth noting that although it is generally agreed that semantic relatedness is *symmetric*, this is not always true for semantic similarity (Tversky, 1977). Asymmetric similarity is often perceived between a concept and its superclass concepts. Similarity from a concept to its superclass is usually considered greater than the opposite. For example, 'a pear is similar to a fruit' is more agreeable than 'a fruit is similar to a pear'. However, the literature has predominantly taken the assumption of symmetric semantic relatedness and similarity.

In addition, methods for measuring distributional similarity of words (Weeds, 2003) have been widely used as a proxy to address lexical semantic relatedness. In the literature, they are sometimes used interchangeably with *word similarity*, or *co-occurrence similarity*. Briefly, distributional similarity between two words is based on the extent to which the two words tend to occur in similar contexts. By this definition, distributional similarity does not strictly adhere to the notion of lexical semantic relatedness, and its application to assessing semantic relatedness between words has been controversial (Weeds, 2003; Budanitsky and Hirst, 2006). Despite these debates, substantial work has been carried out to develop and apply distributional similarity methods to address the issue of semantic relatedness and related tasks.

To generalise, measuring lexical semantic relatedness requires certain forms of 'background' information of terms or concepts, either as formally, explicitly defined lexical and semantic relations between terms and concepts in the case of semantic relatedness, or as implicit connections given by distributional context in the case of distributional similarity. The resources from which such background information can be obtained are referred to as **background information resources**. Furthermore, measuring lexical semantic relatedness often requires dealing with natural language polysemy. Theoretically, semantic relations are defined for concepts, while the lexical realisation of concepts – words or terms, to be used interchangeably in the following – can be ambiguous and used to refer to different concepts. This may not be an issue when distributional similarity is used as a proxy for this purpose since the background information is collected at lexical level. For semantic relatedness methods that truly take into account concept-level background information, relatedness between two words is generally approximated based on their underlying concepts.

8.2.2 Background information resources

As discussed before, measuring lexical semantic relatedness generally requires certain **background information** about concepts or terms. Such information is often encoded in structured and semi-structured **knowledge bases** that form a graph of concepts, which are lexicalised and indexed by their lexical forms. Concepts are interconnected by **links** or **edges** that denote a certain sense of semantic relations. Several terms are used by the literature when describing knowledge bases: '**taxonomy**' refers to a hierarchical structure, in which nodes are organised by the generalisation-specialisation relationship; '**ontology'** refers to a taxonomic structure enriched with other semantic relationships such as antonymy and synonymy, and class properties or attributes; '**semantic graph**' or '**semantic network**' refers to any kinds of concept graphs connected by any semantic or loose associative relations. In an analogy, distributional similarity can be considered to employ background information of terms in the form of their contexts, which are derived from a large corpus. This source of background information will be referred to as **unstructured corpora**, in the sense that the documents do not provide sense-tagging of words or explicitly organise words or concepts in a structured way encoding their associations.

8.2.2.1 Knowledge bases

Examples of knowledge bases include dictionaries, thesauri, wordnets, and encyclopaedic resources. Some methods, especially earlier ones, have employed dictionaries and thesauri, such as Morris and Hirst (1991), Kozima and Furugori (1993), and Jarmasz and Szpakowicz (2003). Most methods employ wordnets, and encyclopaedic resources. Among these, the most frequently used general purpose knowledge bases include WordNet (Fellbaum, 1998), Wiktionary and Wikipedia.

WordNet has been one of the most popular background knowledge bases for the studies of semantic relatedness. It is a lexicalised ontology of English words. It groups nouns, verbs, adjectives and adverbs into **synsets**, each expressing a distinct concept. Searching for a word in WordNet may return multiple synsets corresponding to different senses or concepts. Each concept in WordNet is provided with a short definition called **gloss**, and is connected to other concepts by a set of semantic relations depending on the word class, such as hypernymy and meronymy for nouns, hypernymy and entailment for verbs, synonymy and antonymy for adjectives.

Since WordNet is designed to provide complete coverage of common, open-class English words, it has little or no coverage of vocabularies from specialised domains, and very limited coverage of proper nouns. This may hinder its application to domain specific contexts and tasks required to deal with proper nouns (Hirst, 1998; Strube and Ponzetto, 2006).

Wiktionary is a multi-lingual free dictionary built and maintained by collaborative effort. It has many commonalities with WordNet: each entry in Wiktionary is an article page about a term and distinguishes one or more word classes. Each word class has one or more senses that correspond to concepts. Each concept is provided with a short definition (similar to WordNet gloss) often accompanied by example sentences. Wiktionary also defines lexical semantic relations that are available in WordNet, such as hypernymy, hyponymy, coordinate terms, synonymy and antonymy. In addition, it encodes information such as alternative forms and etymology at the level of terms; and derived terms and translation at the level of word class.

Meyer and Gurevych (2010) compared resource coverage of English Wiktionary against WordNet. They showed that in general, Wiktionary encodes twice the amount of words than WordNet. Wiktionary outnumbers WordNet by covering a broader range of word classes, a larger number of abbreviations, numerals, symbols and proper nouns. Wiktionary covers a large number of word inflectional forms (nearly 30% of Wiktionary) and neologisms, which are unavailable in WordNet. However, about half the amount of Word-

Net lexicons are missing in the English Wiktionary, of which 50% are found to be Latin words belonging to scientific domains. Meyer and Gurevych also studied the word sense distribution over different word classes and sense alignment in the two resources. They concluded that the distribution in both resources is very similar, despite that on average WordNet encodes more word senses for verbs while Wiktionary encodes more word senses for nouns. Wiktionary has better coverage of slang-related and domain-specific senses, as well as word senses for rarely used terms. Using a sample corpus, they discovered that both knowledge bases share many word senses for words with a medium language frequency; while Wiktionary encodes a large number of word senses for words with a high frequency.

Navarro et al. (2009) showed that Wiktionary suffers from issues such as uneven density of knowledge, imbalanced coverage of different languages and a sparse synonym network. For example, the lexical-semantic information is not always encoded for any words belonging to the same word class. And the amount of encoded information is largely imbalanced.

Wikipedia has been a popular choice of knowledge base in recent work on lexical semantic relatedness. It is a multi-lingual encyclopaedia created and maintained by collaborative effort. It has been briefly mentioned previously in Chapter 6. In general, article pages in Wikipedia describe a vast amount of proper nouns (or entities) and concepts. They do not have a dedicated section of definitions similar to WordNet gloss or Wiktionary definitions. However, research has assumed that the first paragraph of a Wikipedia article provides definitional details and the first sentence often gives a short definition. Wikipedia articles are hyperlinked. The links are not typed; they denote rather general semantic associations and thus create a loosely connected semantic graph. Articles are tagged with multiple category labels, which are general concepts organised in a hierarchical structure, creating a category tree generally resembling the broader and narrower sense of relation between categories. In addition, Wikipedia groups synonyms and aliases using the mechanism of ‘redirect’ – an article page may be linked to a number of alternative names denoting a sense of synonyms, which when searched, will always be redirected to the uniform article page. Polysemous names and phrases are encoded in separate ‘disambiguation’ pages, which list different meanings with links to corresponding article pages. For many Wikipedia pages, a tabular ‘infobox’ of additional metadata is available, usually presenting fact-like information of certain types that are common to similar articles.

Wikipedia offers several advantages over WordNet and Wiktionary. Most of all, it covers a substantial amount of proper nouns and concepts, as well as domain-specific vocabularies. As mentioned before, Holloway et al. (2007) showed that by 2005, Wikipedia already contained 1,069 disconnected clusters of categories of articles each denoting a distinctive subject. Milne et al. (2006) showed that in the domain of food and agriculture, Wikipedia provides excellent coverage of domain terminology and semantic relations that rivals a professional thesaurus. Halavais (2008) in an analysis of topical coverage of Wikipedia by comparing printed books against Wikipedia articles concluded that the coverage of topic-specific knowledge is generally good in Wikipedia. Additionally, the denser connections between article pages and categories as well as longer content also imply richer lexical semantic information.

However, Wikipedia does not annotate article hyperlinks by semantic relations. Likewise, the hierarchical structure of category tree rather represents a loose folksonomy than a strict taxonomy (Strube and Ponzetto, 2006; Ponzetto and Strube, 2011), since it contains relations such as meronymy. Relations are not explicitly defined. Also, due to the encyclopaedic purpose of Wikipedia rather than a lexical knowledge base, its content may be biased towards specialised concepts and instances rather than lexicographic senses of words. In particular, verbs are largely underrepresented, as shown in Zesch and Gurevych (2010a). For example, the closest entry matching the word ‘win’ and its verb sense is the article on ‘victory’, while all the other articles describe domain-specific concepts or entities referred to by the same word.

8.2.2.2 Unstructured Corpora

Unstructured corpora can be considered the background information resource for distributional similarity methods. Some semantic relatedness methods also employ unstructured corpora in certain ways. Unlike knowledge bases, unstructured corpora do not provide sense tagging of words or define lexical semantic relations in an explicit way. As a result, background information is collected at the level of terms rather than concepts, and connects terms in a rather implicit way. With distributional similarity methods, this is often in the form of textual contexts of terms or their co-occurring behaviours within certain contexts based on a sufficiently large collection of documents. The hypothesis is that terms that tend to occur in similar contexts are similar.

Some recent approaches have proposed mining lexical semantic network of terms from unstructured document collections, such that methods that use structured knowledge bases can be applied. For example, Harrington (2010) proposed to parse a corpus to build connected graphs of words based on their syntactic relations, and exploit the link structure using graph-based algorithms to measure semantic relatedness between words. Details of these will be discussed later in Section 8.2.3.

A large number of general purpose document collections have been compiled for the use in NLP research and applications. The most often used include the Brown corpus (Kucera et al., 1967), the British National Corpus (The British National Corpus, 2007), the Penn Treebank corpus (Marcus et al., 1993), the Reuters corpus (Rose et al., 2002), and the newswire articles published by Associated Press (AP newswire articles) and Wall Street Journals (WSJ), some of which are archived by Harman and Liberman (1993). Each of these compiles different document resources of various topics to the order of millions of words.

Recently, an increasing number of approaches (Chen et al., 2006; Matsuo et al., 2006; Cilibrasi and Vitanyi, 2007) have explored the Web as the source of unstructured documents for distributional similarity methods. Typically, queries are composed based on the words in question and are used to retrieve documents that are likely to contain co-occurrences of the words. Compared to pre-compiled document collections, such approaches can benefit from the sheer size of the Web, which generally provides a better coverage. However, they may also suffer from limitations inherited from search engines, such as limited query syntax, and potentially misused counting that is intended for number of pages rather than instances (Kilgarrif, 2007).

8.2.2.3 Biomedical Resources

Lexical semantic relatedness has been a focus of research in the biomedical domain, since it is often a method for creating new lexical knowledge resources and discovering new knowledge (e.g., relations) from data. Studies in this domain often prefer biomedical resources due to their comprehensive coverage of domain specific knowledge. They can also be divided into knowledge bases and unstructured corpora. Biomedical knowledge bases are usually a structured vocabulary of technical terms, which often denote unique, specialised concepts.

The most frequently used knowledge base in this domain is the *Gene Ontology – GO* (The Gene Ontology Consortium, 2000), which defines an ontology of terms representing gene product properties and used by the majority of semantic relatedness methods in the biomedical domain, each term is usually assigned a unique identifier; a term name which is a word or phrase; a short definition similar to that of WordNet gloss; and a name space label indicating the sub-domain that the term belongs to. Terms are inter-linked with other terms within or across sub-domains, by relations such as synonymy, hypernymy, meronymy, and domain specific relations such as regulation. The majority of studies in the biomedical domain exploit the GO as the background information resource (see Pesquita et al., 2009). Also, many have exploited the semantic similarity between GO terms to assess the similarity between genes or gene products annotated by these terms. The idea is that genes are similar if they share same properties, which are described by GO terms. Due to the similarity in its structure with WordNet, many of these methods have adapted WordNet-based methods or are developed based on similar rationales.

The *MeSH* (Medical Subject Headings, 1999) is a comprehensive controlled vocabulary resource for indexing articles and books in the biomedical domain. It is structured as a hierarchy of ‘descriptors’, each of which is essentially a subject heading designed for the purpose of indexing. A descriptor is accompanied with a definition of the description; a list of synonyms or very similar names known as ‘entry terms’, in the sense that the same descriptor can be looked-up using such terms. This is similar to the Wikipedia redirect mechanism. The descriptors are split into 16 categories representing sub-topics. Descriptors of each category are organised as sub-hierarchies from most general to most specific levels. A MeSH descriptor may appear in multiple places in a hierarchy.

The *SNOMED-CT* (SNOMED CT, 2002), standing for ‘Systematized Nomenclature of Medicine – Clinical Terms’ is a controlled vocabulary of medical terminology. The terminology is organised into 13 hierarchies based on different topics and contains over 1 million concepts. Each concept is assigned a unique identifier, and provided with multiple ‘descriptions’, each of which is a name used to refer to the concept. Thus they are similar to the ‘entry terms’ in MeSH. The descriptions are divided into three types: a unique ‘Fully Specified Name (FSN)’, a ‘preferred term’, and one or multiple synonyms. Preferred terms and synonyms are names that are not unique to the concept, but can be shared by multiple concepts. Concepts in SNOMED-CT are organised as taxonomies following the hypernym relation. In addition, a number of domain-specific relations (e.g., ‘due to’, ‘causative agent’) are defined to connect concepts.

The *UMLS – Unified Medical Language System* (Bodenreider, 2004) is a resource that maps a wide range of biomedical knowledge bases. It contains three main knowledge bases: Metathesaurus, Semantic Network and SPECIALIST lexicon. The Metathesaurus contains over 1 million biomedical concepts and 5 million concept names integrated from over 100 knowledge bases (source vocabularies) including GO, MeSH, and SNOMED-CT. One of its main purposes is to group different names for the same concept from different source vocabularies. Each concept is assigned a unique identifier, one or multiple concept names, and pointers to their source vocabularies. Many relationships are encoded between concepts, including the hypernymy, meronymy and synonymy relations. Relations encoded in source vocabularies are also retained. The Semantic Network defines a set of subject categories called semantic types, such as organisms, biologic functions and chemicals. They are organised into a hierarchy representing the hypernymy relation, and also interlinked by many other non-hierarchical relations. The SPECIALIST lexicon is intended to be an English lexicon of both common English words and biomedical vocabulary. An entry is defined for each word or phrase, and records the syntactic, morphological and orthographical information that can be used by NLP systems.

Furthermore, two other biomedical knowledge bases used by some studies are the *Human Phenotype Ontology (HPO)*, a vocabulary of approximately 9,000 terms referring to phenotypic abnormalities encountered in human disease (Kohler et al., 2009), and the *ChEBI (Chemical Entities of Biological Interest)*, an ontology of molecular entities focused on small chemical compounds (Degtyarenko et al., 2007).

In the biomedical domain, corpora are usually pre-processed, with distributional statistics stored in relational databases. The statistics are gathered for domain-specific vocabularies, usually terms defined in biomedical knowledge bases such as the GO. Examples of such databases include the *SWISS-PROT/UniProtKB* (Boutet et al., 2007), the *Saccharomyces Genome DB* (Cherry et al., 1998), and the *Gene Ontology Annotation database* (Camon et al., 2004). Unlike counting word frequencies in a general corpus, for many of these databases the statistics are not based on the lexical realisation of concepts, but according to their usage in gene product annotations. Typically, biomedical terms are used to annotate mentions of gene products in the corpus if they represent a property of the gene products. Thus the frequency of a term is determined by the frequencies of the gene products it annotates, and terms are said to co-occur if they are used to annotate same gene products.

Table 8.1 below summarises different knowledge bases (both generic and domain specific) that have been widely used in the literature.

8.2.3 Lexical Semantic Relatedness Methods

This section discusses different methods for measuring lexical semantic relatedness, including distributional similarity methods that have been used as proxy for the task. The studies from both the general and biomedical domains will be included. Due to the significant amount of studies available in this field, the discussion will focus on the rationale that connects different methods, rather than detailing every single method. In the formulae, **SemRel** denotes semantic relatedness, **DistSim** denotes distributional similarity. Many methods specifically address semantic similarity while some measure semantic distance as the inverse of relatedness or similarity. They are denoted as **SemSim** and **SemDist** respectively. Also, some methods apply to concepts, while others apply to words or terms. To distinguish these cases, *c* denotes a concept, *w* denotes a polysemous word, phrase or term. Concepts that are represented by the same word are denoted as $c \in C(w)$. Thus *SemRel*(c_1, c_2) denotes semantic relatedness between concepts c_1 and c_2 while *SemRel*(w_1, w_2) denotes relatedness between words w_1 and w_2.

Given a pair of terms which can be polysemous, semantic relatedness methods typically calculate relatedness between their underlying concepts (i.e., *C(w)*), and then adopt a strategy to derive an aggregated score for their lexical expressions. The literature largely differs in terms of the approaches to measuring concept relatedness, while the method for deriving term relatedness is generally based on some rather *de facto* practice. Three techniques are commonly used for this purpose: (1) maximum pairwise concept relatedness, which assigns the maximum relatedness score obtained for every pair of $c_1 \in C(w_1)$ and $c_2 \in C(w_2)$; (2) average pairwise concept relatedness, which takes the average of relatedness scores obtained for every concept pair; and (3) sum of pairwise concept relatedness. To avoid repetition, the remainder of this section will not explain for each individual method its choice of method for deriving word relatedness from concept relatedness, unless a different technique has been used.

Background Information	**Coverage of Knowledge**		**Structure and content**
	Focus of Coverage	**Limited Coverage**	
WordNet (GN)	Common English words including nouns, verbs, adjectives and adverbs	Limited coverage of specialised vocabularies, very few proper nouns	- gloss: a definitional description - synsets are linked by semantic and lexical relations; such as hypernymy, meronymy, synonymy, coordinate etc.
Wiktionary (GN)	Nouns, verbs, adjectives, adverbs, other word classes; abbreviations, neologisms, inflectional words	Covers some specialised vocabularies and more proper nouns than WordNet	- definitional description with example sentences - most WordNet lexical and semantic relations are encoded, plus additional information such as etymology
Wikipedia (GN)	Proper nouns and concepts, covering a broad range of topics	Not focusing on lexicographic senses of words, which can be underrepresented (e.g., verbs)	- dedicated content pages for concepts and proper nouns - heavily hyperlinked article pages - hierarchical category tree for classification of articles - redirect system grouping synonyms and aliases
GO (DS)	Concepts and vocabularies representing gene product properties.		- definitional description - hypernymy, meronymy, synonymy, and domain specific rela-
MeSH (DS)	Descriptors denoting unique concepts of the medical domain		- definitional description and synonymous terms, which are names used to refer to the same descriptor concept - hierarchical structure representing generalisation/specification
SNOMED-CT (DS)	Concepts and vocabularies for the clinical domain		- synonymous terms (similar to those in MeSH) - concepts linked by hypernymy and domain specific relations
UMLS (DS)	Concepts and vocabularies integrated from other biomedical resources		- Metathesaurus maps concepts across over 100 knowledge bases - Semantic Network defines a category tree to classify concepts - SPECIALIST lexicon defines syntactic, morphological and or-
HPO (DS)	Concepts and vocabulary of phenotypic abnormalities		- definitional descriptions, hypernymy, meornymy and synonymy relations - cross-reference to UMLS concepts
ChEBI (DS)	Molecular entities focused on small chemical compounds		- entities linked by hypernymy, meronymy and synonymy relations

Table 8.1. Summary of frequently used knowledge bases

Semantic relatedness – particularly similarity – is often measured with respect to a taxonomic structure, denoted by *T*. Figure 8.1 shows an example taxonomy of arbitrary food concepts. A number of common notions shared by these methods are defined below.

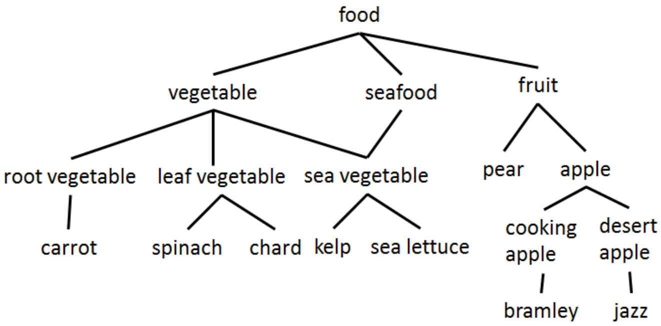


Figure 8.1. A sample taxonomy

- **node** – a node corresponds to a concept in the semantic graph.
- **edge** – a single link connecting two adjacent nodes in a semantic graph. The type of an edge is defined as the relation it represents.
- **root** – the root node in *T*, in this case, the node 'food'.
- **parent(c)** – returns the concept(s) that immediately subsumes the concept in the taxonomy. For example, *parent*(*sea vegetable*) = {*vegetable*, *seafood*}.
- **subsumer(c)/ancestor(c)** – returns the **subsumers or ancestors** of a node in a recursive manner. For example *ancestor {kelp}* = *{vegetable, seafood, food, sea vegetable}.*
- **child(c)** – returns the concept(s) that are immediately subsumed by the concept in the taxonomy. For example, *child(food) = {vegetable, fruit, seafood}.*
- **descendant(c)** – returns the descendants of a node in a recursive manner. For example *descendant {vegetable}={leaf vegetable, spinach, chard, sea vegetable, kelp, sea lettuce, root vegetable, carrot}.*
- **cs(c$_1$, c$_2$)** – returns the shared or common subsumers of concepts c_1, c_2. In mathematical terms, $cs(c_1, c_2) = subsumer(c_1) \cap subsumer(c_2)$. In some literature, this is called *common ancestor*.
- **lcs(c$_1$, c$_2$)** – returns the **least common subsumers** of concepts c_1, c_2. There are different definitions of *lcs* in the literature. The majority of these adopt a definition by Resnik (1995), which defines *lcs* as the member in *cs(*c_1, c_2*)* at the lowest level of the taxonomy. Thus *lcs(chard, kelp)* = *{vegetable}*, and *lcs(kelp, sea lettuce)={sea vegetable}.* In some literature this is called *most specific subsumer* (Budanitsky and Hirst, 2006), or **lowest common ancestor** (Schickel-Zuber and Faltings, 2007). Following this definition, theoretically two concepts may have multiple *lcs*, which could happen if each concept has multiple parents and two of these are shared between them. In practice, this is not very common. For example, a concept in WordNet has on average 1.03 parents ac-

cording to Schickel-Zuber and Faltings (2007). Despite this low possibility, the authors proposed a revised definition to resolve such cases, and their work is the only one that applies a different definition of *lcs*. This will be introduced in Section 8.2.3.1.

Based on the rationales of semantic relatedness methods, the literature can be divided into *path based, Information Content (IC) based, gloss based, feature vector based, distributional similarity methods*, and *hybrid methods* that combine multiple purebred measures in certain ways.

8.2.3.1 Path based methods

The fundamental rationale behind path based methods is that the relatedness between concepts can be determined based on their distance, or the length of the paths connecting them in a semantic graph following a given type of edges, or relation. The length of a path is typically calculated by counting the number of edges or nodes along the path. For this reason, they are also referred to as **edge based methods** (Pesquita et al., 2009).The majority of these methods exploit taxonomic links in a semantic network; therefore, they measure similarity rather than relatedness.

The earliest work of this type is Rada et al. (1989), which measures semantic relatedness between two concepts using the shortest path length in a semantic graph. When applied to taxonomic structures, the shortest path is typically the one that connects the concepts by their *least common subsumer (lcs)* – the nearest concept that subsumes both concepts in the taxonomy (e.g., *lcs(pear, sea vegetable) = {food}*). This simple method has been used by a number follow-up studies using other knowledge bases such as Jarmasz and Szpakowicz (2003), Gentleman (2005), and Bhattacharya et al. (2010). Hirst and St-Onge (1998) suggested limiting the length of a valid path and discriminating the change of directions (*upward* such as hypernymy, *downward* such as hyponymy and *horizontal* such as antonymy) along the path. The motivation is that the strength of relatedness correlates negatively with path length and frequency of changes of direction along the path. Yang and Powers (2005) proposed to calculate similarity as the product of edge weights rather than their sum, where different types of edges are treated with different weights. The rationale is that they contribute differently to semantic similarity.

Notions of specificity, depth and density

A widely recognised limitation of these methods is that they do not account for the specificity of nodes in a taxonomy. Typically, edges at any levels in the taxonomy are assumed to represent uniform length, and that nodes are distributed uniformly across the hierarchy. These assumptions are rarely true in practice, particularly for biomedical ontologies (Pesquita et al., 2009). For example, in Figure 8.1, the same path length at different levels (e.g., the shortest path between ‘seafood’ and ‘fruit’, and that between ‘spinach’ and ‘chard’) can denote different distances.

To overcome these issues, many have incorporated the notion of **depth** or **density,** or both, to account for specificity. The motivation is that two nodes are semantically closer if they reside deeper in the hierarchy or are more densely connected locally. Given a node *c* in a taxonomy *T*, the depth of *c* denoted by *depth(c)* is usually the number of nodes along the longest path between *c* and *root*, and the depth of the taxonomy is the depth of the deepest node. The density of a node denoted by *den(c),* is usually defined as the number of its child nodes, or the number of its sibling nodes. In Figure 8.1, *depth(spinach) = 4, depth(T)=5*. The density of *root vegetable* is 1 following the first definition, or 2 following the second definition.

A recent study by Wang and Hirst (2011) has shown these classic definitions can misrepresent depth and density. In terms of depth, the classic definition assumes a linear function that returns the depth of a node as an ordinal integer. However, experiments have shown that the “notion of depth is relative to the distribution of number of nodes over depth value”, and the distribution of nodes over depth conforms to a normal distribution. Also, there is “no definitive, sufficient and necessary relation between depth and similarity”. In particular, examples are given to show that semantically similar words (and their underlying concepts) are not necessarily deeper in the hierarchy. More experiments were carried out to show that the correlation between density (as by number of sibling nodes) and similarity is found to be even weaker. The distribution of density values is found to generally follow the Zipf’s law, with more than 90% of nodes in WordNet having density values not greater than 3. This means that for the majority of concepts in WordNet, there are “only three integer values to distinguish the varying degrees of similarity”. Wang and Hirst proposed new measures of depth and density that reflect their true distributional nature. These are denoted as $depth_{WH}$ and den_{WH} respectively:

$$depth_{WH}(c) = \frac{\sum_{c' \in T} |\{c' : depth(c') \leq depth(c)\}|}{|T|}$$ **Equation 8.1**

$$den_{WH}(c) = \begin{cases} 0, & c = root \\ \frac{\sum_{c' \in subsumer(c)} den_{WH}(c')}{|subsumer(c)|} + den(c), & else \end{cases}$$ **Equation 8.2**

Due to the recency of this work, existing path based methods still exploit the classic definition of depth and density.

Addressing specificity in path based methods

Several studies (Ye et al., 2005; Yu et al., 2005; Lei and Dai, 2006) have proposed methods that simply determine semantic similarity using the *depth* of concepts or their *lcs*. Similar to methods purely based on path lengths, these may also lead to spurious predictions since they would tend to suggest concepts at the same level of a taxonomy (i.e., same depth) are similar, which is not necessarily true (e.g., pair-wise similarity between 'spinach', 'chard', and 'cooking apple' tend to be similar using the taxonomy in Figure 8.1). The majority of path based methods combine path length with depth or density, or both in certain ways. Four widely cited methods in this direction are Sussna (1993), Wu and Palmer (1994), Jiang and Conrath (1997) and Leacock and Chodorow (1998). Leacock and Chodorow (1998) normalised the shortest path length between two nodes by the depth of the taxonomy:

$$SemSim(c_1, c_2) = -log \frac{\delta(c_1, c_2)}{2 \cdot depth(T)}$$ **Equation 8.3**

where $\delta(c_1, c_2)$ denotes the shortest path length between two concepts, *depth(T)* is the depth of the taxonomy, which is effectively the maximum concept depth in the taxonomy. Wu and Palmer (1994) defined similarity between two nodes based on the depth of their *lcs* normalised with respect to the shortest path connecting them:

$$SemSim(c_1, c_2) \\ = \frac{2 \cdot depth(lcs(c_1, c_2))}{\delta(c_1, lcs(c_1, c_2)) + \delta(c_2, lcs(c_1, c_2)) + 2 \cdot depth(lcs(c_1, c_2))}$$ **Equation 8.4**

Sussna (1993) and Jiang and Conrath (1997) combined both the depth and density factors. In Sussna (1993), each edge connecting two nodes is assigned a weight, the calculation of which treats the edge as a combination of two unidirectional links, one leaving from a node to another and one as the inverse of this. The weight of each directed link is dependent on two factors: (1) the arbitrary weight for the relation it represents and (2) the number of links of the same type leaving the same node (i.e., a measure of density). Then the path connecting two nodes is weighted following this method, and normalised by the depth of the two nodes. The method by Jiang and Conrath (1997) is usually classified as IC based although the rationale is partly related to the path between concepts; this will be further discussed in the next section. In addition, Pekar and Staab (2002) and Liu et al. (2007) also proposed methods that are highly similar to that of Wu and Palmer, based on similar rationales. According to Liu et al., the principle behind Wu and Palmer's method can be considered as computing semantic similarity based on the ratio of two concepts' common features and different features, which can be quantified by the depth of their *lcs* and the shortest path length respectively. Another study by Li et al. (2003) concluded with a measure that linearly combines path length with the depth of *lcs*, each assigned with a scaling factor to control the contribution of each.

Al-Mubaid and Nguyen (2006) proposed to combine the shortest path length and the notion of **common specificity** with different weights using a log function. Common specificity of two concepts is determined as the difference between the depth of their *lcs*, and that of the cluster – essentially the branch, a sub-hierarchy – in the taxonomy that contains both concepts:

$$CSpec(c_1,c_2)=D-depth(lcs(c_1,c_2))$$

Equation 8.5

where *D* is the depth of the cluster containing both concepts in *T*. The intuition is that "the smaller the common specificity score, the more they share information, and thus the more they are similar". Then similarity is defined as a non-linear combination of the shortest path length and common specificity:

$$SemSim(c_1,c_2)=log((\delta(c_1,c_2)-1)^{\alpha}\cdot CSpec(c_1,c_2)^{\beta}+k)$$

Equation 8.6

Wu et al. (2006a) proposed a measure called **relative specificity similarity** which takes into account three different factors: the specificity of the *lcs* of two concepts in the taxonomy (α), the generality of each concept (β), and the local distance between the two concepts relative to their *lcs* (*γ*). The method for calculating *α* is initially introduced in Wu et al. (2005), which effectively has equivalent effect to *depth(lcs)*. The *β* factor assumes that the similarity between two concepts is subject to the one that is more general and equates the path length between the concept to its furthest leaf node. The *γ* factor evaluates the distance between each concept to their shared *lcs*. The final semantic similarity balances all three factors and the depth of the taxonomy:

$$SemSim(c_1,c_2)=\frac{depth(T)}{depth(T)+\gamma}\cdot\frac{\alpha}{\alpha+\beta}$$ **Equation 8.7**

which returns 0 when *α* = 0, indicating that two concepts do not share any ancestors; and returns 1 when *β* =0 and *γ* =0, indicating that two concepts are the same node.

Tsatsaronis et al. (2010) proposed a method that takes into account the weighted *path length* as well as the depth of all nodes along the path, which they call *path depth*. A major difference is that a path can be a combination of different types of edges including both taxonomic and non-taxonomic relations, while different types of relations are given different weights. The path length, which is captured under the notion of **compactness**, is computed as the product of weighted edges connecting the nodes along the path:

$$compactness(\,p(c_1,c_2)\,)=\prod_i^l w(e_i)$$ **Equation 8.8**

where e_1, e_2... e_l are the edges along the path connecting c_1 and c_2, and $w(e_i)$ is a weighting function that assigns a real valued weight to an edge based on the type of relation it represents. The weights associated with each relation are designed to promote those that denote stronger semantic connections. The path depth, which is captured under the notion of **semantic path elaboration,** is calculated as the product of weighted depth of the nodes along the path:

$$spe(\,p(c_1,c_2)\,)=\prod_i^l \frac{2\cdot depth(c_i^{'})\cdot depth(c_{i+1}^{'})}{depth(c_i^{'})+depth(c_{i+1}^{'})}\cdot\frac{1}{depth(T)}$$ **Equation 8.9**

where c'_1, c'_2... c'_l are the concept nodes along the path from c_1 to c_2. The intuition is to promote paths with deeper nodes, since paths with shallower nodes are more general. The final semantic relatedness between two concepts is the maximum product of *compactness* and *spe* given by any paths between them.

While these path based methods assume symmetric semantic relatedness, Schickel-Zuber and Faltings (2007) introduced a metric that allows measuring asymmetric similarity between concepts. It is called the **Ontology Structure based Similarity** (**OSS**), which firstly computes an **a-priori score** (**APS**) of every concept in a taxonomy to reflect its topological property; then views the similarity between two concepts as an effect of transferring the score from one concept to another via a directed path connecting the two concepts by their *lcs*. The *APS* of a concept is calculated based on the inverse of the number of *descendants(c).* As mentioned before, the classic definition of *lcs* may return multiple concepts. The authors introduced a tie-breaking method to handle such situation, which balances the depth of the node with the number of different paths leading to the node from the two concept nodes in question:

$$lcs_{SF07}(c_1,c_2) = max_{c \in lcs(c_1,c_2)}\{(|P(c_1,c)| + |P(c_2,c)|) \cdot 2^{depth(c)}\} \quad \textbf{Equation 8.10}$$

The hypothesis is that “a concept found higher in the ontology can still be more useful … if it has many paths leading to it”. Following this definition, the single *lcs* connects the two concepts with a directed path, along which the score of the starting concept is transferred to the target concept. The amount transferred depends on an upward (*α*) transfer from c_1 to the *lcs*, and a downward (*β*) transfer from the *lcs* to $c_{2.}$. *α* quantifies the amount of information of c_1 that can be generalised by the *lcs* as a ratio of their APS; while *β* quantifies how much information becomes specialised by c_2 as the difference between their APS. Next, both factors are combined in a log function normalised by the depth of the taxonomy to derive a score of semantic distance.

$$SemDist(c_1,c_2) = \frac{log(1+2\beta(c_2,lcs_{SF07}(c_1,c_2)) - log(\alpha(c_1,lcs_{SF07}(c_1,c_2)))}{depth|T|} \quad \textbf{Equation 8.11}$$

Figure 8.2 below summarises the connections between different path-based methods. The initial background information resources used in each study are noted as: WN – WordNet, BIO – a biomedical knowledge base, OTH – other structured knowledge bases

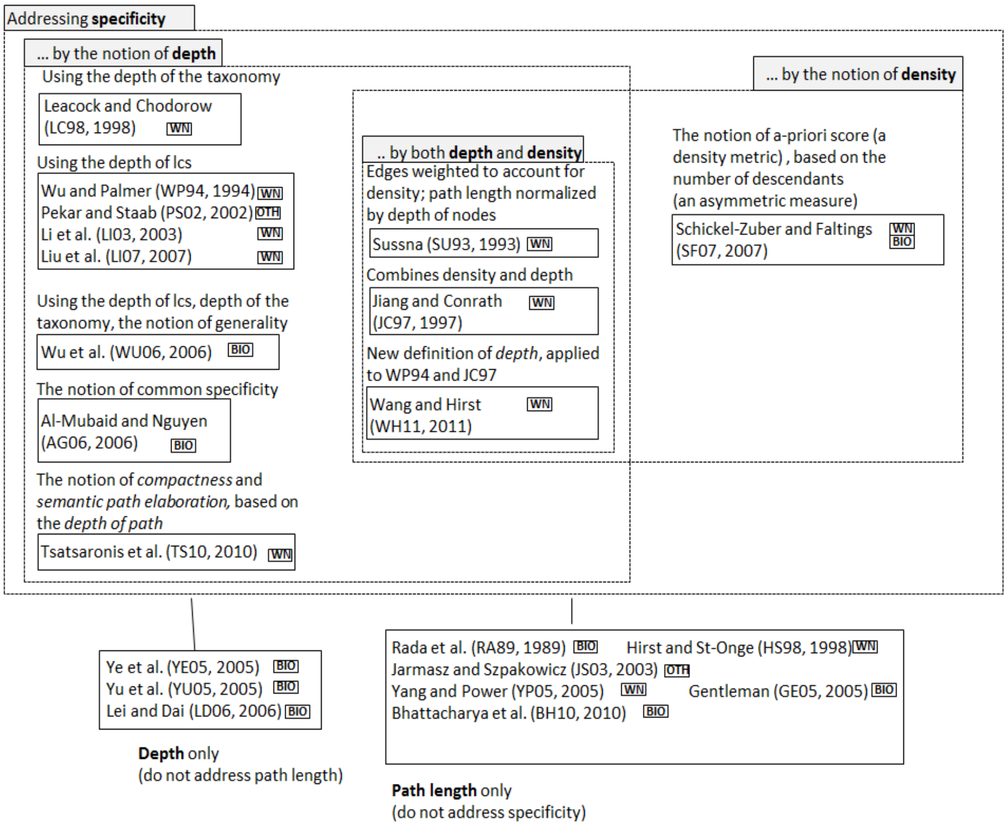

Addressing specificity
... by the notion of depth
Using the depth of the taxonomy
Leacock and Chodorow (LC98, 1998) WN
Using the depth of lcs
Wu and Palmer (WP94, 1994) WN
Pekar and Staab (PS02, 2002) OTH
Li et al. (LI03, 2003) WN
Liu et al. (LI07, 2007) WN
Using the depth of lcs, depth of the taxonomy, the notion of generality
Wu et al. (WU06, 2006) BIO
The notion of common specificity
Al-Mubaid and Nguyen (AG06, 2006) BIO
The notion of compactness and semantic path elaboration, based on the depth of path
Tsatsaronis et al. (TS10, 2010) WN
.. by both depth and density
Edges weighted to account for density; path length normalized by depth of nodes
Sussna (SU93, 1993) WN
Combines density and depth
Jiang and Conrath (JC97, 1997) WN
New definition of depth, applied to WP94 and JC97
Wang and Hirst (WH11, 2011) WN
... by the notion of density
The notion of a-priori score (a density metric), based on the number of descendants (an asymmetric measure)
Schickel-Zuber and Faltings (SF07, 2007) WN BIO
Ye et al. (YE05, 2005) BIO
Yu et al. (YU05, 2005) BIO
Lei and Dai (LD06, 2006) BIO
Depth only
(do not address path length)
Rada et al. (RA89, 1989) BIO
Hirst and St-Onge (HS98, 1998) WN
Jarmasz and Szpakowicz (JS03, 2003) OTH
Yang and Power (YP05, 2005) WN
Gentleman (GE05, 2005) BIO
Bhattacharya et al. (BH10, 2010) BIO
Path length only
(do not address specificity)

Figure 8.2. A summary of path based methods

8.2.3.2 IC based methods

Information content (IC) based methods hypothesize that the relatedness between concepts can be measured by the amount of information they share, which, if in a taxonomy, is often determined with respect to their *lcs*. These methods usually combine knowledge of a concept's hierarchical structure with statistics of its actual usage in text usually derived from a large corpus.

The first IC based method is introduced by Resnik (1995):

$$IC(c) = -\log p(c) \qquad \textbf{Equation 8.12}$$

where *p(c)* is the probability of encountering an instance of a concept *c*, estimated from noun (which corresponds to the concept) frequencies observed in a large corpus. The counting ensures that the frequency of a noun is added to all of its subsumers in the taxonomy, which guarantees *p* and therefore the IC of a concept, to be monotonic (i.e., *p(c)* ≤ *p(c')* and *IC(c)* ≥ *IC(c')* if *c IS-A c'*). The semantic similarity between concepts is defined as the IC of their *lcs*. Due to the monotonic nature of the IC measure, it can be considered as a measure of specificity of concepts (Li et al., 2003).

Resnik's definition of IC is widely adopted by later methods. Most of these improved Resnik's in different ways to overcome a number of limitations. **The first limitation** is that any two pairs of concepts having the same *lcs* will receive the same similarity (e.g., the pairs 'seafood – fruit' and 'bramley – kelp' have the same *lcs* 'food'), which is not necessarily appropriate. To address this problem, Jiang and Conrath (1997), and Lin (1998b) proposed to incorporate the IC of each concept in question as a way to balance that of their *lcs*. Briefly, Jiang and Conrath (1997) proposed a measure based on the rationale of the shortest path but formulates the distance as a function of the IC measure:

$$SemDist(c_1,c_2) = IC(c_1) + IC(c_2) - 2 \cdot IC(lcs(c_1,c_2)) \qquad \textbf{Equation 8.13}$$

Lin's (1998b) semantic similarity measure aims to address the commonality of two concepts as well as their difference, both measured in terms of IC:

$$SemSim(c_1,c_2) = \frac{2 \cdot IC(lcs(c_1,c_2))}{IC(c_1) + IC(c_2)} \qquad \textbf{Equation 8.14}$$

In its initial form the measure applies to taxonomic structures only. Maguitman et al. (2005) generalised Lin's measure such that it is applicable to both hierarchical and non-hierarchical links. A number of studies have adapted these methods to the biomedical domain with small modifications. For example, Speer et al. (2004) converted Lin's semantic similarity measure into a distance metric; while Schlicker et al. (2006) transformed Jiang and Conrath's distance metric to a measure of similarity.

The second limitation with Resnik's definition of IC is related to the calculation of *p(c)*, which is dependent on the choice of a corpus. As a result of this, given different corpora, it is possible to obtain a different IC for a concept. Seco et al. (2004) argued that the IC of a concept should be related to its hierarchical structure, and introduced an alternative IC measure called '**intrinsic IC**', calculated solely based on a taxonomy:

$$IC_{int}(c) = 1 - \frac{log(|descendant(c)|+1)}{log(|T|)} \qquad \textbf{Equation 8.15}$$

The rationale behind is that taxonomic structures usually organise concepts in such a way that the information expressed by a concept is inversely proportional to its descendants. Therefore, the IC value of a concept can be assessed using a function of the descendants it has. This notion of IC is adopted by Pirro et al. (2009), who introduced a method based on the same hypothesis of Lin (1998b) to quantify semantic similarity between two concepts based on their commonality and difference.

While these methods generally define semantic similarity of two concepts with respect to their *lcs*, which is usually a single concept, some methods have proposed to consider multiple *common subsumers* of the concepts. Couto et al. (2005) proposed the GraSM measure which takes the average of IC of all **common disjunctive subsumer** of two concepts. Two common subsumers are said to be *disjunctive* if there are ‘independent paths from both ancestors to the concept’, where an independent path is one that contains at least one concept unused by the other paths. The motivation is that each common disjunctive subsumer provides a different interpretation of the concepts that can be equally important. For example, with an *lcs* based method, the similarity between ‘leaf vegetable’ and ‘sea vegetable’ in the taxonomy of Figure 8.1 depends on their *lcs* ‘vegetable’. GraSM also takes into account the following independent paths to ‘food’: ‘leaf vegetable – vegetable – food’ and ‘sea vegetable – seafood – food’, and thus also considers the IC of ‘food’. On the other hand, there are no independent paths from ‘root vegetable’ and ‘leaf vegetable’ to ‘food’, and therefore, their similarity is only dependent on their *lcs* ‘vegetable’. Thus following the formula, GraSM will return a higher similarity for ‘root vegetable’ and ‘leaf vegetable’ than ‘leaf vegetable’ and ‘sea vegetable’. The intuition is that ‘sea vegetable’ has another interpretation (‘seafood’).

Wang et al. (2007) also argued that firstly, all ancestors of a concept should contribute to the semantics of the concept, and therefore, should be accounted for when measuring semantic similarities between concepts; secondly, the significance of the contribution should follow an inverse relationship with their distance to the concept. The method is based on two core notions: the **semantics** of a concept *c’* with respect to a target concept *c* – denoted by $semantics_c(c')$, and the **semantic value** of *c*. Given a hierarchical structure containing a concept *c*, their method firstly extracts a sub-hierarchy of *c* to be used for measuring semantic similarity. Formally, $T_c(c, C_c, E_c)$ is the extracted sub-hierarchy for the concept *c*, where C_c denotes all concepts in the sub-hierarchy including *subsumer(c)* and *c* itself, and E_c denotes the set of edges that connect any two concepts in the sub-hierarchy. The $semantics_c(c')$ for each $c' \in C_c$ is defined as a recursive function that aggregates the *semantics* of the descendants of *c'* (i.e., *semantics*$_{c'}$*(descendant(c'))*, where the contribution from one concept to its subsumers is modified by weighted edges and *semantics*$_c$*(c)* = 1, i.e., the semantics of a concept with respect to itself is 1.

$$semantics_c(c') = \begin{cases} 1, & c = c' \\ max_{ch \in children(c'), e \in E_c}\{weight(e) \cdot semantics_c(ch)\}, & c \neq c' \end{cases} \quad \textbf{Equation 8.16}$$

Next, the *semantics*$_c$*(c')* of all *c'*∈ *C*$_c$ is summed up as a measure of semantic value of *c*, denoted by *sv(c)*. The final semantic similarity between two concepts is calculated by their 'semantics in common', which follows similar rationale with other IC based methods:

$$SemSim(c_1, c_2) = \frac{\sum_{c \in C_{c_1} \cap C_{c_2}} (semantics_{c_1}(c) + semantics_{c_2}(c))}{sv(c_1) + sv(c_2)} \quad \text{Equation 8.17}$$

Figure 8.3 below summarises the connections between different IC-based methods. The initial background information resources used in each study are noted as: WN – WordNet, BIO – a biomedical knowledge base, OTH – other structured knowledge bases

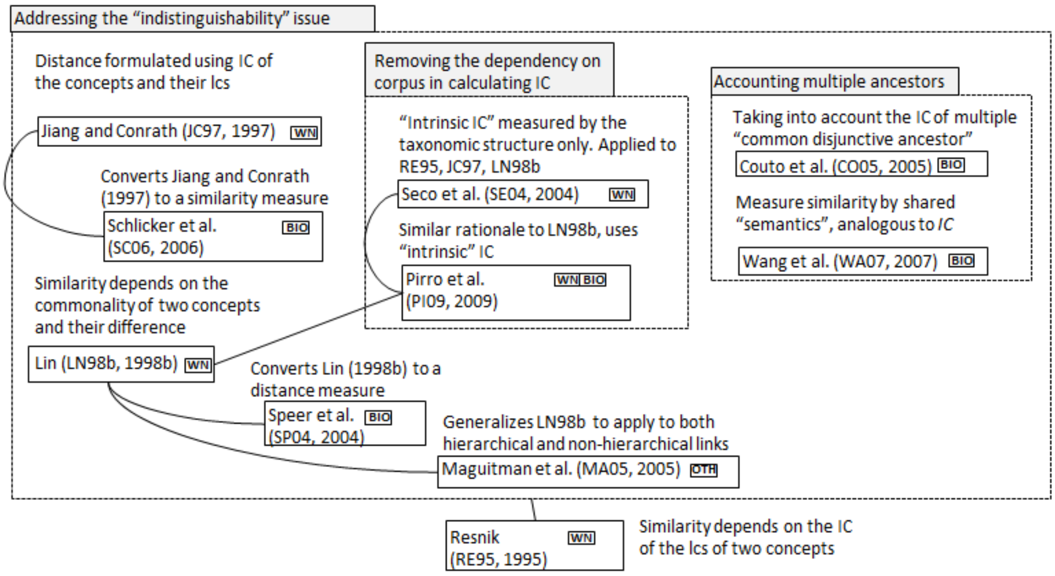

Addressing the "indistinguishability" issue
Distance formulated using IC of the concepts and their lcs
Jiang and Conrath (JC97, 1997) WN
Converts Jiang and Conrath (1997) to a similarity measure
Schlicker et al. (SC06, 2006) BIO
Similarity depends on the commonality of two concepts and their difference
Lin (LN98b, 1998b) WN
Removing the dependency on corpus in calculating IC
"Intrinsic IC" measured by the taxonomic structure only. Applied to RE95, JC97, LN98b
Seco et al. (SE04, 2004) WN
Similar rationale to LN98b, uses "intrinsic" IC
Pirro et al. (PI09, 2009) WN BIO
Accounting multiple ancestors
Taking into account the IC of multiple "common disjunctive ancestor"
Couto et al. (CO05, 2005) BIO
Measure similarity by shared "semantics", analogous to IC
Wang et al. (WA07, 2007) BIO
Converts Lin (1998b) to a distance measure
Speer et al. (SP04, 2004) BIO
Generalizes LN98b to apply to both hierarchical and non-hierarchical links
Maguitman et al. (MA05, 2005) OTH
Resnik (RE95, 1995) WN
Similarity depends on the IC of the lcs of two concepts

Figure 8.3. A summary of IC based methods

8.2.3.3 Gloss based methods

As introduced previously in Section 8.2.2.1, concepts represented in knowledge bases are often provided with definitions and examples. Gloss based methods generally refer to these as **glosses**, and hypothesize that the relationship between concepts is often implied by the shared words in their glosses. Therefore, gloss based methods propose to measure semantic relatedness with respect to the *overlap of words* in two concepts' glosses. Following this definition, gloss based methods assess relatedness rather than similarity, since they do not make use of the taxonomic links. However, there are also exceptions.

Lesk (1986) introduced the first gloss based measure, which simply evaluates the relatedness between concepts by the number of overlapping words in their glosses, denoted by *gloss(c)*. It has been adopted by several later studies including Banerjee and Pedersen (2003), Mihalcea and Moldovan (1999), and Gurevych (2005), which are distinguished by the construction of *gloss(c)*. For example, Banerjee and Pedersen argued that the gloss of a concept should be extended by including the glosses of others that are related to the concept; while Gurevych adapted the method to knowledge bases that do not provide a gloss of a concept by building a pseudo-gloss, which concatenates concepts in close relation (e.g., hypernym, synonym) to the concept.

Further to the motivation behind Gurevych (2005), the definition of gloss can be lessened to allow different methods of gloss construction. In this way a few other methods can also be classified as gloss based. Gentleman (2005) proposed to measure semantic similarity as the ratio between the overlap and the union of the subsumers of two concepts extracted from a taxonomic structure, where *subsumer(c)* can be considered as a gloss representation of *c*. The general hypothesis is that each subsumer of a concept provides an interpretation. Thus the overlap is the amount of interpretations common to both concepts, while the union is the amount required to fully interpret both concepts. Turdakov and Velikhov (2008) adopted the same measure but using Wikipedia article links. Briefly, a pseudo-gloss for each concept is constructed by concatenating the outgoing and ingoing links of the article page describing that concept, and placed into the same formula. While the method by Gentleman (2005) can be considered as quantifying similarity with respect to the shared knowledge of two concepts, Batet et al. (2010) proposed a method that quantifies similarity with respect to 'non-shared knowledge'. The method represents each concept as a gloss of their subsumers in the same way, and the non-shared knowledge is simply the difference between the union and the overlap of the subsumers of the two concepts.

To some extent, the rationale behind these methods is related to that of IC based methods, in the sense that relatedness is quantified based on the information that two concepts share in common. However, they are classified as *gloss* based methods rather than IC based for two reasons: (1) they do not quantify the *information content* of a concept formally; (2) the calculation is generally based on certain forms of vocabulary overlap, which is the key feature of gloss based methods.

Figure 8.4 below summarises the connections between different gloss-based methods. The initial background information resources used in the studies are noted as: WN – WordNet, WK – Wikipedia, BIO – a biomedical knowledge base, OTH – other knowledge bases.

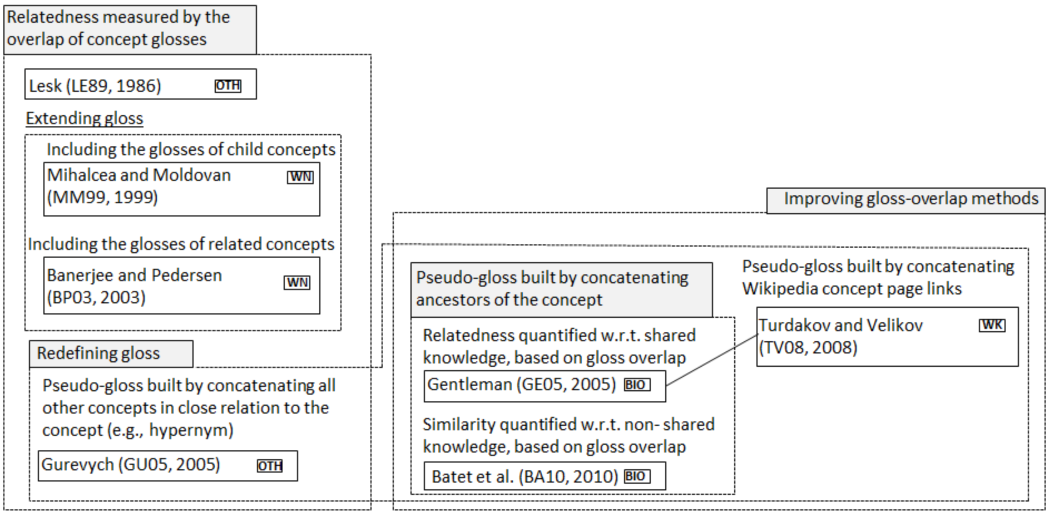

Relatedness measured by the overlap of concept glosses
Lesk (LE89, 1986) OTH
Extending gloss
Including the glosses of child concepts
Mihalcea and Moldovan (MM99, 1999) WN
Including the glosses of related concepts
Banerjee and Pedersen (BP03, 2003) WN
Redefining gloss
Pseudo-gloss built by concatenating all other concepts in close relation to the concept (e.g., hypernym)
Gurevych (GU05, 2005) OTH
Improving gloss-overlap methods
Pseudo-gloss built by concatenating ancestors of the concept
Relatedness quantified w.r.t. shared knowledge, based on gloss overlap
Gentleman (GE05, 2005) BIO
Similarity quantified w.r.t. non- shared knowledge, based on gloss overlap
Batet et al. (BA10, 2010) BIO
Pseudo-gloss built by concatenating Wikipedia concept page links
Turdakov and Velikov (TV08, 2008) WK

Figure 8.4. A summary of gloss based methods

8.2.3.4 Feature vector based methods

Feature vector based methods refer to methods that represent a term or concept using a feature vector derived from a structured knowledge base, rather than using co-occurrence counts or contexts. For vector based methods, concepts and their lexicalised forms can be represented based on features encoded in knowledge bases. For example, in the taxonomy shown in Figure 8.1, concepts can be described by the features *parent concepts* and *child concepts*.

With feature vector representations, assessing relatedness between concepts can be achieved by comparing the similarity between their feature vectors. The most well-known and widely used measure for this purpose is the cosine similarity function, which measures the similarity between two vectors by the cosine of the angle between them:

$$cosine(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{||\vec{v}_1||||\vec{v}_2||}$$ **Equation 8.18**

The cosine similarity measure is used by the majority of vector based methods. However, they differ significantly in terms of how the feature vectors are constructed. Because such methods do not exploit the hierarchical nature of a knowledge base, they generally quantify relatedness rather than similarity.

Generally, feature vector based methods can be further divided into two types. The first type of methods build a concept vector using the lexical semantic content (e.g., a concept's gloss and synonyms in WordNet) directly defined for that concept in a knowledge base. Intuitively the lexical semantic content can be considered as features relevant to a concept. The second type of methods builds a concept vector using other *related* concepts in the knowledge base. The intuition is that semantically related concepts are related to a similar set of concepts. In these methods, features for each relevant concept are exploited indirectly and usually contribute to the weight of the relevant concept in the corresponding vector of the target concept.

Patwardhan and Pedersen (2006) represented a concept by a second-order gloss vector using WordNet glosses. Firstly, WordNet is turned into a corpus made of the set of glosses in WordNet. Next, a first-order context vector based on co-occurrence is created for every word in this corpus, where words are said to co-occur if they are found in the same gloss. Then for each concept, a second-order gloss vector is constructed by combining the first-order context vectors of words that appear in its gloss. The intuition is that the orientation of the second order gloss vector indicates the domains or topics it is associated with. The relatedness between two concepts is then measured as the cosine similarity between the two vectors. Zesch et al. (2008b) suggested that different kinds of lexical semantic content defined for a concept can be used as features to build concept vectors. They demonstrated this using three different knowledge bases with the generic cosine similarity function. Concept vectors are built using the first paragraph of Wikipedia, WordNet gloss and Wiktionary gloss separately.

One of the methods that represent a term or concept as a vector of related concepts is introduced in Ziegler et al. (2006). They firstly queried each term on Google Directory[9], which returns a ranked list of websites each annotated with a category defined in the

[9] Google Directory, originally available at http://www.google.com/dirhp, which has been closed

Open Directory Project (ODP) taxonomy. The ODP taxonomy organises category concepts in a tree structure, and classifies websites using the categories. The term is then represented as a vector $\vec{v}$ of ODP categories, such that $|\vec{v}| = |T|$ the size of the ODP taxonomy. The initial weight of each category *cat* in the vector is dependent on the rank of the website that the corresponding category is associated with. Next, for each *cat*, its weight is propagated upwards to the subsumers of *cat* along the ODP taxonomic links, where the amount of propagation depends on the density of the parent node at each step. Then the final weight of *cat* is adjusted to combine the weights of all of its subsumers, resulting in the final weighted vector representation of the word. The intuition behind this is that a category could be counted as an instance of its subsumers, and therefore its weight should reflect the contribution from all its subsumers.

A similar approach is Explicit Semantic Analysis (ESA) introduced by Gabrilovich and Markovitch (2007). ESA computes word relatedness using vectors constructed using Wikipedia concepts. The method firstly builds an inverted index of words and Wikipedia articles, then represents each word as a high-dimensional vector of Wikipedia articles. Since each article usually focuses on a certain topic, the vector can be viewed as a vector of concepts, where each element in the vector corresponds to a concept and the dimension is the number of Wikipedia articles. Each element in the vector is weighted by the tf.idf score of the word in the associated article. The intuition is that the weight denotes the degree of relevance between the word and the concept. Semantic relatedness is also calculated using the cosine function.

ESA was later extended by Hassan and Mihalcea (2009) and Radinsky et al. (2011). Hassan and Mihalcea (2009) introduced three modifications to the original ESA method: 1) normalise vector elements to account for the length of the associated concept article, since the original method is biased towards long articles; 2) scale the weights of vector elements according to the corresponding concept's depth in the Wikipedia category tree to promote concepts that are lower down (and thus more specific); 3) replaced the cosine similarity metric with an overlap based metric similar to that by Lesk (1986) in order to place more emphasis on the overlap of vectors. Radinsky et al. (2011) proposed to incorporate the 'temporal behaviour' of words in computing their relatedness. The idea originates from the observation that semantically related words do not necessarily co-occur in the same articles, however, they are likely to be mentioned roughly around the same time. For example, by studying the distributional patterns of words over time using a collection of New York Times articles spanning over 130 years, they found that the words 'war' and ‘peace’ tend to correlate in frequency of use over time. However, they might rarely be mentioned at the same time in the same articles. To exploit this feature they proposed to modify the ESA method by modelling the ‘temporal dynamics’ of each non-zero weighted concept in the vector, and scaling their temporal dynamics according to the concept’s original weight in the vector. The temporal dynamics of a concept is modelled based on a series of its usage frequency over the corpora:

$$dynamics(c) = \left\langle \frac{\{d \in D_1 \mid freq(c,d)|\}}{|D_1|}, \ldots, \frac{\{d \in D_n \mid freq(c,d)|\}}{|D_n|} \right\rangle \quad \textbf{Equation 8.19}$$

where $D_1 \ldots D_n$ can be viewed as a history represented by a collection of corpora ordered chronologically, D_i is a corpus at time point *I*, *freq(c, d)* returns the frequency of observing the lexicalised words of the concept in *d*. Details of the counting method can be found in Radinsky et al. (2011). Next each concept in the ESA vector is represented by its temporal dynamics, the weight of which inherits that of the concept in the original ESA vector. The final relatedness between the two words can be computed as the sum of pairwise concept relatedness of non-zero weighted concepts in their vectors. And the pairwise concept relatedness is computed using their temporal dynamics by two different methods for measuring time series similarity.

Milne and Witten (2008) employed the hyperlinked Wikipedia article graph and proposed to represent a Wikipedia concept as a vector of concepts linking to it or a vector of concepts it links to. The cosine similarity function and the measure proposed by Cilibrasi and Vitanyi (2007) were applied to each vector representation respectively for evaluation. Liu and Chen (2010) also represented Wikipedia concept as a vector of links, but defined the relatedness between two concept vectors as the sum of the pair-wise relatedness between the links in the two vectors, which is then computed using either gloss overlap, or the Wu and Palmer’s (1994) measure using the category graph.

The method by Milne and Witten only considers the adjacent concepts (e.g., neighbours) in a semantic graph. Another stream of work based on the link structure of a semantic network exploit both direct and indirect connections between concepts. Gouws et al. (2010) proposed to apply the spreading activation algorithm (Collins and Loftus, 1975) to a semantic network of concepts as a method to derive concept vectors. Briefly, given a semantic network of nodes connected by directed and weighted edges, spreading activation allows propagating the ‘activation value’ from a given node (or a set of nodes) to any other nodes in the network across the path that connecting them. The amount of value propagated drops as the number of edges increases (depending on a so-called 'decay factor'). For the purpose of measuring semantic relatedness between two concepts c_1 and c_2, Gouws et al. built a semantic network of Wikipedia concepts connected by ingoing links from the article page. They firstly set a non-zero initial activation value to the node of concept c_1, while all other nodes receive an initial value of 0. Spreading Activation is started to propagate the value to c_2 through each node along the set of paths connecting them. After the activation terminates (subject by individual methods), c_2 receives a final activation value, denoted by *act(*$c_2|c_1$*)*. Meanwhile, a concept vector $\vec{v}_1$ is created for c_1 by collecting the final activation values of all nodes in the network. These values can be considered as a measure of the relevance between each node and c_1. Next, the same process is repeated from the opposite direction, by propagating an initial activation value from c_2 to c_1, to obtain *act(*$c_1|c_2$*)* and $\vec{v}_2$. Finally, the authors computed the semantic relatedness between the two concepts in three different ways: (1) as a function of the sum of *act(*$c_2|c_1$*)* and *act(*$c_1|c_2$*)*, in which case each value is considered to be a distance from one to another; (2) as the cosine similarity of the two vectors; (3) using a distance-based formula introduced in Cilibrasi and Vitanyi (2007).

Harrington (2010) and Wojtinnek and Pulman (2011) applied the same technique but derived a semantic and syntactic structure from an unstructured corpus, by applying NLP techniques such as NER and syntactic parsing to a corpus to generate a syntactic and semantic network of entities and concepts. The authors argued that the advantage of using a corpus is that they may provide better coverage of domain specific information than a general purpose knowledge base.

Another popular stream of methods that exploit the link structure of a semantic network for constructing concept vectors is using Personalised PageRank (Haveliwala, 2002), a generalisation of the well-known PageRank algorithm (Page et al., 1999), which is based on the principle of random walk on graphs. Random graph walk formalises the intuitive idea of taking successive steps in a graph, each in a random direction (Lovász, 1993). The output of a *t*-step random walk process on a graph can be represented as a matrix of probability distribution, often referred to as a transitional distribution, where each row encodes the probability of a random walker reaching this particular node starting from any other nodes on the graph after *t* steps. PageRank, originally well-known for its usage in ranking webpages by Google, is the most typical form of random graph walk. It views the Web as a graph of webpages, connected by hyperlinks. A surfer randomly clicks on hyperlinks, where each click is equivalent to one step of random walk and takes the surfer from one node to another. Then the PageRank algorithm builds on a random walk process that takes a very large number of steps, so large that the resultant transitional distribution at any nodes on the graph becomes converged, resulting in a stationary distribution. The stationary distribution at a node typically consists of uniform values for every starting node, which means that the probability of reaching this node is constant, regardless of starting nodes. This is considered to be indicative of the amount of time a surfer will take to reach that node, while the value is only affected by the connectivity of the graph. Intuitively, the longer it takes the surfer to reach the node, the less important the node is. Therefore, the results can be used for ranking webpages.

In personalised PageRank, the ranking (or importance) of each page is biased (or 'personalised') towards a particular query, such that webpages more relevant to the query receive higher importance and therefore, higher ranks. Such ideas have been adapted to a semantic network of concepts to develop semantic relatedness methods (Hughes and Ramage, 2007; Yeh et al., 2009; Yazdani and Popescu-Belis, 2010). Generally, the PageRank algorithm is applied to the semantic network of concepts to obtain the stationary probability distribution to represent the likelihood of reaching any node in the network. Next, this distribution is personalised against each question concept node following Haveliwala's method, which effectively places more weight to nodes closer to the target concept in the graph. Intuitively, this can be viewed as re-ranking all nodes (similar to webpages) in the network (similar to the Web) with respect to their connection (similar to relevance) with the target concept node (similar to a query). The two resultant distributions, biased towards c_1 and c_2 respectively, can be represented as two vectors, the dimensionality of which is the number of nodes in the network. Then the relatedness between them can be computed with a measure of vector similarity.

A major difference between these methods is how the semantic network is constructed. Hughes and Ramage (2007) constructed a graph of WordNet synsets, tokens (e.g., polysemous words), and token-with-part-of-speech (tokenPOS, e.g., click#noun, click#verb). Edges are established for any relations between synsets defined in WordNet, for a synset node and a tokenPOS node if the synset 'uses' the tokenPOS (e.g., synset 'click#noun#mouse click' uses 'click#noun'), and for a tokenPOS node with a token node following the same strategy. Yeh et al. (2009) and Yazdani and Popescu-Belis (2010) constructed a graph of Wikipedia concepts. An edge is established between two concepts if they are connected by links in their article page, by sharing a category, or sharing a link in their infoboxes. Yazdani and Popescu-Belis (2010) assigned higher weights to certain types of edges to gain better performance.

Figure 8.5 below summaries feature vector based methods. The initial background information resources used are noted as: WN – WordNet, WK – Wikipedia, OTH – other knowledge bases, C – a general purpose corpus,

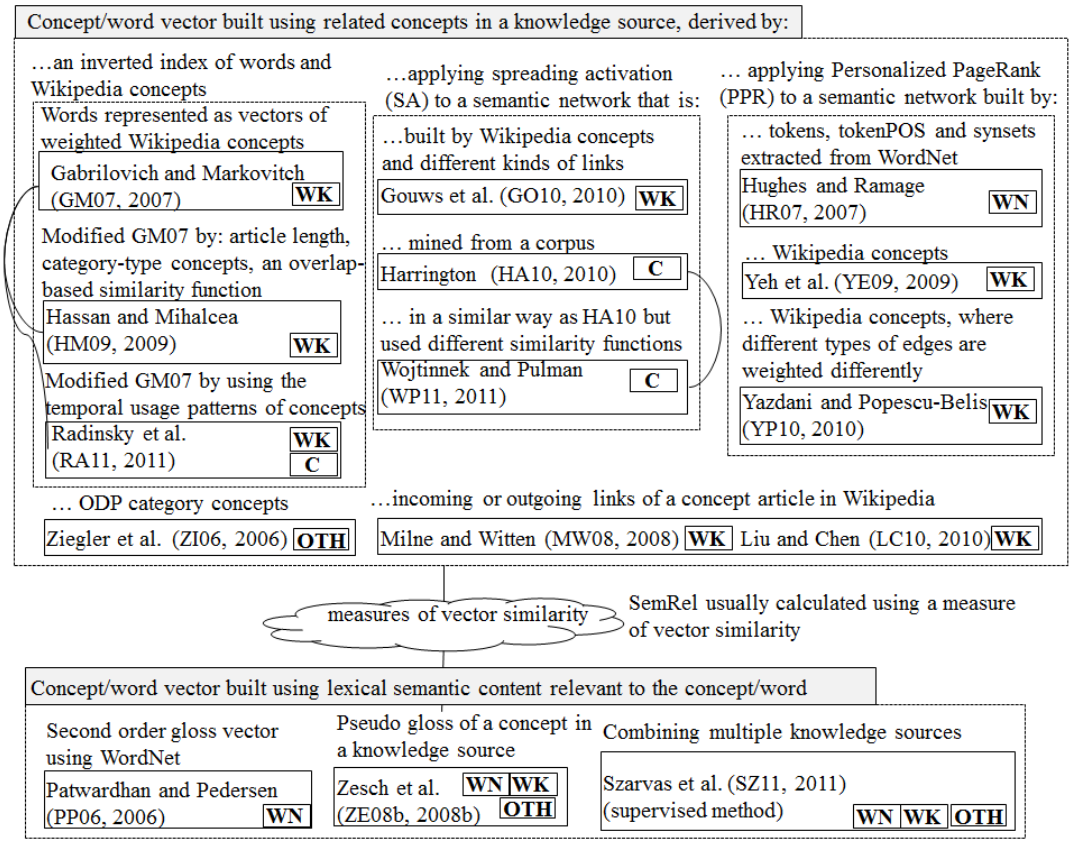

Concept/word vector built using related concepts in a knowledge source, derived by:
…an inverted index of words and Wikipedia concepts
Words represented as vectors of weighted Wikipedia concepts
Gabrilovich and Markovitch (GM07, 2007) WK
Modified GM07 by: article length, category-type concepts, an overlap-based similarity function
Hassan and Mihalcea (HM09, 2009) WK
Modified GM07 by using the temporal usage patterns of concepts
Radinsky et al. (RA11, 2011) WK C
…applying spreading activation (SA) to a semantic network that is:
…built by Wikipedia concepts and different kinds of links
Gouws et al. (GO10, 2010) WK
… mined from a corpus
Harrington (HA10, 2010) C
… in a similar way as HA10 but used different similarity functions
Wojtinnek and Pulman (WP11, 2011) C
… applying Personalized PageRank (PPR) to a semantic network built by:
… tokens, tokenPOS and synsets extracted from WordNet
Hughes and Ramage (HR07, 2007) WN
… Wikipedia concepts
Yeh et al. (YE09, 2009) WK
… Wikipedia concepts, where different types of edges are weighted differently
Yazdani and Popescu-Belis (YP10, 2010) WK
… ODP category concepts
Ziegler et al. (ZI06, 2006) OTH
…incoming or outgoing links of a concept article in Wikipedia
Milne and Witten (MW08, 2008) WK
Liu and Chen (LC10, 2010) WK
measures of vector similarity
SemRel usually calculated using a measure of vector similarity
Concept/word vector built using lexical semantic content relevant to the concept/word
Second order gloss vector using WordNet
Patwardhan and Pedersen (PP06, 2006) WN
Pseudo gloss of a concept in a knowledge source
Zesch et al. (ZE08b, 2008b) WN WK OTH
Combining multiple knowledge sources
Szarvas et al. (SZ11, 2011) (supervised method) WN WK OTH

Figure 8.5. A summary of feature vector based methods

8.2.3.5 Distributional similarity as proxy for lexical semantic relatedness

Weeds (2003) summarised the literature and defined that two words are distributionally similar if (1) they tend to occur in each other's context; or (2) the contexts each tends to occur in are similar; or (3) that if one word is substituted for another in a context, its 'plausibility' is unchanged. Different methods have adopted different definitions of contexts, but usually a context is the set of words collected from the window around the word, or an entire document, or syntactic relationship as introduced in Lin (1998a) and Weeds (2003). Following this definition, a multitude of measures and algorithms are available for measuring distributional similarity.

There is a widely accepted hypothesis that distributional similarity can predict semantic similarity and thus relatedness in general. As a result, many studies have proposed to use distributional similarity methods as a proxy for lexical semantic relatedness. However, Budantisky and Hirst (2006) argued that there are three essential differences between the two paradigms. Firstly, semantic relatedness is inherently a relation on concepts, while distributional similarity is a relation on words; secondly, semantic relatedness is typically symmetric, whereas distributional similarity can be potentially asymmetric; finally, semantic relatedness depends on a structured lexicographic or knowledge bases, distributional similarity is relative to a corpus. Nevertheless, due to the popularity of their application to semantic relatedness, in the following, several recent studies in this direction are discussed. The criteria for selection is that the method is either evaluated using methods or datasets usually used for evaluating lexical semantic relatedness, or that it is later ported to address relevant tasks in other studies. A popular research direction taken by these approaches is exploiting the Web as the background information corpus.

Matsuo et al. (2006) computed distributional similarity of two terms using statistics collected from search engines. The method queries each term using a search engine to retrieve the number of pages containing each term denoted by *freq(w)*. Next, the number of pages containing both term s *freq*(w_1, w_2) is counted as the page counts for a query concatenating both terms. Then using these figures, the distributional similarity can be calculated using the Point-wise Mutual Information (PMI) measure or the Chi-Square (x^2) statistical test. Essentially, in their method, the context is equivalent to a webpage.

Cilibrasi and Vitanyi (2007) introduced the Normalised Google Distance (NGD) measure, which builds on a data compression related theory that is rather intricate. The resultant method is however, very simple and also utilises page counts returned for the pair of terms and each term separately:

$$SemDist(w_1, w_2) = \frac{max\{log(freq(w_1)), log(freq(w_2))\} - log(freq(w_1, w_2))}{log N - min\{log(freq(w_1)), log(freq(w_2))\}} \quad \textbf{Equation 8.20}$$

As mentioned before, this method is adopted by Milne and Witten (2008) for measuring semantic relatedness between concepts using Wikipedia. Briefly, they substituted *freq(w)* as the number of incoming links leading to a concept article and *freq*(w_1, w_2) as the number of incoming links leading to both concept articles.

While these methods based on page counts can be effective, a potential issue is that they do not consider the relative position of terms or multiple occurrences of the term in a single page. Two terms occurring in a page may not be related at all if they are distant. Therefore, methods based on page counts are prone to errors (Bollegala et al., 2007). Many methods cope with this by using the returned snippets for the query. Chen et al. (2006) proposed to count term co-occurrences in the top *N* snippets returned by a search engine. They hypothesize that two terms w_1 and w_2 are associated if it is possible to find w_2 from w_1 (a forward process) and find w_1 from w_2 (a backward process) by web search. Therefore, their method queries each term in turn, and count the occurrences of the other in the top *N* snippets returned for the query (i.e., $freq(w_1|w_2)$ and $freq(w_2|w_1)$). This process gives two different co-occurrence figures for the two terms, which they call a 'double checking' process. The distributional similarity is then computed as:

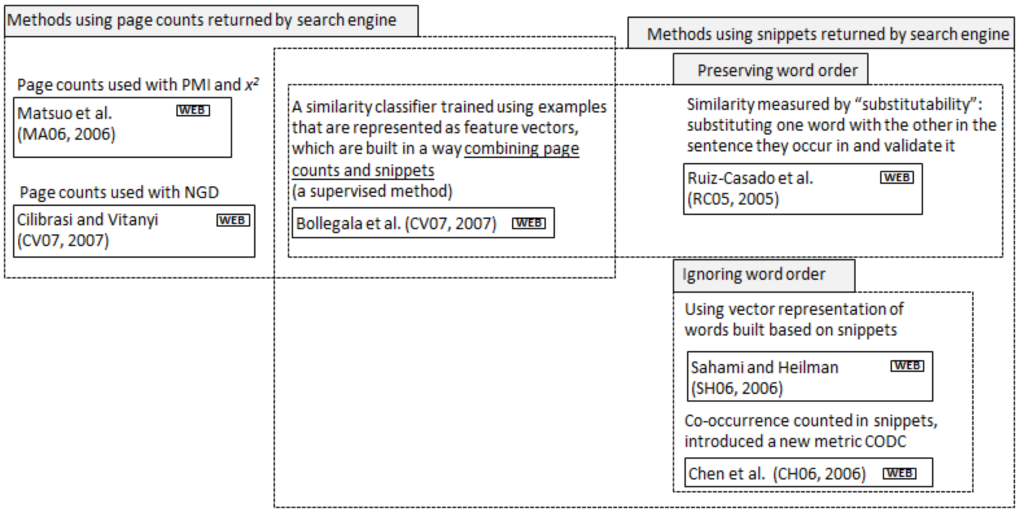
$$DistSim(w_1,w_2) = \left\{ \begin{array}{cc} 0, & if\ \ freq(w_1 \mid w_2) = 0\ or\ \ freq(w_2 \mid w_1) = 0 \\ e^{log(\frac{freq(w_2|w_1)}{freq(w_1)} \cdot \frac{freq(w_1|w_2)}{freq(w_2)})^{\alpha}}, & else \end{array} \right\} \quad \textbf{Equation 8.21}$$

where α is a control parameter. Sahami and Heilman (2006) query each word to obtain a set of snippets. For each word, each of its corresponding snippets is represented as a weighted term vector, and the centroid of the set of vectors is computed. Similarity between the two words is defined as the inner product between the corresponding centroid vectors.

Although these methods can tackle the limitation of page counting based methods to certain extent, one of their limitations, as suggested by Ruiz-Casado et al. (2005), is that they ignore word-order and phrasal structures. The authors thus proposed a method that assesses the substitutability of two terms as a measure of their similarity. It firstly collects a set of sentences (S_1, S_2) from the snippets returned for each word (w_1, w_2) as a query to a search engine. Next, it counts in how many of sentences in S_1 it is possible to substitute w_1 with w_2. This is done by replacing w_1 with w_2 to create a new sentence, which is queried using the search engine for validation. The same process is repeated for S_2 and the final similarity score is derived based on the percentage of sentences that are substitutable.

Bollegala et al. (2007) proposed a method that combines both page counts and sentence-level contexts in snippets. Their method is based on supervised classification trained on examples, which is uncommon in semantic relatedness and distributional similarity methods. An SVM based classifier is trained using a set of synonym pairs as positive examples and a set of non-synonym pairs as negative examples, both of which are randomly selected from WordNet. Firstly, they apply a pattern induction process to extract lexical patterns that are likely to indicate synonyms and non-synonyms in texts. For this purpose, the positive and negative examples are used as queries to retrieve snippets from a search engine. Then lexical-syntactic patterns are extracted from the data and those that more often represent positive examples are selected to be used as features for the next learning task. Next, each pair of words is queried using the search engine, and each of the selected pattern features are searched within the returned snippets. A feature vector is created for the pair based on the frequencies of each pattern and four additional scores computed using page counts based on similar methods to Matsuo et al. (2006). Using this feature representation, an SVM model is trained using the training data, and then used to predict the similarity between any new pairs of terms given a feature representation created in the same way.

Figure 8.6 below summaries distributional similarity methods and their connections. The initial background information resources used are noted as: WEB – the Web



Methods using page counts returned by search engine
Page counts used with PMI and x^2
Matsuo et al. (MA06, 2006)
WEB
Page counts used with NGD
Cilibrasi and Vitanyi (CV07, 2007)
WEB
A similarity classifier trained using examples that are represented as feature vectors, which are built in a way combining page counts and snippets (a supervised method)
Bollegala et al. (CV07, 2007)
WEB
Methods using snippets returned by search engine
Preserving word order
Similarity measured by "substitutability": substituting one word with the other in the sentence they occur in and validate it
Ruiz-Casado et al. (RC05, 2005)
WEB
Ignoring word order
Using vector representation of words built based on snippets
Sahami and Heilman (SH06, 2006)
WEB
Co-occurrence counted in snippets, introduced a new metric CODC
Chen et al. (CH06, 2006)
WEB

Figure 8.6. A summary of distributional similarity methods used for lexical semantic relatedness

8.2.3.6 Hybrid methods

Hybrid methods are those that combine multiple semantic relatedness methods or distributional similarity methods to arrive at a single measure of semantic relatedness. Many of these methods firstly calculate semantic relatedness using an assembly of different purebred methods, and then derive the aggregated score as the average, sum or maximum of all scores. This kind of simple **combination** is intuitive – if each distinctive method provides a different perspective of semantic relatedness, it is natural to combine them to create a full picture. For example Alvarez and Lim (2007) calculated semantic relatedness as the maximum score given by either a method that inverts the shortest path length (weighted by depth of concepts along the path), or a method based on the depth of *lcs* and the taxonomy, or the gloss overlap in WordNet. Riensche et al. (2007) also adopted a similar strategy. Sheng et al. (2010) linearly combined the distributional similarity of two words using a PMI based method with the sematic similarity score calculated using Wu and Palmer's method (Wu and Palmer, 1994). Similarly, Gracia and Mena (2008) linearly combined the score calculated using the Cilibrasi and Vitanyi's measure with one that calculated using an overlap based method.

Other hybrid methods employ scores given by individual methods as an integral part of the hybrid model, usually as some kind of features. To distinguish them from the above combination methods, we call this type of hybrid method the **integration methods**. Rodr ǵuez and Egenhofer (2003) combined a gloss based method with depth as a normalisation factor. Given the gloss created for two concepts, semantic relatedness is determined by the common characteristics *α* measured by set overlap as $gloss(c_1) \cap gloss(c_2)$, and non-common characteristics measured by set difference as $\beta = gloss(c_1)/gloss(c_2)$, *and* $\gamma = gloss(c_2)/gloss(c_1)$*:*

$$SemRel(c_1, c_2) = \frac{|\alpha|}{|\alpha| + k(c_1, c_2) \cdot |\beta| + (1 - k(c_1, c_2)) \cdot |\gamma|} \qquad \textbf{Equation 8.22}$$

where $k(c_1, c_2)$ is the relative importance factor for non-common characteristics, calculated using the depth of the two concepts in the taxonomy:

$$k(c_1, c_2) = \left\{ \begin{array}{ll} \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)} & ,if\ depth(c_1) \le depth(c_2) \\ 1 - \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)} & else \end{array} \right\} \qquad \textbf{Equation 8.23}$$

This measure is asymmetric. The effect of weighting non-common characteristics with respect to the depth of two concepts is that relatedness from deeper concepts to shallower concepts is higher than the opposite, which is consistent with the common perception of asymmetric semantic relatedness. This method is later adapted by Petrakis and Varelas (2006) with different definitions of gloss.

Othman et al. (2007) defined semantic similarity of two concepts in a taxonomy as a function of their individual distance to their *lcs*. This distance is calculated using a measure that combines the *IC*, depth and local density in order to address specificity of concepts:

$$dist(c_1,c_2)=\sum_i^l D(c_i')\cdot E(c_i')\cdot(IC(c_{i+1}')-IC(c_i')) \qquad \textbf{Equation 8.24}$$

where c'_1, c'_2... c'_l are the list of concepts along the path from c_1 to c_2, *D(c)* and *E(c)* are functions that return the depth and local density of the concept node respectively. They are slightly modified based on the classic definition of depth and density. The distance is effectively the sum of weighted edges along the path from c_1 to c_2, which has a similar notion to semantic path elaboration proposed by Tsatsaronis et al. (2010). The final semantic similarity is computed as:

$$\begin{aligned}&SemSim(c_1,c_2)=\\&\quad 1-min\{1,\frac{dist(c_1,lcs(c_1,c_2))+dist(c_2,lcs(c_1,c_2))}{max_{c\in T}\{IC(c)\}}\}\end{aligned} \qquad \textbf{Equation 8.25}$$

Pozo et al. (2008) proposed to derive a taxonomic structure of words from a corpus and apply path-based methods to the taxonomy. Their motivation is that even if a well-curated knowledge base is available, many structural relations may not be encoded but may be hidden in a large corpus. Thus they proposed to uncover hidden structural relations between words by applying hierarchical clustering to the words based on their distributional features observed from corpora. The approach consists of four steps: (1) representing each word extracted from a corpus by a vector of its contextual words, and compute pairwise similarity of words using the cosine similarity function; (2) creating a connected graph of words, where edges are established for two words if the pairwise similarity is above a threshold; (3) apply spectral clustering to partition the graph; (4) apply agglomerative hierarchical clustering to generate a clustering tree of words, which is used as a taxonomy. Then the semantic distance between any words is simply the depth of their *lcs*.

Han and Zhao (2010) proposed the *Structural Semantic Relatedness* (*SSR*) method, which makes use of three methods including that by Lin (1998b), Milne and Witten (2008) and the NGD method by Cilibrasi and Vitanyi (2007). Given a collection of words, the method begins with a pre-processing step that calculates semantic relatedness for each pair of words using all three methods. Due to the coverage of background information resource used in each method, a pair may receive between 1 and 3 relatedness scores. For multiple scores, the score returned by the most ‘reliable’ method is retained. This is arbitrarily determined in the order of preference as Lin, Milne and Witten, and NGD. Next, the words are plotted in a connected graph, where the score is used to weigh edges between them. Let *e(*w_1*,*w_2*)* denote the weight of the edge connecting two words, *neighbour(w)* returns the immediate neighbours (connected by an edge) of a word, the *SSR* is then computed as:

$$Semrel(w_1,w_2) = \alpha\sum_{w'\in neighbou(w_1)}\left(\frac{e(w_1,w')}{d_{w_1}}\cdot SemRel(w',w_2)\right)+\beta\cdot e(w_1,w_2)$$ **Equation 8.26**

where d_w is the degree of the node representing a word on the graph, and α and β are control parameters. The computation of SSR is recursive. It formulates the intuition that two words are similar if they are semantically related to a similar set of neighbours. The recursive equation can be solved by applying matrix algebra. The authors claimed that this method takes into account both explicit semantic relations and implicit semantic connections, and overcomes the coverage limitations of individual background information sources.

Another hybrid method that does not fall under either of the two categories is Agirre et al. (2009a). They proposed to combine a WordNet based method adapted from Hughes and Ramage (2007) with a distributional similarity method based on context similarity. The method adapted from Hughes and Ramage (2007) exploits the link structure of WordNet. However, due to the limited coverage of WordNet, a small proportion of testing data are not covered. To cope with this, for each word that is unknown to WordNet, they firstly applied the distributional similarity method to find several similar words. These are then used to substitute the unknown word and the WordNet based approach is re-applied.

Figure 8.7 below summarises hybrid methods and their connections. The initial background information resources used are noted as: WN –WordNet, WK – Wikipedia, BIO – biomedical knowledge bases, OTH – other structured knowledge bases, WEB – the Web, C – a general purpose corpus, C-BIO – a biomedical corpus.

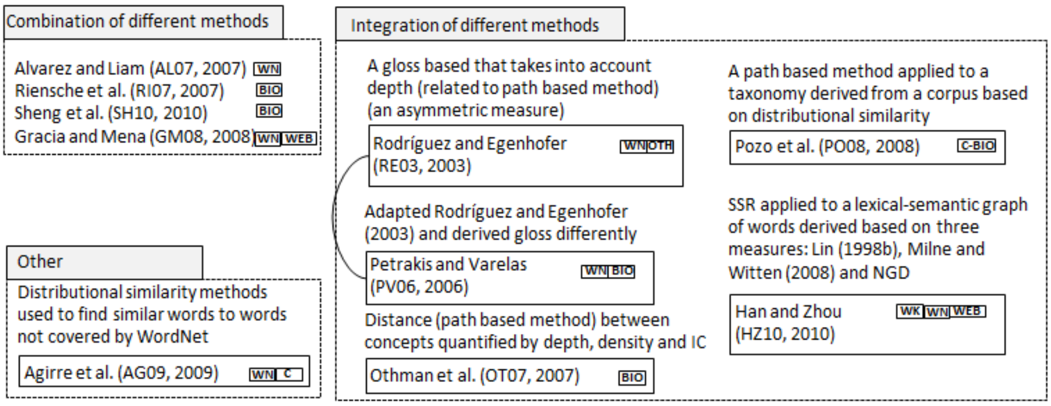

Combination of different methods
Alvarez and Liam (AL07, 2007) WN
Riensche et al. (RI07, 2007) BIO
Sheng et al. (SH10, 2010) BIO
Gracia and Mena (GM08, 2008) WN WEB
Other
Distributional similarity methods used to find similar words to words not covered by WordNet
Agirre et al. (AG09, 2009) WN C
Integration of different methods
A gloss based that takes into account depth (related to path based method) (an asymmetric measure)
Rodríguez and Egenhofer (RE03, 2003) WN OTH
Adapted Rodríguez and Egenhofer (2003) and derived gloss differently
Petrakis and Varelas (PV06, 2006) WN BIO
Distance (path based method) between concepts quantified by depth, density and IC
Othman et al. (OT07, 2007) BIO
A path based method applied to a taxonomy derived from a corpus based on distributional similarity
Pozo et al. (PO08, 2008) C-BIO
SSR applied to a lexical-semantic graph of words derived based on three measures: Lin (1998b), Milne and Witten (2008) and NGD
Han and Zhou (HZ10, 2010) WK WN WEB

Figure 8.7. A summary of hybrid semantic relatedness methods.

8.2.4 Remark

Given the availability of such a wide range of different methods of lexical semantic relatedness, a natural question is how do they compare and how to select the most appropriate method for a task. The remainder of this literature review presents an analysis from several perspectives that can help to answer this question: limitations of different categories of methods, background information resources available and domains, and purpose of the task.

8.2.4.1 Limitations of different categories of methods

Path based methods can be very sensitive to the taxonomic structure of a knowledge base and the density and depth of the structure. They "heavily depend on the degree of completeness, homogeneity and coverage of the semantic links" represented in the structure (Batet et al., 2010). Al-Mubaid and Nguyen (2006) has shown much higher accuracies can be obtained using MeSH than using SNOMED-CT for their method. According to the authors this is attributed to the higher level of specificity (granularity) in the SNOMED-CT concept hierarchy, which has penalised methods that do not address specificity adequately. Strube and Ponzetto's study (Strube and Ponzetto, 2006) has shown similar observations when adapting path based methods from the WordNet structure to a less strict hierarchical structure, the Wikipedia category tree. Earlier path based methods have assumed uniform distance of any edges regardless of the specificity of relevant concepts. This has been the major issue since the assumption proves to be untrue in most real taxonomies. It has been the focus of research and addressed in different ways in later methods. One remaining issue is that most path based methods are based on a single path following a single type of relation, typically *IS-A*. As a result, other useful semantic evidences are overlooked (Wang et al., 2007) and such background information may be insufficient to represent conceptual distance or relatedness between concepts in a semantic network (Lee et al., 1993). It has been shown (Resnik, 1995) that when compared to some of the other categories of methods, path based methods can lead to spurious results, which may be partially attributed to this issue. Nevertheless, compared to other methods, path based methods are generally simple (Budanitsky and Hirst, 2006; Pirro, 2009).

IC based methods inherit some limitations of the path based methods since they exploit the ancestors of two concepts as background information, which is dependent on the taxonomic structure. As a result, they can be also sensitive to the chosen taxonomic structure and a drop of accuracies of some IC based methods has also been observed when they are adapted from WordNet to Wikipedia (Strube and Ponzetto, 2006). To some extent, this sensitivity may be offset by the use of corpus statistics in some IC based methods. The use of a large corpus provides additional useful background information and may have contributed to superior performance in some IC based methods as shown in Batet et al. (2010). However, in some cases such corpora are not always available, especially in the clinic domain where patient records are often highly confidential. In this case, as well as to eliminate corpus pre-processing, methods that approximate the IC of a concept based on the taxonomic structure (e.g., intrinsic IC) may be preferred. Similar to path based methods, IC based methods ignore other potentially useful semantic evidence but employ the *IS-A* relation. Additionally, path based and IC based methods are generally more suitable for measuring semantic similarity, due to their emphasis on the hierarchical relations. Empirically, evaluations by previous studies have shown that path based and IC based methods generally perform better on measuring semantic similarity than relatedness.

Gloss based methods present arguably a cheaper alternative to other kinds of methods since the gathering of background information and the computation are generally less intensive. Due to the lack of comparable evaluations, it is difficult to discuss their limitations with respect to their performance. However, the study by Zesch and Gurevych (2010a) has shown that the Lesk’s method (Lesk, 1986) can be equally sensitive to the underlying background information source as path based and IC based methods.

Vector based methods are generally more extensible than others and adaptation across different background information resources are generally straightforward. Most vector based methods differ in terms of how the concept vectors are created, while share a large degree of commonality in terms of algorithmic calculation. Zesch et al. (2008b) also showed that a concept vector based method can be easily adapted across three different knowledge bases (WordNet, Wikipedia and Wiktionary) and obtaining comparable results. In terms of performance, vector based methods are more balanced for both semantic similarity and relatedness tasks. They are also less sensitive to underlying background information resources, as shown by the results in Zesch et al. (2008b) and Zesch and Gurevych (2010a). This may suggest that vector based methods can be a better option when addressing new domains and datasets. However, the methods of constructing concept vectors differ significantly, and may lead to substantial difference in their performances. Methods that construct concept vectors using other relevant concepts in the semantic graph (Gabrilovich and Markovitch, 2007; Harrington, 2010; Radinskty et al., 2011) require extensive pre-processing of the entire background information resource.

Distributional similarity methods offer a major advantage over other methods – the flexibility in the choice of background information resource. Semantic relatedness methods typically employ a structured knowledge base, whose structure and the coverage and completeness of knowledge may limit the capability of the methods. Distributional similarity, however, is in theory free from such limitation since the underlying corpora can be substituted without incurring changes to the method. This also makes it easily extensible to other domains. Pre-processing a large corpus of millions of documents will be a major issue to be considered with these methods as it creates substantial computational cost (Pantel et al., 2009). In addition, the intrinsic difference between distributional similarity and semantic relatedness should also be considered when they are used as a proxy for semantic relatedness.

Hybrid methods are usually created to combine the strength of different measures. Based on the results which are somewhat inadequate to draw a final conclusion, it seems that they can lead to marginal improvement to their purebred competitors. Thus, before committing to a hybrid method one should consider whether the complexity of the methods can be justified in the particular context. Another consideration is that hybrid methods may inherit the limitations of the purebred methods that are combined, which might explain the decreased performance of some methods (Rodr ǵuez and Egenhofer, 2003; Petrakis et al., 2006) when they are compared against some path based and IC based methods.

Within each category of methods, the modifications introduced in later methods generally lead to higher accuracies than the earlier basic methods. For example, based on the previously published results in individual work and also the comparative evaluation in Zesch and Gurevych (2010a), in the case of path based methods, those (WP94, LC98, LI03, LI07, SF07, TS10, WH11, see corresponding keys on Figures 8.2 - 8.7) addressing concept specificity in a hierarchy generally obtain higher accuracies than those (RA89, HS98) ignoring specificity. In the case of IC based methods, later models (JC97, LN98b, SE04, PI09) have further improved over their ancestor (RE95). In particular, Pirro et al. (2009) demonstrated the superiority of the intrinsic IC (Seco et al., 2004) to the original definition by Resnik (1995) by comparative experiments.

8.2.4.2 Background information resources and domains

Since computing semantic relatedness depends on some background information resources, the choice of such resources will have a major impact on the performance of the methods. Apparently the choice of background information resources is often bound by the methods. Nevertheless, when multiple choices are available, several factors such as the types of information encoded, and the focus as well as the coverage should be considered, and matched against the requirements of the task.

Generally compared to unstructured corpora, structured knowledge bases encode explicit semantic and lexical relations between concepts and entities, which are essential for assessing relatedness and similarity. As discussed in Section 8.2.2, WordNet and Wiktionary are knowledge bases of common words, focusing on nouns, verbs, adjectives and adverbs. They have very limited coverage of proper nouns and specialised concepts. For this reason, they can be a good choice for tasks related to words, such as WSD. Also, the availability of well-defined hierarchical relations allows the knowledge bases to be tailored for semantic similarity other than just relatedness. In comparison, Wikipedia is a vast knowledge base of concepts and entities, which makes it better suited for tasks such as NED. However, the coverage of word knowledge can be very limited, as shown in Zesch and Gurevych (2010a) where methods based on Wikipedia obtained poor performance on the a dataset based on verb pairs. Its broad coverage of a large number of topics also suggests that Wikipedia can be a better choice for domain specific tasks. As an encyclopaedic resource, Wikipedia focuses on covering fact-like information related to concepts and entities and present them as articles. Although Wikipedia articles are intensively hyperlinked, the semantic relations represented by such links are undefined. The category tree used to tag the articles is also a non-strict taxonomy. For this reason, it may be better suited for measuring semantic relatedness; while using Wikipedia for semantic similarity may require adaptation in order to obtain competitive results to WordNet based methods, which has been shown in Ponzetto and Strube (2011).

Research in the biomedical domain generally prefers biomedical knowledge bases. They usually encode hierarchical relations between concepts, and have been used mostly for measuring semantic similarity. Compared to general-purpose knowledge bases, the distribution of knowledge is more uneven, resulting in different densities of regions and incomparable connections (Pozo et al., 2008). This is largely due to the nature of biomedical science, where knowledge is constantly updated, and the empirical knowledge of each concept is highly variable (Li et al., 2010). As a result, methods exploiting the hierarchical structure of such knowledge bases should take into account its unbalanced structure, for example, by addressing specificity of concepts in path based methods.

A major limitation of using structured knowledge bases is that their scope and coverage can limit the capability of the methods. Besides, such resources are usually expensive to maintain, and often unavailable in specific domains. In contrast, unstructured corpora are generally much easier to obtain, which offers the advantage of easier domain adaptation when knowledge bases are unavailable. Background information is often gathered in an implicit form, such as co-occurrence statistics primarily used for distributional similarity. As a result, important semantic evidences encoded in a structured knowledge base are often neglected. Although some methods (Pozo et al., 2008; Harrington, 2010) have proposed to mine hierarchical structures of words from corpora to address this, this often comes at the price of a computationally expensive pre-processing of the entire corpus. Another specific issue in the biomedical domain is “shallow annotation” in biomedical corpora. As mentioned before in Section 8.2.2.3, term frequencies and co-occurrences in the biomedical domain are often gathered based on their usage in annotating gene mentions in a corpus. A potential issue as noted by Sevilla et al. (2005) is that annotators sometimes use more general concepts for annotation even if a more specific concept is more suitable, possibly by mistake or due to their lack of knowledge. As a result, the statistics can be biased towards more general concepts, leading to spurious prediction in methods that depends on corpus statistics (e.g., some IC based methods and distributional similarity methods). Thus the authors suggested that methods that are only based on the topological structure of a semantic graph (e.g., path based methods) should be a better alternative in this domain.

8.2.4.3 Purpose of the task

Another aspect to take into account is the purpose of the task. As discussed before, methods that are particularly tailored for measuring semantic similarity can be found to have inferior performance in assessing general relatedness. They may be better options for tasks such as synonym detection and taxonomy learning; but less effective for WSD or NED where relatedness are more important. Likewise, methods that assess relatedness may also produce spurious predictions of similarity and become unsuitable for some tasks. The lexical units (e.g., words or entities) involved in a task will also have an impact on the choice of underlying background information resources, due to the different focuses of knowledge in these resources.

For tasks built on top of lexical semantic relatedness, the trade-off between accuracy and complexity of the methods will be a major factor particularly in large scale tasks since computing semantic relatedness adds extra pre-processing cost. There is limited work on comparing applications of lexical semantic relatedness. Curran and Moens (2002) compared several distributional similarity metrics in a thesaurus generation task. Budanitsky and Hirst (2006) compared five WordNet based methods in a malapropism application, while Patwardhan and Pedersen (2006) performed similar studies in a WSD task. Bollegala et al. (2007) compared their method against Sahami and Heilman (2006) and Chen et al. (2006) in an entity clustering task. Zesch and Gurevych (2010a) and Tsatsaronis et al. (2010) carried out comparative evaluations of several methods in a word choice application. Where semantic relatedness methods are evaluated both standalone and in applications, there is no strong evidence of a positive correlation between the accuracies of semantic relatedness methods (as in standalone evaluation) and their contribution to the application. For example Patwardhan and Pedersen's method obtained an improvement of 0.15 point in correlation over the Jiang and Conrath's method (Jiang and Conrath, 1997) on a dataset by Rubenstein and Goodenough (1965); however, it was outperformed by this method when applied to a WSD task. Similarly Zesch and Gurevych (2010a) showed that several better performing methods in the in-vitro evaluation achieved lower accuracies in a word choice application.

Given these observations, one may want to re-consider their requirements for the accuracies of semantic relatedness methods when choosing one for their applications; particularly when dealing with a large amount of data such as in WSD or sense clustering.

8.2.5 Towards a new method

With respect to the method of lexical semantic relatedness to be proposed in the next sections of this study, it has been noted that different methods have largely focused on using a single source of background information. In particular, where a structured knowledge base is used, literature has been predominantly based on a single knowledge base.

The benefits of using multiple resources have already been recognised. For example, IC based methods aims to combine the structural information of a concept with its actual usage patterns for measuring semantic relatedness. The former is assessed based on a structured knowledge base, and the latter is assessed based on corpus statistics. However, all IC based methods have only used a single structured knowledge base. The use of corpus statistics as a second source of evidence has also been criticised (Seco et al., 2004), since it makes the method more dependent on the selection of an appropriate corpus. The work by Han and Zhao (2010) makes one step closer: three state-of-the-art methods each based on a single source of background information are used, and their scores are aggregated and further used as features for evaluating semantic relatedness in a different measure. However, in this approach, each method is isolated from one another and different resources are used separately. Essentially, each component method still uses only a single resource. The approach does not combine different background information resources in a uniform approach.

Considering these limitations and the observations that different knowledge bases share a high degree of commonality and their possible complementary nature, this study further proposes a different method of lexical semantic relatedness based on the principles of combining different background information resources – particularly knowledge bases – in a uniform semantic relatedness measure. Details of this are to be presented in the following sections.

8.3 Hypothesis

The proposed method of lexical semantic relatedness is based on the following hypothesis:

H3.2 Lexical semantic relatedness: lexical semantic relatedness measures can benefit from combining different background information resources since they complement each other in certain ways.

As discussed in the previous section, the benefits of combining evidence from multiple background information resources have been recognised and partially attested by IC based methods that use corpus statistics in additional to structured knowledge bases. However, it has also been shown that such kind of methods can be overly dependent on the external corpus, which, if not selected properly, can harm the accuracies of a measure (Seco et al., 2004; Pirro, 2009).

Instead, this work studies the combination of multiple *knowledge bases* in measuring lexical semantic relatedness for two reasons. On the one hand, no work has explored the potential of combining multiple knowledge bases in a single semantic relatedness measure. On the other hand, as discussed in the previous sections, many knowledge bases have shown two properties that suggest potential advantages by combination: the commonality and the complementary natures.

The commonality nature

It has been discussed before that many knowledge bases share a high degree of commonality in terms of the knowledge units covered, the types of information encoded and the structure used for organising the information. For example, both WordNet and Wiktionary are lexicalised ontology of words. They share a large overlap of vocabulary and encode similar types of lexical and semantic relations. Wikipedia is a knowledge base of entities and concepts. It also includes a hierarchical structure of concepts that encodes taxonomic and meronymy relations. Such commonality has been partially demonstrated in studies that adapt semantic relatedness methods to different knowledge bases (Strube and Ponzetto, 2006; Zesch et al., 2008b; Zesch and Gurevych, 2010a). Typically, these studies have shown that different knowledge bases share an overlap of vocabulary and often encode equivalent or similar lexical semantic information. As a result, same semantic relatedness methods can be adapted to different knowledge bases, while achieving reasonable results.

The complementary nature

It has also been discussed before that different knowledge bases are designed with different purposes, thus placing different focuses in terms of their content and structure. For example, due to the collaborative nature of Wiktionary, it has covered in general a larger vocabulary than WordNet, particularly neologisms and acronyms. Both serves as a general purpose dictionary and therefore, have limited coverage of proper nouns (e.g., entity names) and specialised concepts; but focuses on the lexical and semantic relations between words and concepts. Wikipedia, on the other hand, is an encyclopeadia that focuses on fact-like knowledge of entities and specialised concepts. Wikipedia content is highly interlinked. However, the links are not semantified and Wikipedia does not explicitly define lexical semantic relations between entities and concepts.

Such differences have led to different performances of same methods when adapted to different resources and applied to different datasets. On the other hand, this may suggest that different knowledge bases can be complementary. For a concrete example, the word 'mouse' and 'fox', and their 'mammal' senses are considered. The Wikipedia articles of 'mouse' and 'fox' describes the two concepts as:

mouse*: A mouse (plural: mice) is a small mammal belonging to the order of rodents.*

fox*: Fox is a common name for many species of omnivorous mammals belonging to the Canidae family.*

A simplistic overlap based method that determines relatedness based on the word overlap between the two descriptions will return a numeric value of 1, since they share only 1 word 'mammal'. The WordNet[10] glosses of the two concepts are:

mouse*: Any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails.*

fox*: Alert carnivorous mammal with pointed muzzle and ears and a bushy tail; most are predators that do not hunt in packs.*

Using the same approach based on these two descriptions, the relatedness score will be 2, since they share two words 'ear' and 'tail'. Apparently, the descriptions from both resources are correct and in fact present complementary information. If the two descriptions for each concept are merged, a higher relatedness score of 3 can be obtained.

Following these discussions, it is hypothesized that lexical semantic relatedness methods can benefit from a combination of multiple knowledge bases. One natural solution is to create a joint representation of a concept based on the information encoded in different

[10] All examples based on WordNet use WordNet version 2.1 for Windows OS.

knowledge bases. This lays the fundamental principle of the proposed method, which is to be detailed in the next section.

8.4 Lexical Semantic Relatedness based on Multiple Knowledge Bases

The proposed method is based on the idea of creating a joint feature representation of concepts using the background information encoded in different knowledge bases. It is divided into four steps. Firstly (Section 8.4.2), each word or phrase is searched in each knowledge base to identify their **contexts** that is specific to that knowledge base. A context is defined as the description of meaning or a concept for a word. It is associated with a unique entry in the knowledge base, often denoting a distinct concept. Formally, an input word *w* can map to multiple contexts in a knowledge base $CTX_{KB}(w)$, i.e., $CTX_{KB}(w) = \{ctx_{KB}(c_1), \dots\ ctx_{KB}(c_n)\}$. Secondly (Section 8.4.3), for each context of an input word, different features are extracted to create a representation of the context. Thirdly (Section 8.4.4), cross-source contexts are mapped where they refer to the same meaning, thus their features from different resources can be combined to derive a joint representation. This creates a final, uniform feature representation, which is then used to compute semantic relatedness using a vector based similarity function (Section 8.4.5).

8.4.1 Choice of Knowledge Bases

Three knowledge bases are selected for this study: Wikipedia, WordNet and Wiktionary. Wikipedia and WordNet have been widely used for measuring lexical semantic relatedness. In particular, WordNet has also been used for WSD and Wikipedia has shown to be an effective knowledge base for NED. Although little work has explored Wiktionary for this task, Zesch et al. (2008b) showed that it can be equally competitive. Also, considering the high degree of commonality between Wiktionary and WordNet, the combination of the two resources can be straightforward.

8.4.2 Context retrieval

Given a pair of words or phrases (each denoted by *w*), the contexts $CTX_{KB}(w)$ representing the underlying meanings or concepts from each knowledge bases are retrieved.

WordNet (*wn*)

In WordNet, a context is a single synset, which corresponds to a concept. For each word in WordNet all possible synsets are extracted. Let $CTX_{wn}(w)=\{ctx_{wn}\ (syn_1),\ \ldots\ ctx_{wn}\ (syn_{\mathrm{n}})\}$ denote the set of different contexts of the word *w*, each defined by a synset in WordNet. $ctx_{wn}(syn_{\mathrm{n}})$ is a function that builds a feature representation of a context based on the synset *n* from WordNet. This will be explained further in Section 8.4.3.

Wiktionary (*wkt*)

In Wiktionary, a context corresponds to a single Wiktionary entry. As previously described in Section 8.2.2.1, each entry in Wiktionary defines all possible meanings for one *word class (wc)* of a word. For example, the Wiktionary[11] entry for word 'dog' with the noun word class defines 12 different senses, mapping to 12 different concepts. The lexical and semantic relations are defined between word classes, rather than senses. Let $CTX_{wkt}(w) = \{ctx_{wkt}\ (wc_1),\ \ldots\ ctx_{wkt}\ (wc_{\mathrm{n}})\}$ denote the set of different contexts of *w*, each defined by a word class in Wiktionary. $ctx_{wkt}(wc_{\mathrm{n}})$ is a function that builds a feature representation of a context based on the word class *n* from Wiktionary. This will be explained further in Section 8.4.3.

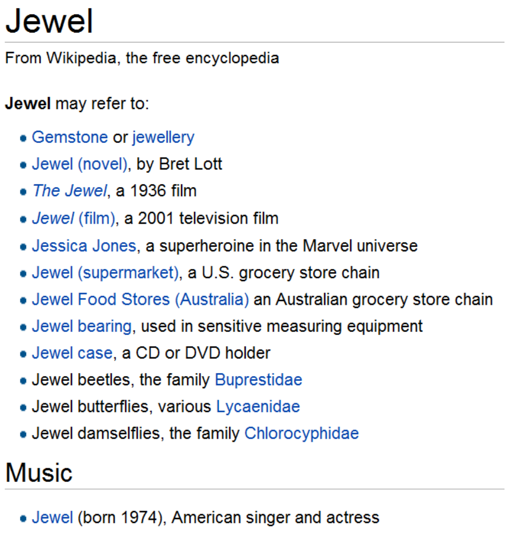
Wikipedia (*wk*)

In Wikipedia, a context is an article that describes a unique concept or entity. A word or phrase *w* is firstly searched in Wikipedia. As previously discussed in Chapter 6, three situations can be anticipated: a single non-disambiguation page describing a single concept or entity is returned if the search term is not ambiguous to Wikipedia, or there is a commonly used sense defined in Wikipedia; a disambiguation page listing all underlying concepts and entities referenced by *w*, if *w* is highly ambiguous (Figure 8.8); or nothing if there are no concepts or entities referenced by *w* in Wikipedia. In the first case, the single article page is used as the context for *w*. In the third case, the most relevant page is retrieved by searching *w* as keyword(s) in an inverted index of all Wikipedia pages (this is done by searching *w* on Google and select the first Wikipedia page whose title matches *w*). In the second case, the disambiguation page is further processed by following heuristics to select the candidate pages:

- Select the article page listed as the first entry on the page – often, the first entry refers to the most commonly used sense for *w*.

[11] All examples based on Wiktionary use a version obtained on the 14 Feb 2012.

- For all the other entries that do not link to further disambiguation pages, select those whose link text (parentheses and contents within are trimmed, e.g., 'Jewel (film) => Jewel') matches *w* – because many of these in fact link to a concept relevant to *w*, but not necessarily a candidate sense of *w* (e.g., 'jewel case').



Jewel

From Wikipedia, the free encyclopedia

Jewel may refer to:

- Gemstone or jewellery
- Jewel (novel), by Bret Lott
- *The Jewel*, a 1936 film
- *Jewel* (film), a 2001 television film
- Jessica Jones, a superheroine in the Marvel universe
- Jewel (supermarket), a U.S. grocery store chain
- Jewel Food Stores (Australia) an Australian grocery store chain
- Jewel bearing, used in sensitive measuring equipment
- Jewel case, a CD or DVD holder
- Jewel beetles, the family Buprestidae
- Jewel butterflies, various Lycaenidae
- Jewel damselflies, the family Chlorocyphidae

Music

- Jewel (born 1974), American singer and actress

Figure 8.8. An example of a Wikipedia disambiguation page for the word 'jewel'.

- The second step above can discard many named entities, which are often referred by shorthands. For example, surnames of a person are often used to refer to a person entity instead of the person's full name (e.g., 'Jackson', 'Jordan'). Also, if *w* is capitalised, it is likely to denote a named entity. In this case, the candidate selection phase is expanded by selecting entries that are likely named entities. This is done by selecting entries whose link texts are multi-word capitalised phrases that contain *w* (e.g., 'The Jewel' will also be selected if the input word is 'Jewel' instead of 'jewel').
- For any entries that link to further disambiguation pages, retrieve those disambiguation pages and repeat the above candidate selection processes. If more disambiguation links are found on these pages, they are not further processed.
- Pages that contain less than 100 words are discarded – this is to eliminate pages that are incomplete, or created as skeletons to invite contributors for further editing.

Thus the search for a word *w* in Wikipedia can return one or multiple non-ambiguous article pages, each describing a unique concept or entity. Let $CTX_{wk}(w) = \{ctx_{wk}\ (p_1),\ \ldots\ ctx_{wk}\ (p_n)\}$ denote the set of different contexts of the word *w*, each defined by an article in

Wikipedia. $ctx_{wk}(p_n)$ is a function that builds a feature representation of a context based on the article page *n* from Wikipedia. This will be explained further in Section 8.4.3.

8.4.3 Feature extraction and representation

Next, for each context identified from a knowledge base for a word *w*, features are extracted to represent the context.

Wikipedia

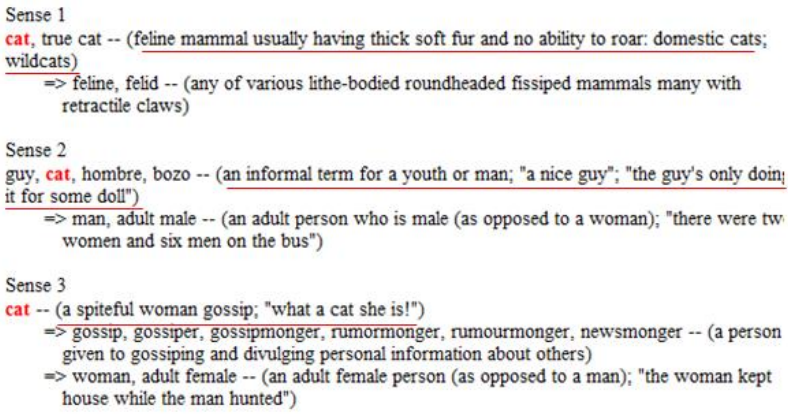
For each element in $CTX_{wk}(w) = \{ctx_{wk}(p_1), \ldots ctx_{wk}(p_n)\}$, the following types of features are extracted from the Wikipedia article page:

- *wiki-title*: These are words from the title of the page, and the redirection links of the page. As discussed before, the redirection links encode name aliases associated with a term. These can be considered as equivalent to the title.
- *wiki-cat*: These are words from the category labels assigned to the page. Category labels have been used as equivalence to hypernyms of the concept described by a Wikipedia article page (Strube and Ponzetto, 2006). The same filtering strategies introduced previously in Chapter 6 are adopted to discard noisy category labels that are created for data archiving purposes by Wikipedia.
- *wiki-link*: Words from links on the page. Other Wikipedia articles linked by the current page are often relevant entities or concepts. For each link, the 'target' of a link rather than the 'surface' of a link is taken. For example, in the page about 'Queen (chess)', the first sentence 'The Queen is the most powerful piece in the game of chess' contains a link 'Chess_piece' with surface form 'piece'; in such case, the target phrase 'Chess_piece' is taken and converted to two words 'Chess' and 'piece'. A frequency threshold of 2 is used to filter *wiki-link* words. This is because, Wikipedia links can represent a wide range of semantic and non-semantic associations, the strength of which can vary to a large degree. A simple frequency threshold can filter out words that are potentially less relevant.
- *wiki-word*: Words that have at least two occurrences on the page are selected.

For all feature types, the same stopwords list used by the previous chapters is used to filter out noisy and often meaningless words.

WordNet

Previous studies (Gurevych, 2005) have proposed to represent WordNet synsets by concatenating concepts in close relations such as hypernymy and synonymy. Based on this principle, for each element in $CTX_{wn}(w) = \{ctx_{wn}(syn_1), \dots ctx_{wn}(syn_n)\}$, ten types of features are extracted to represent a WordNet synset: hypernyms, hyponyms, meronyms, holonyms, synonyms, antonyms, attributes, 'see also' words, 'related' words, and gloss. For example, Figure 8.9 below shows three synsets defined for the word 'cat', with their gloss highlighted by underlines and synonyms indicated by '=>'.



Sense 1
cat, true cat -- (feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats)
=> feline, felid -- (any of various lithe-bodied roundheaded fissiped mammals many with retractile claws)

Sense 2
guy, cat, hombre, bozo -- (an informal term for a youth or man; "a nice guy"; "the guy's only doin; it for some doll")
=> man, adult male -- (an adult person who is male (as opposed to a woman); "there were tw women and six men on the bus")

Sense 3
cat -- (a spiteful woman gossip; "what a cat she is!")
=> gossip, gossiper, gossipmonger, rumormonger, rumourmonger, newsmonger -- (a person given to gossiping and divulging personal information about others)
=> woman, adult female -- (an adult female person (as opposed to a man); "the woman kept house while the man hunted")

Figure 8.9. Example synsets and features for the word 'cat' in WordNet

Wiktionary

Wiktionary and WordNet share a high degree of similarity in terms of the structure and content encoded for words. Eight out of the ten WordNet feature types are also defined in Wiktionary, namely, hypernyms, hyponyms, meronyms, holonyms, synonyms, antonyms, 'see also' words, and gloss. Therefore, each word class in $CTX_{wkt}(w) = \{ctx_{wkt}(wc_1), \dots ctx_{wkt}(wc_n)\}$ is represented using these features.

Furthermore, the ten WordNet feature types and eight Wiktionary feature types are regrouped to create four types of features for two reasons. First, WordNet and Wiktionary features can be very sparse. For example, the WordNet gloss can be very short since the focus of WordNet is to define word senses and their lexical and semantic relations. Also, certain relations are only defined for certain word classes. For example, hypernyms are only defined for nouns and verbs. Examples will be given later in the experiment section. Wiktionary features suffer from the same limitation. Grouping different types of features effectively increases possible feature values for each feature type, improving the possibility of shared features between different WordNet synsets or Wiktionary word classes. Second, the features are regrouped in a way to enable mapping similar feature types across different knowledge bases, which will be further discussed in Section 8.4.4.2. For WordNet, the ten types of features are regrouped as:

- *wn-synant* merges WordNet synonyms and antonyms.
- *wn-hypoer* merges WordNet hypernyms and hyponyms, collectively representing features by '*is-a'* semantic relation
- *wn-link* merges WordNet meronyms, holonyms, related and 'see also', which are features corresponding to any associative relations
- *wn-word* merges WordNet gloss and attributes that generally describe a concept.

Similarly, the eight types of Wiktionary features are re-grouped to create four equivalent types of features:

- *wkt-synant* merges Wiktionary synonyms and antonyms.
- *wkt-hypoer* merges Wiktionary hypernyms and hyponyms, collectively representing features by '*is-a'* semantic relation
- *wkt-link* merges Wiktionary meronyms, holonyms, and 'see also', which are features corresponding to any associative relations
- *wkt-word* which is the Wiktionary gloss that generally describe a concept.

8.4.4 Combining Knowledge Bases

Next, the representations created for the contexts defined by each knowledge base are mapped to create a joint context and feature representation. This involves solving two subtasks: mapping the contexts and the feature types defined by different knowledge bases.

8.4.4.1 Context mapping

In order to create a joint feature representation based on multiple knowledge bases, the first step is to map the contexts across different knowledge bases such that they refer to similar meanings. To do so, a reference knowledge base is firstly selected and the contexts extracted for the word from this knowledge base is chosen as the base set of contexts; then contexts extracted from other knowledge bases are mapped to reference contexts of similar meanings. Empirically, Wikipedia is chosen as the reference knowledge base for two reasons: 1) it has the broadest coverage of knowledge units, particularly specialised concepts and named entities, which can be critical for the task of NED; 2) the encyclopaedic nature means a richer feature representation, since the articles contain much more detailed descriptions that those available in WordNet or Wiktionary.

Following this strategy, $CTX_{wk}(w)$ is chosen as reference contexts, and for each article page in $CTX_{wk}(w)$, the closest synset and word class is selected from $CTX_{wn}(w)$ and $CTX_{wkt}(w)$ respectively, creating a collection $CTX_{mapped}(w)= \{ ctx_m^1 \dots ctx_m^n \}$, where $|CTX_{mapped}(w)|=|CTX_{wk}(w)|$ and each element ctx_m^n is a triple $\{ctx_{wk}(p), ctx_{wn}(syn), ctx_{wkt}(wc)\}$ denoting a Wikipedia article page (p) mapped with a WordNet synset (syn) and a Wiktionary word class (wc). To select the closest synset and word class from WordNet and Wiktionary, a simple maximum feature overlap based measure is used. Let $F(\cdot)$ be a function that returns all feature values of a context as bag-of-words, then for each $ctx_{wk}(p)$, it is mapped to a synset such that $|F(ctx_{wk}(p))| \cap |F(ctx_{wn}(syn)|$ is maximised among all candidate synsets in $CTX_{wn}(w)$. The same procedure is performed to select a single context from $CTX_{wkt}(w)$. It is also likely that no WordNet-contexts or Wiktionary-contexts can be mapped when $CTX_{wkt}(w)= \emptyset$ or $CTX_{wn}(w)= \emptyset$, or when $F(\cdot) = \emptyset$.

8.4.4.2 Feature mapping

Next, given the set of cross-mapped contexts of a word $CTX_{mapped}(w)$, a joint feature representation is created for each cross-mapped context ctx_m^n. This is done by mapping different types of features extracted from each knowledge base to create a joint set of features to represent a context. Two different approaches are proposed.

Feature integration

With feature integration, the re-grouped four WordNet feature types and four Wiktionary feature types are mapped to the equivalent Wikipedia feature types, and collapsed to create a set of four joint types of features:

- *merged-synant* merges *wiki_title*, *wn-synant* and *wkt-synant*. Intuitively, *wiki_title* contain words that can be aliases to a concept or entity, denoting a sense of ‘synonyms’.
- *merged-hypoer* merges *wiki_cat* with *wn-hypoer* and *wkit-hypoer*. Wikipedia categories are treated as hypernyms.
- *merged-link* merges *wiki_links*, *wn-link* and *wkt-link*, all of which represent a general sense of association.

- *merged-word* merges *wiki_word, wn_word* and *wkt_word*, all of which can be considered as general descriptions of a context.

Therefore, with feature integration, each mapped context is represented using the four *joint feature types*, creating a joint feature representation.

Feature combination

While feature integration creates a joint feature representation by merging similar feature types from different knowledge bases, an alternative approach would be to simply collect different types of features extracted from each knowledge base while retaining the diversity in feature types, i.e., a mapped context will be represented by 12 types of features, including four types of Wikipedia features, four types of WordNet features and four types of Wiktionary features.

The difference between the two approaches is that feature combination introduces more *types* of features, whereas feature integration retains a small number of feature types but increases the number of possible feature values for each type. Since two contexts only share features via same feature types, feature combination improves this possibility by adding more feature types, while feature integration achieves the same goal by adding more feature values for each type.

Feature diversification

In addition, a different feature of the proposed approach from many existing methods is feature diversification, i.e., a representation using multiple feature types rather than a simple bag-of-words model. Existing methods mostly employ a single type of feature. For example, most gloss based methods (Lesk, 1986; Banerjee and Pedersen, 2003; Gurevych, 2005; Gentleman, 2005) create a bag-of-words representation that combines words gathered by all kinds of lexical semantic relations defined in a knowledge base; Zesch et al. (2008b) proposed to use only the first paragraph of Wikipedia. To contrast, this single-typed feature representation will be referred to as **feature unification**. To compare the two different designs, the effects of feature unification are also studied. This is done by simply collapsing all feature types into a single type that includes all feature values. Detailed settings will be discussed in the experiment section.

8.4.5 Computing Semantic Relatedness

Following the above discussion, given a pair of words or terms, each word is mapped to a set of contexts $CTX_{mapped}(w)$ where each element is a triple $\{ctx_{wk}(p), ctx_{wn}(syn), ctx_{wkt}(wc)\}$ denoting that a Wikipedia article page (*p*) is mapped with a WordNet synset (*syn*) and a Wiktionary word class (*wc*) that are likely to describe similar meanings with *p*. A joint feature representation is created for each element. The next step is to compute the relatedness between pair-wise contexts for the two words. For this, the simple cosine vector similarity function is used.

Specifically, given two mapped contexts and their joint feature representations, two feature vectors are created. For each feature type *ft*, the unique feature values found in both representations are firstly gathered to create a set of unique feature values for that type, i.e., $\{fv \mid Type(fv) = ft\}$ where *Type(fv)* states that the type of the feature *fv* is *ft*. Next, for each context, a feature vector $\vec{v}$ with the size equal to the sum of all unique values of all feature types is created, where each element corresponds to a unique feature of a specific type that is found in the feature representations of either context. The value of this element is assigned as below:

- 0 if the feature value does not exist in the representation;
- $freq(fv) \cdot \frac{1}{\sum_{Type(fv')=ft} freq(fv')}$, where *freq(fv)* returns the number of occurrences of a feature value among all features that belong to the same feature type. This feature weighting function effectively places a uniform weight for each occurrence of a feature of a specific type. As a result, the feature that is found more frequently receives a higher weight.

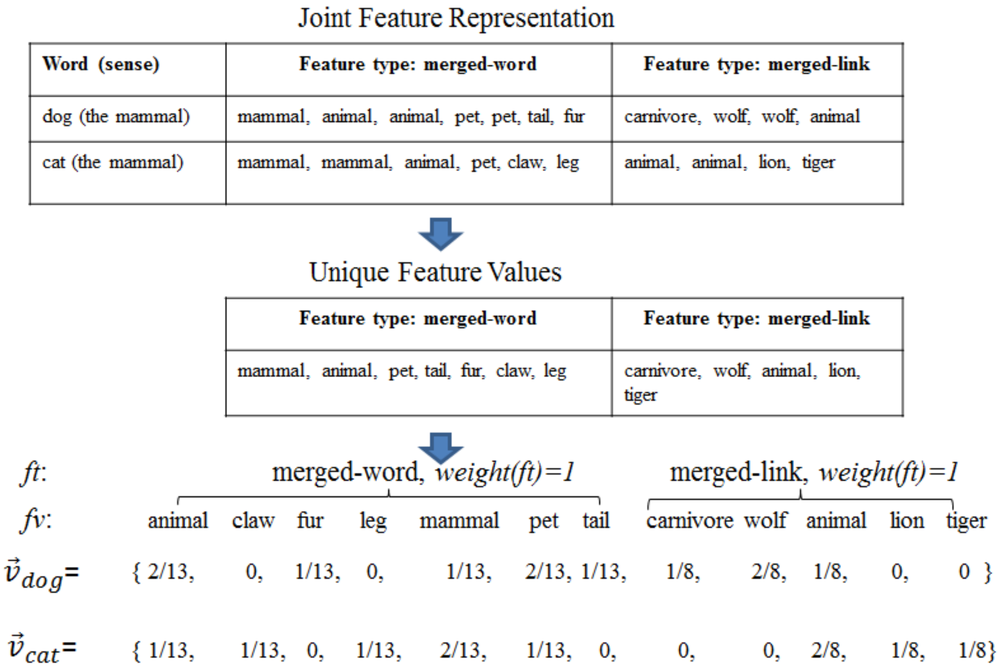
An example of the feature vector creation is illustrated in Figure 8.10.

Thus the semantic relatedness between two joint contexts is computed using the cosine similarity function (Equation 8.18) using their feature vectors. The relatedness between two polysemous words is derived as the maximum pairwise context relatedness for their contexts $CTX_{mapped}(w_1)$ and $CTX_{mapped}(w_2)$. The cosine function is chosen because it is the most widely used metric for assessing vector based similarity and is found generalisable across different problems and very effective (Gabrilovich and Markovitch, 2007; Zesch and Gurevych, 2010a). Although other metrics such as KL-divergence (Hughes and Ramage, 2007) can also be used for the same purpose, they are not studied in this work since the focus of this work is feature engineering with multiple knowledge bases for semantic relatedness, not comparing the effectiveness of different similarity metrics.

8.5 Evaluation and Discussion

8.5.1 Method

Methods for measuring lexical semantic relatedness are typically evaluated by two types of approaches: *in-vitro* and *in-vivo*. In *in-vitro* experiments semantic relatedness scores of concepts or words are compared directly against a gold standard. In *in-vivo* experiments the methods are evaluated indirectly by the performance of an application built on top of it. This study focuses on in-vitro evaluation, while Chapter 9 presents an evaluation of a NED method that is built on top of the proposed semantic relatedness method.



Joint Feature Representation

Word (sense)	Feature type: merged-word	Feature type: merged-link
dog (the mammal)	mammal, animal, animal, pet, pet, tail, fur	carnivore, wolf, wolf, animal
cat (the mammal)	mammal, mammal, animal, pet, claw, leg	animal, animal, lion, tiger

Unique Feature Values

Feature type: merged-word	Feature type: merged-link
mammal, animal, pet, tail, fur, claw, leg	carnivore, wolf, animal, lion, tiger

ft:	merged-word, *weight(ft)=1*							merged-link, *weight(ft)=1*				
fv:	animal	claw	fur	leg	mammal	pet	tail	carnivore	wolf	animal	lion	tiger
$\vec{v}_{dog}$=	{ 2/13,	0,	1/13,	0,	1/13,	2/13,	1/13,	1/8,	2/8,	1/8,	0,	0 }
$\vec{v}_{cat}$=	{ 1/13,	1/13,	0,	1/13,	2/13,	1/13,	0,	0,	0,	2/8,	1/8,	1/8}

Figure 8.10. Creation of feature vectors

Typically in an *in-vitro* evaluation, a dataset containing a set of pairs of concepts or words are presented to human judges, who subjectively estimate the relatedness between each pair within a certain scale. Next, the semantic relatedness method is applied to the same dataset. The results are correlated against the human judgement to derive an indication of the accuracy of the method.

Two correlation functions are often used in the literature, the Pearson correlation coefficient and the Spearman rank order correlation coefficient. The Pearson correlation compares the scores computed by a semantic relatedness method with the numeric scores of the gold standard. For example, the pair *table-chair* might get a human judgement of 8/10. With Pearson correlation, a machine computed score of 7/10 will be awarded higher than a score of 4/10, since the first score is closer to the gold standard. Pearson correlation returns a value of 1 for perfect correlation and a value of 0 for no correlation. Given *X* the list of scores assigned to a list of term pairs by humans, and *Y* the scores assigned to the same list by a semantic relatedness method, it is calculated as below:

$$Pearson(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}} \qquad \textbf{Equation 8.27}$$

The Spearman rank order correlation coefficient is based on rankings of data series. For this method, the list of term pairs are ranked by their scores given by the human judge and a measure respectively, and correlation strength is assessed based on how well the ranking given by the measure resembles that by the human. Empirically it is calculated by the same formula for the Pearson correlation coefficient by simply replacing the lists of scores *X* and *Y* with the lists of ranks of these scores.

Zesch and Gurevych (2010a) discussed the limitations of each correlation measures and argued in favour of the Spearman measure. Firstly, the Pearson correlation function is very sensitive to outliers – a single outlier may produce a significantly different result. They further justified this in an experiment and showed that on the dataset by Miller and Charles (1991), the method by Lesk (1986) was significantly penalised by a single word pair that received an extraordinarily high score representing an outlier in the dataset. Secondly, it measures the strength of the linear relationship between two data series and can produce flawed results when the relationship between human judgements and computed scores is non-linear. Thirdly, it requires normal distribution of two random variables (scores given by a human judge and a measure) and scores to be normalised within certain interval scales. However, studies have shown that semantic relatedness scores are not always interval scaled (Budanitsky and Hirst, 2006; Zesch and Gurevych, 2007). For these reasons, the Spearman rank order correlation coefficient is used for evaluation.

8.5.2 Dataset

Gold standard datasets for correlation analysis are typically created by asking multiple human annotators to rate the semantic relatedness of a pair of concepts or terms within a certain scale, and then averaging their interpretations. Since the interpretation is subjective, inter-annotator agreement (IAA) should also be studied.

Studies of lexical semantic relatedness in the general domain have predominantly evaluated their methods using general purpose datasets. Several biomedical datasets are also available in the same format, but have not been used in these studies. For a thorough evaluation, the proposed method is evaluated using both general purpose datasets and biomedical datasets.

In the general domain, three datasets and their variants have been widely used. These are the Rubenstein and Goodenough (1965) dataset of 65 pairs of nouns (RG65), the Miller and Charles (1991) dataset that is a subset of the RG65 dataset and containing 30 pairs (MC30), and the Finkelstein et al. (2002) dataset containing 353 pairs of words (Fin353). No IAA figures were reported for the original RG65 and MC30 datasets. Pirro and Seco (2008), and Resnik (1995) re-created these datasets with IAA analyses. Also, the Fin353 dataset contains two subsets, each containing 153 pairs and 200 pairs respectively and annotated by different groups of annotators. Zesch and Gurevych (2010a) carried out IAA analysis and discover largely varying figures for the two subsets, and therefore suggested treating them separately (Fin153 and Fin200) in evaluation. The RG65 and MC30 datasets are assessed based on similarity, and are therefore originally used for evaluating semantic similarity methods. Fin153 and Fin200 are assessed based on relatedness.

One widely known limitation of these datasets is that they do not contain multi-word phrases, and only the Fin153 and Fin200 datasets contain very limited number of named entities. To address this issue, Ziegler et al. (2006) created two English datasets of concept instances and named entities. The first dataset (Zie25) contains 25 pairs, annotated by 23 people; the second (Zie30) contains 30 pairs, annotated by 51 people. 87% of annotators are German native speakers. These datasets have rarely been used. Since the semantic relatedness method will be applied to an NED task, it is important to evaluate its capability for finding related named entities. Therefore, these datasets are also included for evaluation.

In the biomedical domain, two datasets have been used for evaluating semantic relatedness between terms. Petrakis et al. (2006) compiled a set of 49 MeSH term pairs and asked 12 experts to assess the relatedness of these pairs. Pairs with standard deviation above 0.8 were excluded, resulting in a total of 36 term pairs (MeSH36). Pedersen et al. (2006) created a set of 120 pairs of medical terms extracted from the SNOMED-CT terminology annotated by 13 experts. However, they obtained a very low IAA of 0.51 on this dataset. To create a more reliable testbed, they selected a subset of 30 pairs and asked another two groups of experts – including a group of 3 physicians and a group of 9 medical coders – to re-annotate them. The final dataset (Ped30p and Ped30c) has IAA figures of 0.68, 0.78 respectively and a cross-group IAA of 0.85.

In total, nine datasets are selected for evaluating the semantic relatedness method. The RG65, MC30, Fin153 and Fin200 datasets focus on *general domain common English words*; the Zie25, Zie30datasets focus on *named entities*; and the MeSH36, Ped30p and Ped30c focus on *domain specific terminologies*, most of which are multi-word units. Table 8.2 below summarises these datasets, N – noun, V – verb, A – adjective, NE – named entity, T – domain specific terminology.

Dataset	Pairs	PoS	Similarity/Relatedness	IAA
RG65	65	N	Similarity	0.80
MC30	30	N	Similarity	0.90
Fin153	153	N, V, A	Relatedness	0.73
Fin200	200	N, V, A	Relatedness	0.55
Zie25	25	NE	Relatedness	-
Zie30	30	NE	Relatedness	-
MeSH36	36	T	Relatedness	-
Ped30p	30	T	Similarity	0.68
Ped30c	30	T	Similarity	0.78

Table 8.2. Datasets for evaluation

8.5.3 Setting

The versions of knowledge bases and corresponding APIs used for accessing their contents are as below:

- Wikipedia – a version dated 6 Feb 2007 is used and the JWPL API (Zesch et al., 2008a) is used for accessing the data. This version is chosen since it is also used by Zesch and Gurevych (2010a). Therefore the results can be directly compared.
- WordNet – WordNet 2.1 for Windows OS is used. This is the most recent version of WordNet for Windows OS and is consistent with WordNet 3.0 for Linux OS used by Zesch and Gurevych (2010a).

- Wiktionary – a version dated 21 Mar 2011 is used and the JWKTL API (Zesch et al., 2008a) is used for accessing the data.

The main goal of the experiment is to study how the different feature representations created based on multiple knowledge bases affect the performance of the proposed semantic relatedness method. It is further divided into three sub-objectives: *1) test the effect of combining multiple knowledge bases; 2) test the effect of feature integration v.s. combination; 3) test the effect of feature diversification v.s. unification.* Different feature representation settings are created for these purposes.

In order to test the effect of combining multiple knowledge bases, the joint feature representation is created on an incremental basis and tested separately:

- Wikipedia/WordNet/Wiktionary features only (*wk/wn/wkt*)
- Features from Wikipedia + WordNet (*wkwn*)
- Features from Wikipedia + Wiktionary (*wkwkt*)
- Features from Wikipedia + WordNet + Wiktionary (*wkwnwkt*)

As a result, a total of 18 settings are created for experiments. For each individual knowledge base:

- Four Wikipedia features only (*wk4*), one Wikipedia feature only (*wk1*)
- Four WordNet features only (*wn4*), one WordNet feature only (*wn1*)
- Four Wiktionary features only (*wkt4*), one Wiktionary feature (*wkt1*),

For the proposed joint feature representations obtained by using multiple knowledge bases, different feature mapping methods, and feature diversification or unification:

Two knowledge bases:

- Wikipedia + WordNet, feature *combination, four* types of features from *each* knowledge base (*wkwn4c*)
- Wikipedia + WordNet, feature *integration, four joint* types of features (*wkwn4i*) from *both* knowledge bases
- Wikipedia + WordNet, feature *combination*, *one* type of features from *each* knowledge base (*wkwn1c*)

- Wikipedia + WordNet, feature *integration, one joint* type of feature from all knowledge bases (*wkwn1i*)
- Wikipedia + Wiktionary, feature *combination, four* types of features from *each* knowledge base (*wkwkt4c*)
- Wikipedia + Wiktionary, feature *integration, four joint* types of features (*wkwkt4i*) from *both* knowledge bases
- Wikipedia + Wiktionary, feature *combination, one* type of features from *each* knowledge base (*wkwkt1c*)
- Wikipedia + Wiktionary, feature *integration, one joint* type of feature from all knowledge bases (*wkwkt1i*)

Three knowledge bases:

- Wikipedia + WordNet + Wiktionary, feature *combination, four* types of features from *each* knowledge base (*wkwnwkt4c*)
- Wikipedia + WordNet + Wiktionary, feature *integration, four joint* types of features from *all* knowledge bases (*wkwnwkt4i*)
- Wikipedia + WordNet + Wiktionary, feature *combination, one* type of features from *each* knowledge base (*wkwnwkt1c*)
- Wikipedia + WordNet + Wiktionary, feature *integration, one joint* type of features from *all* knowledge bases (*wkwnwkt1i*)

Experiments are performed with each of these settings, using the nine datasets described before. Results are presented and discussed in the following.

8.5.4 Results and Discussion

8.5.4.1 Coverage

The coverage of each knowledge base is firstly studied. For each dataset, each pair of words or terms is searched in each knowledge base. Table 8.3 shows *the number of pairs that include at least one word or term that are not covered* by a knowledge base. For brevity, this is referred to as *pairs with non-covered words*, or *PwNC.*

Overall, all datasets are fully covered by Wikipedia, while WordNet and Wiktionary have very limited coverage of domain-specific terminologies and named entities. Therefore, any knowledge base combination that includes *wk* (i.e., *wkwkt, wkwn, wkwnwkt*) has co-

Dataset	**MC 30**	**RG 65**	**Fin 153**	**Fin 200**	**Zie 25**	**Zie 30**	**MeSH 36**	**Ped 30p**	**Ped 30c**
wk	-	-	-	-	-	-	-	-	-
wn	-	-	2	4	21	24	13	15	15
wkt	-	-	2	4	18	23	18	15	15

Table 8.3. Coverage per KB per dataset: pairs with non-covered words (PwNC)

mplete coverage of word pairs such that *PwNC*=0.

PwNC provides an overview of knowledge base coverage in terms of *knowledge units (e.g., words, terms, concepts)* covered. However, as discussed before, different knowledge bases may focus on different types of lexical and semantic content, and cover different depth of content. Therefore, to gain a quantitative view of the content encoded for each knowledge unit, *the number of pairs that receive a zero relatedness score* is also analysed. This is referred to as *pairs with zero scores or PwZS* in the following. In theory, two terms may have a zero relatedness score if they are totally unrelated. However, in practice, absolute non-relatedness may be too difficult to define. In fact, none of the semantic relatedness datasets contain pairs that have zero scores, while some pairs may have very low scores.

For the proposed method, a zero score will be obtained if the feature representations of two terms do not share anything in common. Therefore, the number of zero scored pairs may vary depending on how a feature representation is created. Consider the different settings of feature representation proposed before, the number of zero scores may depend on: 1) the knowledge base(s) used; 2) whether feature integration of combination is used for feature mapping; 3) whether feature diversification or unification is used for representation. Empirically, experiments have shown that the number only depends on the underlying knowledge bases used, not feature mapping strategy or feature diversification or unification. Namely, the same number is found for *wk1* and *wk4*, or *wkwn4*$_i$ or *wkwn1*$_c$, as long as the compared settings use the same underlying knowledge bases. Therefore, Table 8.4 below shows the number of zero scored pairs found with settings of different knowledge bases, where *?* is a wildcard. Note that the zero scores are partially attributed by the non-covered word pairs (see numbers in Table 8.3).

As shown in Table 8.4, with the proposed method, all feature representation settings suffer from certain degrees of feature sparseness depending on the underlying knowledge bases used. In terms of each individual knowledge base, Wikipedia has the best coverage of content thanks to its encyclopaedic nature. As a result, a feature space created based on

Dataset	MC30	RG65	Fin153	Fin200	Zie25	Zie30	MeSH 36	Ped 30p	Ped30c
Single Knowledge Base (KB)									
wk?	-	1	1	11	-	-	1	1	1
wn?	7	23	28	49	24	30	19	21	21
wkt?	13	30	57	115	24	30	23	22	22
Multiple KBs									
wkwn?$_?$	-	1	-	5	-	-	1	1	1
wkwkt?$_?$	-	1	-	6	-	-	1	1	1
wkwnwkt?$_?$	-	1	-	4	-	-	1	1	1

Table 8.4. Pairs with zero scores (PwZS) per setting per dataset

Wikipedia can be high-dimensional, effectively increasing the possibility of shared features and therefore reducing the number of zero relatedness scores. In contrast, the feature space created based on WordNet or Wiktionary is very sparse, causing a large number of zero relatedness scores. This is because both knowledge bases serve the purpose of general dictionaries, and focus on lexical and semantic relations between concepts and word classes. The content defined for each knowledge unit is much less. For example, among all 447 distinctive words in all general domain datasets, only 69% have multiple synonyms in WordNet. Features such as *attributes* and '*see also*' are present for less than 20 words. The problem appears to be more acute for Wiktionary, which causes a larger number of zero scores than WordNet. This confirms the earlier findings of Navarro et al. (2009), that the amount of encoded information is largely imbalanced in Wiktionary; in particular, the synonym network can be very sparse for certain groups of words. For example, based on the MeSH36 dataset, among all terms covered by both WordNet and Wiktionary, all of them have synonyms defined in WordNet, while only 18% have synonyms in Wiktionary.

Due to the poor coverage of named entities in WordNet and Wiktionary and their sparse feature representation, only 1 pair of entity names in the Zie25 and Zie30 datasets received a non-zero score based on the two knowledge bases. Similarly, the coverage of domain specific terminologies in both resources is poor and the corresponding feature representation appears to be ineffective. As a result, semantic relatedness cannot be calculated for over 60% of the biomedical datasets. In contrast, Wikipedia outperforms by covering all knowledge units, while giving only one zero score possibly due to sparse feature representation.

8.5.4.2 Accuracy

Table 8.5 shows the accuracy (Spearman correlation) obtained with each of the 18 settings of feature representation introduced before. These figures are based on *ignoring*

PwNCs under each setting, i.e., word pairs that include non-covered words by a knowledge base are discounted. As an example, the score of 0.494 for the setting *wn1* on dataset Fin153 is obtained when the *2 PwNCs* are ignored according to Table 8.3, resulting in a total of 151 pairs accounted. However, in the remaining 151 pairs, 26 pairs (28 minus 2) have zero scores (i.e., *PwZS*), according to Table 8.4. Similarly, the highest score of 0.801 on MC30 given by *wn1* is obtained when 7 pairs have zero scores; and the highest score of 0.742 on RG65 given by *wn4* is obtained when 23 pairs have zero scores. In general, a large number of zero relatedness scores damages the resulting correlation scores; however exceptions are also noted and discussed below.

Table 8.5 is divided into three sections. The top section (**Single KB**) shows the correlation scores obtained when a single knowledge base is used. The *PwNC* figures associated with each knowledge base can be found in Table 8.3. In general, *wk* covers all pairs from all datasets; while varying numbers of *PwNCs* are ignored for *wn* and *wkt* depending on datasets. The middle section (**Mult. KBs**, 1 feature type) shows the results obtained when multiple knowledge bases are used with *feature unification*, i.e., 1 feature type from each knowledge base under *feature combination* or 1 joint feature type under *feature integration*. Due to the inclusion of *wk* in the underlying knowledge bases, all word pairs are covered, i.e., *PwNC* for any datasets is zero. For a fair comparison, these figures are compared against the figures obtained with *wk1* in the top section. The bottom section shows the results obtained when multiple knowledge bases are used with *feature diversification*, i.e., 4 feature types from each knowledge base or four joint feature types (**Mult. KBs**, 4 feature types). Similarly, all word pairs are covered and results are compared against the setting of *wk4*.

Single KB

Results obtained with single knowledge bases seem to suggest different levels of suitability of each knowledge base in terms of datasets. *WordNet* has led to better results on the MC30 and RG65 datasets, which are designed for measuring similarity rather than relatedness (see Table 8.2). On the contrary, correlations obtained on the relatedness datasets (Fin153 and Fin200) are much lower. Although it has led to reasonably good results on two of the three biomedical datasets (i.e., Ped30p, Ped30c), a large number of term pairs in these datasets have received zero relatedness scores (see Table 8.4). Further investigation shows that these zero-scored pairs have favourably biased the resulting correlation. For example, when zero scores are ignored, the *wn1* setting obtained correlations of

0.319 (-0.197), 0.506 (-0.032), 0.523 (-0.08) on the MeSH36, Ped30p and Ped30c datasets respectively. The *wn4* setting also suffers from drop in correlation when zero scored pairs are discounted. The correlation on the Zie25 and Zie30 datasets cannot be calculated for *wn1* or *wn4* settings since only 1 pair has non-zero relatedness score. Results obtained with *Wiktionary* (i.e., *wkt1* and *wkt4*) have shown similar patterns. Higher correlation scores are obtained on the general purpose similarity datasets (MC30, RG65), while scores are lower on the relatedness datasets (Fin153, Fin200). The results on the biomedical datasets are even poorer than the *wn*-based, and it is also not possible to calculate correlation on the Zie25 and Zie30 datasets due to poor coverage. In contrast, results obtained with *Wikipedia* are found to be more balanced among different datasets. It has led to good results on both named entity datasets and three biomedical datasets. However, results on the general domain similarity datasets are generally lower than those obtained with WordNet or Wiktionary under the same number of feature types, i.e., *wk1 v.s. wn1/wkt1, wk4 v.s. wn4/wkt4.*

Dataset	**MC30**	**RG65**	**Fin153**	**Fin200**	**Zie25**	**Zie30**	**MeSH 36**	**Ped30p**	**Ped30c**
	Single KB (the best corr. score is highlighted in **Bold**)								
wk1	0.761	0.604	**0.711**	**0.475**	**0.675**	**0.670**	**0.684**	0.454	0.483
wk4	0.693	0.614	0.645	0.464	0.532	0.545	0.682	**0.647**	**0.686**
wn1	**0.801**	0.723	0.494	0.353	-	-	0.516	0.538	0.603
wn4	0.788	**0.742**	0.461	0.332	-	-	0.486	0.656	0.719
wkt1	0.715	0.691	0.593	0.383	-	-	0.119	0.428	0.320
wkt4	0.697	0.657	0.461	0.333	-	-	0.170	0.397	0.358
	Mult. KBs, 1 feature type**,** scores **lower** than the Ref.set are in red								
Ref.set=wk1	0.761	0.604	0.711	0.475	0.675	0.670	0.684	0.454	0.483
wkwn1c	0.792	0.668	0.749	0.515	0.760	0.708	0.717	0.520	0.535
wkwn1i	0.773	0.647	0.727	0.510	0.665	0.652	0.744	0.460	0.492
wkwkt1c	0.810	0.769	0.648	0.509	0.695	0.582	0.658	0.498	0.548
wkwkt1i	0.827	0.711	0.725	0.496	0.675	0.670	0.729	0.456	0.485
wkwnwkt1c	0.910	0.797	0.699	0.521	0.730	0.582	0.603	0.514	0.525
wkwnwkt1i	0.824	0.713	0.734	0.505	0.672	0.665	0.744	0.476	0.506
	Mult. KBs, 4 feature types**,** scores **lower** than the Ref.set are in red								
Ref.set=wk4	0.693	0.614	0.645	0.464	0.532	0.545	0.682	0.647	0.686
wkwn4c	0.756	0.719	0.660	0.467	0.575	0.536	0.636	0.618	0.688
wkwn4i	0.781	0.646	0.668	0.498	0.603	0.546	0.729	0.620	0.705
wkwkt4c	0.787	0.749	0.655	0.464	0.516	0.518	0.685	0.649	0.734
wkwkt4i	0.747	0.678	0.646	0.500	0.559	0.549	0.767	0.636	0.691
wkwnwkt4c	0.801	0.737	0.671	0.477	0.496	0.509	0.682	0.648	0.706
wkwnwkt4i	0.785	0.673	0.660	0.517	0.596	0.545	0.733	0.623	0.710

Table 8.5. Results per setting per dataset, PwNC ignored.

Based on these observations and also the findings from the coverage analysis, this study concludes that indeed different knowledge bases have varying levels of suitability for different tasks in terms of domain and focus (relatedness/similarity). This is mainly due to the different focuses in knowledge units covered and different types of information encoded for these knowledge units by different knowledge bases. WordNet and Wiktionary focus on general domain common English words, and define a rich set of lexical and semantic relations. As a result, they are more suitable for measuring lexical semantic *similarity* between common English words. However, due to poor coverage of named entities and domain specific terminologies, they can be very ineffective for tasks concerning these areas. On the other hand, Wikipedia focuses on named entities and specialised concepts from both general and specific domains. As an encyclopaedia, it features long descriptive articles and loosely defined associations and links between articles. For these reasons, it is more suitable for measuring semantic *relatedness* rather than similarity. Its lack of word-level lexical and semantic information may reduce its effectiveness in tasks involving general English words. However, it is better suited for tasks involving named entities and domain specific terminologies.

Multiple KBs

Concerning ***general domain common English words*** (i.e., datasets MC30, RG65, Fin153 and Fin200), the results in both the ***middle*** and ***bottom*** sections in Table 8.5 show that using *multiple knowledge bases* (*wkwn*/*wkwkt/wkwnwkt*) generally improves over the method using the same method based on a *single knowledge base*. The only exception is *wkwkt1c* on Fin153 and *wkwnwkt1c* on Fin200, on which the results are lower than the scores obtained with only *wk1*. Considering the conclusion of knowledge base suitability drawn in the above section, this is possibly due to the nature of Wiktionary being less suitable for measuring relatedness, which the two datasets are designed for. Comparing results obtained with *all three knowledge bases* (*wkwnwkt*) against those obtained with *two knowledge bases* (*wkwn/wkwkt*), it is shown that adding more knowledge bases does not always lead to further improvement. Table 8.6 summarises changes of correlation to the best performing two-knowledge-bases setting when all three knowledge bases are used, based on the results for the general domain datasets in Table 8.5.

Although in more cases (10 instances), the combination of all three knowledge bases have led to further improvement, it also caused drop in accuracy. Overall for the general domain common word datasets, by combining multiple knowledge bases and using joint feature representations, the proposed method has gained more balanced and improved performances on both similarity and relatedness tasks when compared against single-KB based settings. This confirms that combining knowledge bases can effectively help over-

Dataset	**MC30**	**RG65**	**Fin153**	**Fin200**
	1 feature type			
Best two-KBs, feature combination (fc)	0.810, *wkwkt1c*	0.769, *wkwkt1c*	0.749, *wkwn1c*	0.515, *wkwn1c*
wkwnwkt1c	+0.1	+0.03	-0.05	+0.06
Best two-KBs, feature integration (fi)	0.827 *wkwkt1i*	0.711 *wkwkt1i*	0.727 *wkwn1i*	0.510 *wkwn1i*
wkwnwkt1i	-0.003	+0.002	+0.01	-0.005
	4 feature types			
Best two-KBs, (fc)	0.787, *wkwkt4c*	0.749, *wkwkt4c*	0.660, *wkwn4c*	0.467, *wkwn4c*
wkwnwkt4c	+0.054	-0.012	+0.011	+0.01
Best two-KBs, (fi)	0.781 *wkwn4i*	0.678 *wkwkt4i*	0.668 *wkwn4i*	0.500 *wkwkt4i*
wkwnwkt4i	+0.004	-0.005	-0.008	+0.017

Table 8.6. Comparison between results based on two KBs and three KBs.

come the limitations of each individual knowledge base, resulting in an improved feature representation that contributes to higher accuracy.

For the ***named entity datasets*** (Zie25, Zie30), due to the poor coverage of named entities in WordNet and Wiktionary, in many cases, combining *wn*, *wkt* or both with *wk* have damaged the accuracy of the method. Although improvement is noted in some cases, the overall pattern is inconsistent and the improvement may be opportunistic. For the ***biomedical datasets*** (MeSH36, Ped30p, Ped30c), in most cases a combination of multiple knowledge bases have contributed to improvement over a single knowledge base for measuring both relatedness (MeSH36) and similarity (Ped30p, Ped30c). However, in several cases, the poor coverage of biomedical knowledge in WordNet and Wiktionary has also damaged accuracies when they are combined with Wikipedia to create joint feature representations.

Based on these observations, this study concludes that under the proposed feature vector based method, combining multiple knowledge bases is most effective when individual knowledge bases have reasonable coverage of the knowledge concerned for the task. In general, it improves performance over the method based on a single knowledge base in terms of both coverage and accuracy. However, when an individual knowledge base for combination has very limited representation of knowledge required for a task, adding the knowledge base for combination may not always lead to improvement.

Feature combination/integration, diversification/unification

In terms of the ***feature mapping methods,*** *i.e., feature combination or integration*, figures in Table 8.5 do not show strong evidence that supports one over another. Both methods can enhance the feature representation, which then improves accuracy. To study the effect of *the number of feature types, i.e., feature diversification or unification*, the results in Table 8.5 are re-organised to present a different perspective in Table 8.7. Table 8.7 shows the number of feature types used with the best performing settings, for each combination of knowledge base and feature mapping methods. For example, *wk?* selects the best performing setting between *wk1* and *wk4*; while *wkwn?i* selects the best setting between *wkwn1i* and *wkwn4i*.

	Relatedness datasets					Similarity datasets			
Dataset	**Fin153**	**Fin200**	**Zie25**	**Zie30**	**MeSH 36**	**MC30**	**RG65**	**Ped30p**	**Ped30 c**
	where *?* =					where *?* =			
wk?	1	1	1	1	1	1	4	4	4
wn?	1	1	-	-	1	1	4	4	4
wkt?	1	1	-	-	4	1	1	1	4
wkwn?c	1	1	1	1	1	1	4	1	4
wkwn?i	1	1	1	1	1	4	1	4	4
wkwkt?c	4	1	1	1	4	1	1	4	4
wkwkt?i	1	4	1	1	4	1	1	4	4
wkwnwkt?c	1	1	1	1	4	1	1	4	4
wkwnwkt?i	1	4	1	1	1	1	1	4	4

Table 8.7. Number of feature types of the best performing settings

According to Table 8.7, in most cases, *feature unification* – i.e., a single type of feature from each knowledge base or a single joint type of features from multiple knowledge bases – has contributed to the best accuracies on the *relatedness* datasets. For *similarity* datasets, there is no consistent overall pattern. However, based on the biomedical datasets (Ped30p, Ped30c), *feature diversification* – i.e., four types of features from each knowledge base or four joint types of features from the combined knowledge bases – appears to be more effective.

These are possibly due to the linkage between the design of features as well as the definitions of semantic relatedness and similarity. Initially, the four different types of features are designed to represent different semantics, i.e., hypernymy relations, synonymy (and antonymy) relations, any other semantic associations (e.g., Wikipedia links, other semantic relations in WorldNet and Wiktionary), and implicit connections (words and gloss). According to the definition of semantic similarity in Section 8.2.1, semantic relatedness encompasses all kinds of connections between two concepts; while semantic similarity depends on the 'degree of synonymy', which is usually determined based on taxonomic relations (e.g., hypernymy, synonymy, and antonymy). Therefore, among the original four feature types, the hypernymy and synonymy features are more useful for quantifying similarity; while the other two are generally more useful for quantifying relatedness.

Since concepts only share features of the same feature types, preserving the different feature types essentially helps preserve the potential taxonomic links between concepts. On the other hand, collapsing feature types breaks the explicit semantic boundaries and links, creating simplistic bag-of-words representations that can be linked by any connections, which is biased towards relatedness.

For these reasons, settings with four feature types have achieved higher accuracies on the biomedical similarity datasets Ped30p and Ped30c, possibly because the hypernymy and synonymy features have been relatively denser than the other types of features and played a major role in quantifying similarity. In contrast, for measuring relatedness, in most cases a single bag-of-words representation has been more effective.

Context mapping

Previous analyses have focused on the effects of feature mapping methods, and the number of feature types in the joint feature representation. Another contributing factor to the accuracy of the proposed semantic relatedness method is the effectiveness of context mapping. A direct evaluation of the accuracy of context mapping can be a difficult task. Firstly, there is no gold standard for comparison. Secondly, different knowledge bases do not always cover matching knowledge units – in particular Wikipedia, which lacks coverage of general word senses. Thirdly, due to the lack of context, the interpretation of a pair of words can be subjective, which can affect the judgement of context mapping result. For example, human annotators may be more consistent with the selected sense for the word ‘bank’ when interpreting relatedness of ‘bank – finance’, but more inconsistent with the pair ‘bank – land’.

However, results discussed before can be considered as an indirect evaluation of context mapping. Intuitively, incorrectly or randomly mapped contexts from different knowledge bases should damage the feature representation; eventually harm the accuracy of the semantic relatedness method. Therefore, it can be generally concluded that the proposed context mapping method has been effective.

Nevertheless, the context mapping results on the MC30 dataset are manually inspected to obtain a closer look at this issue. The mapped Wiktionary entries are discounted, since an entry corresponds to a word class, and typically mixes different senses. The correctness of a mapped Wikipedia article – WordNet synset pair is judged subjectively by the author, based on the comparison of their descriptions.

Overall, among all 30 word pairs, 23 pairs are found to have suitable senses defined for each word in both WordNet and Wikipedia. Among these, the word senses in 21 pairs are deemed to be mapped correctly between the two knowledge bases. For an example, the word ‘crane’ has five senses in Wikipedia and 5 noun senses and 1 verb sense in WordNet. The five Wikipedia senses are mapped to three closest WordNet synsets (noun senses), as shown in Table 8.8. The other two WordNet synsets for ‘crane’ refers to a person and a constellation respectively.

Wikipedia Page Title	Closest WordNet Synset Mapped
Crane (bird)	Large long-necked wading bird of marshes and plains in many parts of the world
Crane (machine)	Lifts and moves heavy objects; lifting tackle is suspended from a pivoted boom that rotates around a vertical axis
Crane (railroad)	(same as above)
Cranes (band)	Stephen Crane , United States writer (1871-1900)
Crane (surname)	Stephen Crane , United States writer (1871-1900)

Table 8.8. Context mapping for the word ‘crane’

The remaining 9 pairs are shown in Table 8.9. In most cases, errors are due to the pair containing a word for which no matching senses are defined in the two knowledge bases.

Word 1	Word 2	Comment
journey	voyage	No suitable sense of ‘voyage’ in Wikipedia. All candidates are named entities.
magician	wizard	No suitable sense of ‘wizard’ in Wikipedia.
tool	implement	No suitable sense of ‘implement in Wikipedia.
monk	oracle	The ‘person’ sense of oracle in Wikipedia is incorrectly mapped to the ‘shrine’ sense of oracle in WordNet
monk	slave	No suitable sense of ‘slave’ in Wikipedia.
lad	wizard	No suitable sense of ‘wizard’ in Wikipedia.
chord	smile	No suitable sense of ‘chord’ in Wikipedia.
rooster	voyage	No suitable sense of ‘voyage’ in Wikipedia.
noon	string	The ‘rope’ sense of string in Wikipedia is incorrectly mapped to the ‘chain’ or ‘strand’ sense in WordNet

Table 8.9. Errors in the context mapping processing

8.5.4.3 Overall comparison with SoA

Based on the results presented in Table 8.5, the best figures obtained with combined knowledge bases are selected and compared against state-of-the-art. Figures for the *general domain common words datasets*, the *named entity datasets* and the *biomedical datasets* are presented separately in Table 8.10, Table 8.11, and Table 8.12. For each group of datasets, two groups of figures are compared: 1) *Best figures***:** the best figures obtained across different combinations of knowledge bases, feature mapping methods and feature type numbers; 2) *Best setting*: the figures from one best performing setting based on multiple knowledge bases.

Table 8.10 compares against state-of-the-art on the *general purpose common words datasets*. For each dataset, the best figures reported in Zesch and Gurevych (2010a) are used because on the one hand, these are the most recent figures available for the concerning datasets; on the other hand, the same versions of knowledge bases have been used, making a direct comparison possible. All correlation figures are calculated using the Spearman rank order correlation coefficient.

		MC30	**RG65**	**Fin153**	**Fin200**
Best figures	Corr. Setting	**0.910** *wkwnwkt1c*	0.797 *wkwnwkt1c*	**0.734** *wkwnwkt1i*	**0.521** *wkwnwkt1c*
Best setting (*wkwnwkt1c*)	Corr.	**0.910**	0.797	0.699	**0.521**
Lesk (1986)	Corr. KB	0.78 *wn*	0.72 *wn*	0.47 *wn*	0.33 *wn*
Zesch and Gurevych (2007, same as Zesch and Gurevych, 2008)	Corr. KB	0.77 *wn*	**0.82** *wn*	0.60 *wn*	0.51 *wn*
Gabrilovich and Markovitch (2007)	Corr. KB	0.67 *wk*	0.75 *wk*	0.69 *wk*	0.51 *wk*

Table 8.10. Comparison against state-of-the-art on the general domain common words datasets

Table 8.11 compares against state-of-the-art on the two named entity datasets. The work that published the best results on these datasets is Ziegler et al. (2006), who used Zie25 for parameter tuning and Zie30 for testing. The Pearson correlation function was used by Ziegler et al.

		Zie25	**Zie30**
Best figures	Corr. Setting	**0.760** *wkwn1c*	0.708 *wkwn1c*
Best setting (*wk1*)	Corr.	0.675	0.670
Ziegler et al. (2006)	Corr. KB	0.66 *Web+ODP*	**0.751** *Web+ODP*

Table 8.11. Comparison against state-of-the-art on the named entity datasets

Table 8.12 compares against state-of-the-art on the three biomedical datasets. Figures reported by the original dataset publishers are used, i.e., Petrakis et al. (2006) for MeSH36, and Pedersen et al. (2006) for Ped30p and Ped30c. Both studies used knowledge bases specifically tailored for biomedical data; and they both used the Pearson correlation function.

		MeSH36	**Ped30p**	**Ped30c**
Best figures	Corr.	**0.767**	0.649	0.734
	Setting	*wkwkt4i*	*wkwkt4c*	*wkwkt4c*
Best setting (*wkwkt4i*)	Corr.	**0.767**	0.636	0.691
Petrakis et al. (2006)	Corr.	0.71	-	-
	KB	*MeSH*	-	-
Pederson et al. (2006)	Corr.	-	**0.84**	**0.75**
	KB	-	*SNOMED-CT + a million-words corpus*	

Table 8.12. Comparison against state-of-the-art on the biomedical datasets

Tables 8.10 – 8.12 show very competitive results compared against state-of-the-art. For the *general domain common word datasets*, the proposed method has outperformed (both best figures and figures from the best setting) state-of-the-art on three datasets, i.e., MC30, Fin153, and Fin200. The scores on the RG65 dataset are also very close. For the *named entity datasets*, the best figures are obtained with the *wkwn1c* setting. However, as discussed before, WordNet's coverage of named entities is extremely limited, and the improvement gained by combining *wk* with *wn* may be opportunistic. For this reason, the best setting is chosen to be one that uses Wikipedia only, and in this case, *wk1*. Note that since different correlation functions are used, the results may not be directly comparable. However, studies by Zesch et al. (2010a) show that in many cases, the two correlation functions (i.e., Spearman and Pearson) can in fact result in comparable figures. Therefore, it is believed that the original figures by Ziegler et al. can still be useful reference. Generally, compared against Ziegler et al. (2006), the results are very competitive. Although the proposed method seems to underperform on the Zie30 dataset, the advantage of the proposed method is that it is completely unsupervised, while Ziegler's method requires parameter tuning using some training data (Zie25 in this case). For the *biomedical datasets*, the proposed method again obtained encouraging results. Both the work by Petrakis et al. (2006) and Pedersen et al. (2006) have used some domain specific knowledge bases; while the proposed method only uses general purpose knowledge bases. Therefore, the proposed method offers a competitive advantage over classic lexical semantic relatedness methods: by harnessing knowledge from general-purpose knowledge bases that may have limited domain coverage, it is possible to achieve results that are comparable to methods based on well-curated and specially tailored domain-specific knowledge resources. This is an encouraging finding. Although there are abundant resources in the biomedical domain for this type of tasks, such resources may be scarce in other domains and are expensive to build. However, the results suggest that the proposed method offers a more affordable approach that provides reasonable coverage and quality, even if individual general knowledge bases may be limited in themselves.

8.6 Conclusion

This chapter has addressed methods for measuring lexical semantic relatedness. These methods often play an important role and can be the enabling technology in many complex NLP tasks. It is a fundamental building block of the Named Entity Disambiguation approach to be discussed in the next chapter of this thesis. The study has two focuses: 1) a comprehensive review of the literature aimed at bridging related work from different areas and drawing conclusions concerning the research and applications of lexical semantic relatedness; 2) a novel method of lexical semantic relatedness based on harnessing different knowledge bases under a uniform framework.

The literature review

While methods of lexical semantic relatedness have been extensively studied in the past, a comprehensive review that connects work from different areas is lacking. This has caused expensive duplicate research effort that resulted in similar methods being proposed in different contexts, and difficulties in choosing the most appropriate methods in applications.

To address these issues, this chapter has presented a comprehensive literature review that discusses different background information resources used for lexical semantic relatedness, different categories of semantic relatedness methods and the rationales that connect them. The conclusion is that the choice of semantic relatedness methods often depends on a number of inter-related factors. Each category of methods has some advantages over others but equally suffers from certain limitations. These limitations are often associated with the underlying background information resources used by the methods. On the other hand, the choices of semantic relatedness methods can be limited by the availability of background information resources in relevant domains. Furthermore, different background information resources may cover different types of knowledge units, and focus on different types of information and content, all of which may affect their suitability for a task depending on its goal and the concerning domain. From the application point of view, with limited data it is unclear whether improvement in the accuracies of semantic relatedness methods always leads to positive and proportional improvement in the application built on top of it. For this reason, it is often necessary to balance the trade-off between the potential accuracy of semantic relatedness methods and their complexity when choosing for applications.

The new method

The literature review reveals that the vast majority of lexical semantic relatedness methods have employed a single source of background information, while it has been found that different background information resources can have different strengths and limitations. This motivates the hypothesis that different background information resources can complement each other, based on which a novel method is proposed to harness different knowledge bases in measuring lexical semantic relatedness.

The proposed method exploits three different knowledge bases – WordNet, Wikipedia, and Wiktionary, and is based on the idea of creating a joint feature representation of terms or concepts using the background information encoded in the three different resources, which ultimately improves the feature quality and outperforms a representation that is based on any one of the three resources. Evaluated on nine benchmarking datasets, the method has been shown to produce very competitive results and in many cases, outperform state-of-the-art. A key benefit of the proposed method is that by harnessing knowledge from general-purpose knowledge bases that may have limited coverage of domain specific knowledge, it is possible to achieve results that are comparable to methods based on well-curated and specially tailored domain-specific knowledge resources. This can be particularly useful when adapting the method to new domains where knowledge resources are lacking. The overall positive results have confirmed the validity of the proposed hypothesis.

In addition, several lessons are drawn regarding to combining multiple knowledge bases in measuring semantic relatedness. First, different knowledge bases are found to have varying levels of suitability for different tasks in terms of domain and focus (relatedness/similarity). WordNet and Wiktionary are generally more suitable for measuring semantic similarity between common English words. Their coverage of named entities and domain specific terms is extremely limited. Wikipedia, on the other hand, has good coverage of named entities as well as domain specific terminology. The lack of well-defined lexical and semantic relations between knowledge units makes Wikipedia more suitable for measuring general relatedness than similarity. Second, under the current method, combining knowledge bases is most effective when individual knowledge bases have reasonable coverage of the knowledge concerned for the task. When an individual knowledge base for combination has very limited representation of knowledge required for a task, adding the knowledge base for combination can adversely affect the performance. In terms of the different ways of feature representation, there is no evidence of differences given by different feature mapping methods, i.e., feature combination or integration. On the other hand, feature unification appears to be more effective for measuring relatedness while feature diversification could be more effective for measuring similarity in some cases. This could be because the former approach breaks semantic boundaries and links between different types of features, creating a feature representation that is consistent with the definition of relatedness; while the latter potentially preserves taxonomic links between concepts, thus promoting similarity. This issue will be further explored and additional evidence will be sought in future studies.

Future work

Several directions will be further explored in the future. First, a major limitation of the proposed method is feature sparseness. As discussed before, under the proposed method, all methods of feature representation suffer from certain degrees of feature sparseness, which caused zero relatedness scores. A solution to this problem can be including 'second order' features, i.e., features extracted for concepts or terms related to the current concept or term. For example, to cope with the sparse feature space given by the very short gloss in WordNet, one can expand the gloss by adding the glosses of any synsets in close relation to the current synset. For a Wikipedia article, features can be expanded by adding features of other articles linked to the current article, an approach that is used by Liu et al. (2010). Such approaches will be explored in the future. Second, although experiments have shown no empirical difference between the accuracy given by feature combination and integration, theoretically the two approaches can result in different feature representations that should lead to different relatedness scores for the same pair of words. It is unclear what the difference is and why it did not contribute to performance difference. Further work will investigate into this to uncover additional insights regarding combining knowledge bases in measuring semantic relatedness. Finally, concerning context mapping across knowledge bases, while the set overlap based method is simple and proved successful, its contribution with respect to the accuracy of the semantic relatedness method was not studied in details. For example, a random mapping method can be used as a baseline for reference. This will be explored in the future. Also, alternative mapping methods such as that proposed by Toral and Muñoz (2006) will be investigated and compared.

9 Named Entity Disambiguation[12]

PREFACE

This chapter discusses Named Entity Disambiguation, the task that resolves ambiguous name mentions of entities to unique objects depending on context. It is divided into six sections. Section 1 gives an overall introduction to this task; Section 2 discusses related work on NED; Section 3 discusses the hypothesis of ‘agreement maximisation’ based on lexical semantic relatedness, which leads to the NED method to be introduced in Section 4. Section 5 presents evaluation and discussion. Section 6 concludes this chapter.

[12] This work was carried out in collaboration with Dr. Anna Lisa Gentile.

9.1 Introduction

Natural language is well known for its polysemous nature, a property that describes the phenomenon that a word can have multiple meanings. This causes the problem of ambiguity, when the speaker or writer uses a polysemous word without explicitly stating the meaning of the word. The problem is extensible to named entities. As discussed earlier in Section 2.1 and 3.4, the linguistic realisation of an entity – the name or mention – can be ambiguous since it may be used to refer to multiple entities. For example, using the CoNLL2003 shared task dataset for example, it has been shown in Section 3.4 that 58.4% of annotated entities (which are 30.9% of all unique named entities) have used a name that is ambiguous.

While most of the time ambiguity is not a concern for humans, to enable understanding of natural language for machines, it is often necessary to process textual information in order to determine the underlying meaning. As discussed in the previous chapters, NER identifies mentions of named entities in texts and classify the mentions into semantic categories; while the task of resolving ambiguous names is called Named Entity Disambiguation (NED) and is a step further that should be applied to the output of NER. From a theoretical point of view, the two tasks are closely related, complementary and to some extent share common goals. From a practical point of view, it is often necessary to further process the output of NER by a disambiguation process in order for them to be useful by many applications based on named entities. For these reasons, this study views NED a further entity ‘recognition’ step and should also be addressed.

A large number of studies have addressed NED or related tasks, but are limited in different ways. For example, supervised learning methods (Han et al., 2004) require annotated training data, which are expensive to build and maintain. Unsupervised learning methods (Han et al., 2003; Mann and Yarowsky, 2003) on the other hand, are predominantly based on clustering techniques that only partially address NED with the goal of grouping name mentions that refer to the same entity without learning what the entity is. Knowledge based methods (Hassell et al., 2006; Peng et al., 2006; Bunescu and Pasca, 2006; Cucerzan, 2007) usually exploit the content and structures encoded in a knowledge base (e.g., an ontology). Most of these are tightly coupled with the underlying knowledge base and become constrained in terms of domain, types of entities or input format. As a result, many cannot be generalised for other tasks (Hassell et al., 2006; Peng et al., 2006).

This study proposes an approach to NED that is based on the principle of 'agreement maximisation'. Given a coherent text discourse that contains multiple ambiguous names of entities, the true referent entities of each name are usually semantically related. To exploit this, Cucerzan (2007) suggested that ambiguous names can be resolved by maximising the agreement among the data held for candidate entities in the same discourse. This study argues that this agreement can be measured by lexical semantic relatedness, and disambiguation can be achieved based on the idea of maximising the semantic relatedness between concerning named entities in a text discourse. Following this, the semantic relatedness measure introduced in Chapter 8 is adapted in this study for NED. It is used to measure semantic relatedness between underlying entities of ambiguous names extracted in a textual discourse. Next, a number of algorithms are proposed to select the true entity for each name, ensuring that the semantic relatedness among the resultant entities is maximised among all possible candidate entities for each ambiguous name. The method is evaluated using several standard benchmarking datasets. The results have shown that they achieved very competitive results and outperformed state-of-the-art on one dataset.

The remainder of this chapter is organised as the following: Section 9.2 discusses related work on NED. Section 9.3 discusses the hypothesis behind this work and Section 9.4 introduces the NED method and describes how the semantic relatedness measure is adapted for the task. Section 9.5 describes the experiments for evaluation and discusses results. Section 9.6 concludes this chapter.

9.2 Related Work

NED is a task closely related to Word Sense Disambiguation (WSD) and can be considered as a type of WSD where words to be disambiguated are entity names. It is also related to the field of data mining, where a similar task called reference resolution aims at resolving ambiguous references to existing records (often named entities) in a database. For the clarity of discussion, the task of NED is formalised following a similar fashion as WSD below.

Formally, the task of WSD aims to assign a word w occurring in a document d with its appropriate meaning or sense s. The sense s is selected from a predefined set of possibilities, usually known as **sense inventory**. Generally, a WSD algorithm takes as input a document $d = \{w_1, w_2, . . . , w_h\}$ and returns a one-to-one mapping from words to a list of senses X = $\{s_1, s_2, . . . ,s_h\}$ in which each element s_i is obtained by disambiguating the target word w_i. Following this, the task of NED can be formally defined as associating a reference of an entity – a name, mention, or surface form – *n* occurring in a document *d* with its appropriate identity (the referent entity, meaning or sense) *e*, which is the unique real-world object that *n* refers to. The referent *e* is selected from a predefined set of possibilities, which will be referred to as **entity inventory**. In addition, the disambiguation method may determine that there are no suitable candidates in the entity inventory for a name and assigns 'none' in such case.

While by this definition NED aims to associate name references to entities defined in an entity inventory, many studies however, do not use an entity inventory. They address a rather simplified goal of making clear if mentions of an NE across a given text collection refer to the same entity, then group the mentions without explicitly defining the entity. In a strict sense, this is defined as Named Entity Discrimination. Both address the problem of NE ambiguity resolution (Navigli, 2009).

The next few subsections discuss work related to NED. Work from the areas of WSD and reference resolution in data mining is also briefly introduced.

9.2.1 Word Sense Disambiguation

WSD is a fundamental task in NLP and has seen considerable efforts invested in this research over the last decades. The principles behind most WSD approaches are also the foundation of NED studies, and many WSD methods can be adapted to NED. A multitude of methods are available. Navigli (2009) surveyed the state-of-the-art, which generally, can be categorised into either learning based methods, or knowledge based methods. These are briefly discussed below to provide an overall background about disambiguation tasks.

Learning based methods employ supervised or unsupervised learning algorithms for disambiguation. Generally, WSD is viewed as a classification task, in which word senses are the classes, and each occurrence of a word is assigned one or more classes depending on its context. Following this, any supervised learning algorithms introduced previously in Section 2.3.2.1 can be applied to the problem. Typically, each occurrence of a word is represented by some features, usually based on its textual context. A classifier is trained for each sense of the word using sense-tagged training data. It can then be applied to new instances based on their feature representations generated in the same way. Supervised learning methods have been used by a large number of studies (Mooney, 1996; Pedersen, 1998; Agirre and Martinez, 2000; Keok and Ng, 2002) and continue to be the most frequently used technique for WSD. However, their main limitation is the cost of creating training data to build the learner.

Unsupervised learning methods for WSD specifically address NE discrimination, since the goal is to group word occurrences that reference the same sense rather than explicitly linking them to a sense inventory. These have primarily taken clustering algorithms. Generally, word occurrences can be represented by their features following the same way as in supervised methods. A clustering algorithm is applied to group these occurrences based on the similarity between their feature representations. A single cluster is considered to contain all occurrences of a word that refer to a unique sense. Examples of unsupervised WSD include Lin (1998a), Schütze (1998) and Lin and Pantel (2002). Although unsupervised WSD does not require training data, it only partially addresses NED as the sense of a cluster or group is unknown.

Knowledge based methods exploit knowledge resources to infer the senses of words in context. Many of such methods are based on the principle of comparing the context of a word against the lexical and semantic content encoded for the candidate senses of the word in a knowledge base (Lesk, 1986; Banerjee and Pederson, 2002). Also, semantic relatedness methods that use knowledge bases can be adapted for WSD. Following the principle of 'agreement maximisation', the senses for each ambiguous word in a text is selected to ensure that the sum of the semantic relatedness scores between any pairs of the resultant senses is maximised (Pedersen et al., 2005a). In WSD, the most frequently used knowledge bases is WordNet.

9.2.2 Named Entity Disambiguation

A milestone study that claims to address the problem of NED is carried out by Wacholder et al. (Wacholder et al., 1997). The work essentially addressed NER, but was presented from a perspective of disambiguation. In the task of identifying proper names, Wacholder et al. listed three types of ambiguities: structural ambiguity applies to proper names within prepositional and/or conjunctive phrases, making it difficult to identify the correct boundaries of names (e.g., 'The Museum *of* Modern Art *in* New York City'); semantic ambiguity that concerns correctly identifying the semantic types of named entities (e.g., 'Ford' can be a person, an organisation, or a place); and frequent use of common words in proper names (e.g., 'house/The White House', 'gate/The Gate'). To tackle these problems, they proposed to use 'disambiguation resources', the focus of which consists of a list of special name words that can indicate specific types of named entities, and a list of person first names. Examples of special name words can be personal titles (e.g., Mr, Mrs, Dr) that are often used with person names, or organisation words such as ‘Inc.’, or ‘Ltd’. These resources are used together with a large number of rules to identify boundaries of name mentions and classify them into specific entity types, essentially equivalent to the goals of NER.

In general, due to the strong connection with WSD, methods of NED can also be divided into learning based and knowledge based.

9.2.2.1 Learning based methods

Thanks to the genericity of both supervised and unsupervised learning algorithms, in most cases, the adaptation of learning based methods of WSD to NED is generally straightforward. Ambiguous names can be represented by features based on their context in the text. Then a supervised learner can be trained using sense-tagged training data; or an unsupervised algorithm can be applied to group mentions of the same entities based on the similarity between their feature representations. Some example studies based on learning based methods include Han et al. (2004), Mann and Yarowsky (2003), Pedersen et al. (2005b), and Chen and Martin (2007).

Learning based methods of NED suffer from the same limitations of learning based WSD. For supervised approaches, training data are expensive to obtain; for unsupervised approaches, the name occurrences are only grouped if they refer to the same entity, but the true referent entity is not identified.

9.2.2.2 Knowledge based methods

Knowledge based methods for NED are also based on the same principles of those for WSD. However, due to the lack of a universal entity knowledge base that is equivalent to WordNet or general dictionaries in WSD, these methods are often closely coupled with specific entity knowledge bases depending on the task of interest. They employ very specific features of the underlying knowledge bases, which can make the method non-generalisable across domains and tasks.

Most knowledge based methods are built on the principle of comparing the context of a name extracted from a text against the representations of its underlying referent entities created based on a knowledge base. In reference resolution, this type of method is called feature-based similarity techniques.

Hassell et al. (2006) proposed a method to disambiguate researcher names using various background knowledge encoded in a large scale, real-world ontology extracted from the DBLP bibliography website[13]. The ontology contains instances of researchers and defines different types of metadata for researchers (e.g., names, affiliation, research interest) as well as relationships between them (e.g., co-authors). To disambiguate a polysemous name found in a textual context such as a researcher's webpage, the basic principle is to look for the presence of such metadata in the context. Different types of metadata are given different weights. Thus each candidate entity that can be referenced by the same name receives a different score depending on the types and quantity of metadata found in the textual context, and the result is simply the entity that receives the maximum score.

Peng et al. (2006) proposed a two-level context based method to disambiguate location names. Given a large collection of documents that contain mentions of location names, a small proportion of name mentions are firstly disambiguated by a light-weight process that matches their 'local context' against an ontology of locations. Local context is defined as the preceding and following one word of a location name. For example, in 'Aberdeen, Scotland', 'Scotland' is a local context word of the location name 'Aberdeen'. In this light-weight process, the location name 'Aberdeen' is first searched within the ontology to identify possibly multiple nodes that correspond to different senses, e.g., 'Aberdeen, Maryland' and 'Aberdeen, Scotland'. Then the local contextual words are searched in the hypernyms leading from each node. If a match is found, the name is assigned the corresponding sense represented by the node. This light-weight process generates a sense-tagged corpus that is biased towards high-precision but low recall. The next step is to generate 'global context' of these sense-tagged names to be used for a further step of disambiguation. Global context of a sense consists of frequently co-located words of the sense in a collection of documents. For example, the location 'Washington D.C., USA' may often co-occur with 'President', 'the White House' in a document, but not necessarily in its local context. Such words are extracted for each sense of a location name using the sense-tagged corpus to create a 'profile' to represent the sense. These profiles are indexed by a search engine. To resolve other occurrences of ambiguous names, the local

[13] http://www.informatik.uni-trier.de/~ley/db/, last retrieved on 5 Mar 2012

context and the name are submitted as a query to retrieve and rank different profiles. The highest ranked profile and its corresponding sense is assigned to the name occurrence.

Both Hassell et al. and Peng et al. used domain specific knowledge bases and their methods are bound to the structure and content encoded within these resources. As a result, their methods cannot be easily generalised to other tasks. Recent research has recognised the significant potential of Wikipedia as an entity inventory or a general entity knowledge base due to its good coverage of named entities from various domains. This potential has been attested in a number NED studies that address different types of NEs.

Bunescu and Pasca (2006) proposed to disambiguate person names based on feature vector similarity and used Wikipedia as the name inventory. They firstly extracted all Wikipedia articles that are likely to describe named entities based on some heuristics. Words from these articles are extracted and stemmed, and each article is represented as a feature vector based on the word stems. Given a query that contains an ambiguous name, the name is represented as a feature vector based on its query context and the vector is compared against those of the Wikipedia articles that are potentially candidate entities. The one that gives the highest vector similarity score is chosen as the entity referenced by the name in the query. Further, they proposed to refine the vector representation by incorporating the correlation between certain words and category labels assigned to each article. The motivation is that the presence of certain words in the query often gives strong indicator of a category of the corresponding true entity to be assigned. For example, in ‘John Williams *conducted* a summer Star Wars *concert*…’ the words ‘conducted’ and ‘concert’ can indicate that the true entity for the name ‘John Williams’ should belong to the Wikipedia category ‘Composers’ or ‘Musicians’. The correlation strength between words and a total of 110 category labels applicable to person entities is empirically learnt using a collection of Wikipedia articles, and used to enhance the basic vector similarity model.

Cucerzan (2007) proposed a method that is also based on the ‘agreement maximisation’ hypothesis. They used Wikipedia as the name inventory, and proposed to represent a candidate entity using an extended vector consisting of two principle components, the context and category information, both extracted based on the entity’s Wikipedia article. Three types of category information are extracted. The first is the category labels assigned to an article; the second is extracted from Wikipedia “list” or “table” pages, which usually lists links to similar entities or concepts and has a title in the form of ‘List/Table of […]’, e.g., ‘List of animated TV series’. If an entity’s article is found on such a page, the title of this page is also considered a category of the entity. The third type of category information is extracted based on enumerations of entities. For example, the article 'Music of Scotland' contains a paragraph titled 'Classical Performers', which mentions and links to a list of entities such as 'Evelyn Glennie' and 'Murray McLachlan'. In this case, the entity names extracted from this paragraph are assigned the category label 'Music of Scotland'. The context information consists of appositives of titles of articles, and other mentions of entities in the page. For example, in the title of the article of the entity 'Texas (TV series)', the appositive phrase 'TV series' is considered a context for the entity. And mentions of other entities found on this article are considered as contexts for the entity. Empirically, some heuristics were introduced to reduce the number of contexts extracted to a manageable size.

The method then begins with pre-processing the entirety of Wikipedia to identify all entities, their surface names, and all contexts and categories available for these entities. Then each entity is represented by two vectors, one based on the context information (thus context vector) and the other based on the category information (thus category vector). The input to the disambiguation system is a document, which contains a number of ambiguous name mentions. The document is then represented as a vector of contexts – to be named context vector – such that the elements of the vector is based on the frequencies of a context (as extracted from Wikipedia above) found in the document. The disambiguation of the names in this document is based on the 'agreement maximisation' principle, which ensures that the true entities to be assigned to the ambiguous names in the document can satisfy two conditions: 1) the sum of the similarity between the document's context vector and the context vectors of each of the assigned entities is maximised among all possible entities for each ambiguous name; 2) and the sum of the similarity between any pairs of entities' category vectors are maximised among all possible entities for each name. In comparison, the proposed method in this study is also based on 'agreement maximisation' but a rather simplified version. While Cucerzan's agreement maximisation depends on two factors, which adds complexity to the computation, the proposed method in this study simplifies this by only considering agreement among candidate entities for ambiguous names, and ignoring the document context factor.

Han and Zhao (2009a, 2010) proposed a hybrid method that mixes the usage of knowledge bases and unsupervised learning. The task concerned NE discrimination, in which a set of webpages each describing a particular person entity is to be clustered such that each cluster groups webpages that potentially describe the same person entity. In principle, the method firstly represents each webpage as a weighted vector of Wikipedia concepts, and then clusters webpages based on the similarity between their concept vectors. To do so, they extracted n-grams from a webpage, and then filtered and matched n-grams to Wikipedia articles that represent specific concepts named by the n-gram. Each concept then receives a weight that represents the importance of the concept to the webpage to be disambiguated. This is assigned as the average semantic relatedness between the concept and any other concepts extracted from the webpage. Semantic relatedness between concepts is calculated based on the incoming and outgoing hyperlinks of the Wikipedia article for each concept. Finally, the webpages are clustered based on their vector representations.

9.2.2.3 Graph based methods

An increasing number of studies have proposed to address NED by exploiting both direct and indirect connections between candidate entities, usually represented as a graph. Such connections can be based on co-occurrence distributions, or lexical and semantic relations explicitly defined in a knowledge base. The boundaries between graph based methods and learning or knowledge based methods are not always clear-cut. When a graph is built based on connections defined in a knowledge base, it can be considered as a branch of knowledge based methods; on the other hand, in many cases, graph based algorithms are applied as a pre-process to derive entity similarities, which are then used as input to unsupervised learning methods for NE discrimination. However, graph based methods are characterised by the use of a graph based algorithm to incorporate various connections between candidate entities for disambiguation. It is also a major type of technique used for reference resolution in data mining.

Malin (2005) proposed a random graph walk based method for Named Entity Discrimination. The task concerned resolving ambiguous person names found on webpages. Given a set of webpages containing mentions of person names, some of which are ambiguous while some are not (known a-priori), Malin firstly created a graph that contains each unique unambiguous name and each *occurrence* of ambiguous names. For example, if ‘Alice’ is an ambiguous name in the collection and is found in three different webpages, the graph will have three nodes each corresponding to one occurrence. On the contrary, if ‘Bob’ is a name that consistently refers to a single and unique person entity, the graph will only have one node to represent the entity, despite the number of occurrences of the name in the collection. Then two names are connected by an edge if they co-occur in a webpage. The disambiguation method is based on the intuition that if several occurrences of the same name tend to be connected with the similar sets of other unambiguous names, they are likely to refer to the same entity. This was formalised by a random walk process starting from a node of an ambiguous name; the walk proceeds until another node of an ambiguous name is reached, or a maximum of 50 steps is reached. The resultant transitional probability distribution matrix stores the probability of reaching a node *a* from a node *b*, and vice versa. The two values are used to calculate a similarity score between *a* and *b* – only nodes of ambiguous names are considered. Then the similarity scores are used to cluster ambiguous names.

Minkov et al. (2006) also used a modified version of random graph walk to disambiguate names in emails. Email messages are firstly represented as graphs of different types of nodes (e.g., person, email address, date, word stem), connected by different types of edges (e.g., sent-to, sent-from, has-word). The random walk process is modified by also allowing the random walker to stay at the current node in addition to walking to other nodes in any step. Disambiguation is casted as a retrieval and ranking problem. Given a query containing an ambiguous person name, the query is represented as a vector over all nodes on the graph. It is then used to retrieve most relevant nodes from the graph while only nodes that denote person names are shown and ranked in the results.

Fernández et al. (2007) cast NED as a problem of ranking and proposed to use the personalised PageRank algorithm. Given a list of entity names extracted from news items and an entity inventory that indexes all entities referenced by these names, an initial graph is created and contains nodes denoting unique entities that can be refereneced by the names, and edges denoting co-occurrences of entities in any news items. The information of entity co-occurrence is based on a previously sense-tagged corpus. This graph enables the basic PageRank algorithm, which ranks entities on the graph based on the connections between them. An entity will have high rank if it tends to co-occur with entities that are also highly ranked. They introduced two domain specific factors to ‘personalise’ the ranking process: 1) the ‘semantic coherence’ principle which states that entities of a certain type usually occur in news of a certain category; also the occurrence of an entity in a text gives information about other entities; and 2) the ‘news trend’ principle which states that important events are typically described with several news items covering a certain period of time. As a result of 1), the category of the news item that a name reference belongs to will promote certain candidate entities of particular types, and the presence of certain entities (unambiguous) can have a similar effect; and as a result of 2), there can be a 'burst' in the mentions of particular entities within a given period of time and such entities are promoted. Empirically, such information is encoded by two vectors that 'steer' the mathematical computation of PageRank. In the end of the ranking algorithm, each candidate entity for each name reference will receive a rank. And the true reference entity is assigned to be the one that has the highest rank among all candidates for the name.

Nuray-Turan et al. (2007) proposed a graph-based method that resolves an ambiguous reference based on semantic relations between its underlying entities with other entities occurring in the same context using a pre-defined knowledge base, such as a database or ontology. The basic principle is that, if an ambiguous name reference *r* (e.g., 'J. Smith') is found in the context of an entity *e* (e.g., a scientific paper with the title 'The Paper'), then the true entity referenced by *r* is the one that has the strongest connection with *e* (e.g., the candidate 'John Smith, the researcher' will have stronger connection with *e* than 'Jane Smith, the singer'). Their method requires the availability of a knowledge base that contains candidate entities for *r* and the entity *e*, and various semantic relations defined among different entities. Then a graph is created where nodes represent entities, and edges represent semantic relations between entities. Different relations are given different weights, based on which edge weights are calculated. Then the connection strength between two entities is defined over the weights of edges that establish paths between them. Thus given the above scenario, among the set of entities referenced by *r*, the true entity is the one that has the maximum connection strength with the known entity *e*.

9.2.3 Related Evaluation Campaigns

A couple of related evaluation campaigns are currently organised on a regular basis, which has encouraged research in NED. The Web People Search (WePS, 2007) is a name of a series of workshops dedicated to resolving ambiguous person names in Web search results. The task carries the goal of NE discrimination, with the scenario of clustering webpages returned to a person name query, based on the identities described by these webpages. Initiated in 2007 and until 2010, WePS has now undergone three series and contributed to the publication of a plethora of methods for NE discrimination. Since the goal of the task is NE discrimination by clustering, nearly all of the proposed methods have used unsupervised clustering techniques. However, they differ largely in terms of the algorithms and features used. Therefore, WePS provides an excellent repository of literature for researchers that specialise in person name disambiguation.

The Knowledge Base Population (KBP) track in the Text Analysis Conference (TAC) is another regular event that promotes research related to NED. A subtask of KBP in TAC is entity linking, which aims at associating a name reference in a query to an existing entity defined in an entity knowledge base. The name reference is often ambiguous, and can refer to multiple entities in the knowledge base. Initiated in 2009, it has currently undergone three series. The task describes a typical NED setting, and has contributed to a diversity of methods, including supervised learning based (Fisher et al., 2009; McNamee, 2010; Pinto et al., 2009; Zhang et al., 2010), unsupervised learning based (Srinivasan et al., 2009), knowledge based (Honnibal and Dale, 2009; Lehmann et al., 2010), or a combination of different approaches (Agirre et al., 2009b).

9.3 Hypothesis

This section discusses the hypothesis of 'agreement maximisation', based on which the NED method is proposed:

> **H3.1 Resolving ambiguities: an ambiguous entity name can be resolved based on the semantic relatedness between its referent entities and other named entities it co-occurs with in its context, because contextually co-occurring named entities are semantically related.**

Given a coherent text discourse in which a number of (ambiguous) entity names are found, the true referent entities of each name are usually semantically related. For instance, the sentence 'President Bush attended the opening ceremony of the Olympic Game in Beijing' contains three ambiguous names: 'President Bush', 'Olympic Game', and 'Beijing'. The name reference 'President Bush' can refer to multiple person entities, e.g., the 43rd U.S. President George W. Bush, or the 41st U.S. President George H. W. Bush. 'Olympic Game' is a general term that describes a series of international sports events, and can be used to refer to any one event in a particular year. 'Beijing' is a name for cities, and according to Wikipedia, it is most commonly used to refer to the capital city of the country China (People's Republic of China), but is also a historic name for a number of different locations in different dynasties of the Chinese history. However, to most human readers, interpreting these ambiguous names is most likely very straightforward: in this particular context, 'President Bush' is the '43rd US President George W. Bush' (for brevity, this entity is denoted as e_1), 'Olympic Game' is the 2008 summer Olympic Games event (e_2), and 'Beijing' is China's capital city (e_3). The underlying logic is that these entities are the only combination in this sentence such that the semantic connection between them is at the strongest. Specifically, e_1 *attended* e_2; e_2 was *hosted in* e_3; and e_1 *visited* e_3 in 2008.

Arguably, the contextual words (e.g., ‘attended’, ‘in’) between these names may serve as important clues to a human reader; nevertheless the semantic connections between the underlying entities of the name references can still play a critical role. Imagine that all words are removed from the above sentence but the three name references. It is very likely that human readers can still interpret the names in the same way regardless of the context. This is because humans employ their background knowledge and make inferences based on the assumption that these names are related if they occur in the same context. This is in fact a common practice in human cognition. In practice, experienced human readers with proper background knowledge can quickly understand a text by skimming through it and spotting important phrases and terms, often representing concepts or named entities. In this case, their interpretation of such (potentially ambiguous) phrases and terms can be largely based on the similar assumption as in the example above.

There are two keys to exploiting this for automatic NED. First, there must be a way to establish and measure the semantic connections between entities, a task which matches well with the goal of lexical semantic relatedness measures. Second, there must be a way to select the most suitable entity for each name reference such that the overall semantic connection among them is at the strongest, or simply put, to achieve ‘agreement maximisation’.

In fact, the hypothesis of resolving ambiguities based on agreement maximisation and using measures of semantic relatedness is not new, but infrequently used. Pedersen et al. (2005a) and Cucerzan (2007) introduced methods based on similar assumptions. Pedersen et al. (2005a) proposed a method of WSD that assigns a sense to a target word by maximising the semantic relatedness between the target and its contextual words. Given an ambiguous word (target) and neighbour words from its context window, the semantic relatedness between the candidate senses of the target word and those of each neighbour word is computed. The chosen sense for the target word is the one that maximises the sum of the semantic relatedness between the sense and other neighbour words in the context. The hypothesis proposed in this study is different in the way that the agreement maximisation is based on semantic relatedness between candidate entities of each target ambiguous name. Context of name references are not considered.

Cucerzan (2007) applied a similar hypothesis to the task of NED. As introduced before (Section 9.2.2.2), their strategy for measuring agreement maximisation has two components, one considers the agreement between the context of name references and their candidate entities, the other considers the agreement between candidate entities. Compared to the method proposed in this work, this increases computational overheads.

9.4 NED based on Agreement Maximisation

This section introduces the proposed method of NED using semantic relatedness. In a complete workflow of NED, entity names should be firstly identified from a text document. This is done with NER, which has been the focus of discussion in the previous chapters. To avoid repetition, this study assumes that entity names are already extracted from a text and continues the NED workflow. Thus given a set of (ambiguous) entity names, the semantic relatedness measure introduced in Chapter 8 is adapted to measure semantic relatedness between underlying entities of the extracted names. Next, a number of algorithms are proposed to select the true entity for each name, while ensuring 'agreement maximisation' (Section 9.4.2).

9.4.1 Adaptation of the Semantic Relatedness Measure

As a brief re-cap, the semantic relatedness measure introduced in the previous chapter computes lexical semantic relatedness between two polysemous terms, based on the pairwise semantic relatedness computed for their underlying concepts using the background knowledge extracted from a combination of multiple knowledge bases. Here the method is adapted in three ways: 1) the underlying knowledge base is configured particularly for named entities; 2) context (articles describing candidate entities in this case) retrieval is adapted to ensure all candidate entities are extracted for a name reference; 3) the method is extended to compute relatedness for a set of name references instead of a pair.

9.4.1.1 The underlying knowledge base

The semantic relatedness measure originally uses a combination of three knowledge bases: Wikipedia, WordNet and Wiktionary. The evaluation has shown that, when applied to common words and domain specific terminologies, the three resources can complement each other and eventually improve the accuracy of prediction. However, when applied to named entities, Wikipedia in general is a good knowledge base but both WordNet and Wiktionary have poor coverage of named entities. Further, the quality of the content encoded for named entities in the two knowledge bases seems inadequate, such that when they are combined with Wikipedia, the resulting accuracy of the semantic relatedness measure dropped.

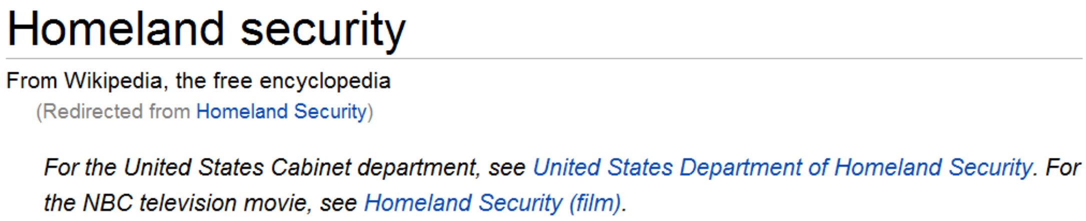
For this reason, both WordNet and Wiktionary are excluded from the semantic relatedness measure. Only Wikipedia is used as the background knowledge base for semantic relatedness and also the entity inventory for NED. Due to this change, context retrieval (previously in Section 8.4.2) and feature extraction and representation (previously in Section 8.4.3) are only based on Wikipedia; and there is no need of cross-resource context mapping and feature mapping (previously in Section 8.4.4).

9.4.1.2 Context retrieval

Initially, to retrieve the contexts for a term from Wikipedia, the term is searched in Wikipedia and if a single page is returned, it is chosen as the only context available for the term. When applied to NED, the term is equivalent to a name reference, and a context is considered to describe an entity. Although in some cases a single page is returned for a name reference, it is not always the case that the name reference is unambiguous. As discussed before, often, a single page that describes the most commonly used sense of a term is returned even if the term can be ambiguous. For example, the search for ‘London’ returns a single page describing the capital city of the U.K. However, ‘London’ can refer to other city entities, such as ‘London, Ontario, Canada’, and ‘London, Texas, U.S.’. This feature can restrict the number of candidate entities for a name reference and thus hinder the capability of the NED algorithm.

To rectify this issue, a search phrase is created for a name reference by appending the suffix ‘(disambiguation)’, which explicitly retrieves the disambiguation page defined for the name reference in Wikipedia. If no disambiguation page is available, the name reference is used instead. So to retrieve candidate entities for ‘London’, the phrase ‘London (disambiguation)’ is searched. If no results are returned, ‘London’ is attempted.

In addition, it has been noted that some non-disambiguation Wikipedia pages are named after ambiguous name references that in fact refer to multiple entities in Wikipedia. These articles start with a short paragraph of particular patterns to outline other senses. For example, there is no entry for the search phrase ‘Homeland Security (disambiguation)’ but one article matching the query ‘Homeland Security’, which begins with a paragraph shown in Figure 9.1. In these cases, the links and the corresponding Wikipedia articles (i.e., ‘United States Department of Homeland Security’ and ‘Homeland Security (film)’ are also extracted and used as candidate entities for the name reference.



Homeland security

From Wikipedia, the free encyclopedia

(Redirected from Homeland Security)

For the United States Cabinet department, see United States Department of Homeland Security. For the NBC television movie, see Homeland Security (film).

Figure 9.1. A non-disambiguation Wikipedia page named after an ambiguous name that refers to multiple Wikipedia articles.

Furthermore, several heuristics have been introduced previously in Chapter 8 to select suitable candidates from a disambiguation page. This is because that many links provided on a disambiguation page do not point to a concept or entity referenced by the search term, but some relevant concepts or entities; and therefore, the links must be processed to discard ‘noisy’ candidates. For this reason and also to ensure maximum recall, all links are extracted from list-like structures on the disambiguation page and are submitted to a filtering process based on the following heuristics:

1. If the link matches the search phrase as a whole or partially, it is selected. Thus the candidate links ‘London’, ‘Greater London’ and ‘London, California’ are all valid candidate entities for the name ‘London’.
2. If the link or search phrase matches (either completely or partially) the other as acronyms, it is selected. Thus ‘LSE’ will match candidates ‘London Stock Exchange’, and ‘London School of Economics and Political Science’.
3. For other links:
   - If the first word in the link is modified by the word ‘a’ or ‘an’, it is ignored. For example, the link ‘tiger’ in ‘A tiger, Panthera tigris …’ is discarded.
   - If the link is within the first three words of a list item, it is selected. This will select links that are referred by different names other than the search phrase, which can be often a more generic term used to refer to more specific senses in particular context. For example, ‘Volkswagen Type 2 (T1)’ will be selected from the list item ‘Volkswagen Type 2 (T1), generation T1 (Microbus, or Split-screen bus)’ found on the disambiguation page of the search phrase ‘VW bus’.

Names that do not match any Wikipedia page titles are searched in an inverted index of Wikipedia pages to obtain the first ten most relevant pages. The titles of these pages are validated using rules 1 and 2 above and only valid candidates are selected.

Next, a final check eliminates links that do not contain capitalised words, links that are anchor points to a particular position of a parent page (i.e., links that uses '#'), and links that point to categorisation or management purpose pages (e.g., 'category: buildings' or 'Lists of U.S. Presidents').

9.4.1.3 The input set of names

Originally the semantic relatedness measure takes the input of a pair of words or terms. In NED, the input will be a collection of name references extracted from a text discourse, e.g., a sentence, paragraph, or document. Following the majority of sense disambiguation studies, this work also adopts the 'one sense per discourse' rule that assumes multiple occurrences of a single name reference in a single text discourse will refer to the same entity. Therefore, the input set only contains unique name references.

Let N denote the set of input name references, $n_i \in N$ denote each unique name reference in the set, $E(n_i)$ denote the candidate entities for a name reference n_i, and $e_{i,j} \in E(n_i)$ denote each candidate entity for n_i, the semantic relatedness measure is applied to every pair of ($e_{i,j}$, $e_{i',j'}$), where $i \neq i'$. That is, the semantic relatedness between every candidate entity of a name and every candidate entity of another name is computed. The results are represented as a matrix, such as that shown in Figure 9.2.

9.4.2 Algorithms for Agreement Maximisation

In this step, one candidate entity is selected for each name reference from the resultant matrix as shown in Figure 9.2 based on the principle of agreement maximisation. Formally, let A be the resultant matrix containing pair-wise entity relatedness scores, a function f is applied to A to select a single element in E (the set of entities) for each element in N (the set of name references to be disambiguated). Three different functions are introduced below.

n_1=George Bush,	$e_{1,1}$: George W. Bush, 43rd U.S. President
	$e_{1,2}$: George H. W. Bush, 41st U.S. President
n_2=Olympic Games,	$e_{2,1}$: 2008 Olympic Games
	$e_{2,2}$: 2012 Olympic Games
n_3=Beijing,	$e_{3,1}$: Beijing, the capital city of P.R. China
	$e_{3,2}$: Beijing, a historic name of nowadays Taiyuan city in P.R. China

	$e_{1,1}$	$e_{1,2}$	$e_{2,1}$	$e_{2,2}$	$e_{3,1}$	$e_{3,2}$
$e_{1,1}$	-	-	0.5	0.2	0.65	0.0
$e_{1,2}$	-	-	0.15	0.07	0.6	0.0
$e_{2,1}$	0.5	0.15	-	-	0.7	0.01
$e_{2,2}$	0.2	0.07	-	-	0.25	0.0
$e_{3,1}$	0.65	0.6	0.7	0.25	-	-
$e_{3,2}$	0.0	0.0	0.01	0.0	-	-

Figure 9.2. Example results after computing semantic relatedness for the input set of name references

9.4.2.1 The combination method

With the combination method, the true referent entity for each name reference is selected as the one that has the maximum sum of relatedness scores with every other name references. Mathematically, this is formulated as:

$$\hat{e}_{i,j} = arg\,max_{e_{i,j}\in E(n_i)} \sum\nolimits_{n_{i'}\in N:i\neq i'} max_{e_{i',j'}\in E(n_{i'})}\{\,Sem\,Re\,l(\,e_{i,j}\,,e_{i',j'}\,)\} \qquad \textbf{Equation 9.1}$$

where *SemRel* is the semantic relatedness measure and the value of *SemRel*($e_{i,j}$, $e_{i',j'}$) corresponds to the value in the cell indexed by ($e_{i,j}$, $e_{i',j'}$) in *A*. This method is based on that used by Pedersen et al. (2005a). The difference is that while Pedersen et al. measure relatedness between a target ambiguous word and its surrounding contextual words, this method evaluates relatedness between target ambiguous name references.

Using the example matrix in Figure 9.2, the two candidate entities $e_{1,1}$ and $e_{1,2}$ for the name reference n_1 each receives a score of 1.15 (0.5+0.65) and 0.75 (0.15+0.6). As a result, the true referent entity for n_1 is assigned to be $e_{1,1}$, which has the highest sum of relatedness scores. Similarly, the true referent entities for n_2 and n_3 are found to be $e_{2,1}$ (1.2) and $e_{3,1}$ (1.35) respectively.

In case of a tie, the candidate entity that has produced most of the highest relatedness scores with other name references is chosen.

9.4.2.2 The voting method

With the voting method, the true referent entity for a name reference n_i is determined based on the number of votes for each of its candidate entity, casted by other name references. For each $n_{i'} \in N$ and that $n_i \neq n_{i'}$, $n_{i'}$ casts its vote to the candidate $e_{i,j} \in E(n_i)$ such that the value of $SemRel_{e_{i',j'} \in E(n_{i'})}(e_{i,j}, e_{i',j'})$ is maximised.

To illustrate, consider the case for n_1 in Figure 9.2, specifically the rows indexed by $e_{1,1}$ and $e_{1,2}$. The goal is to count how many other name references (i.e., among n_2 and n_3 in this case) have casted their votes to $e_{1,1}$ and $e_{1,2}$. The semantic relatedness between n_1 and n_2 are stored in the cells of [$e_{1,1}$, $e_{2,1}$], [$e_{1,1}$, $e_{2,2}$], [$e_{1,2}$, $e_{2,1}$] and [$e_{1,2}$, $e_{2,2}$], with values of 0.5, 0.2, 0.15 and 0.07 respectively. The maximum value is 0.5, between $e_{1,1}$ and $e_{2,1}$. Therefore, n_2 casts a vote to $e_{1,1}$. In the same manner, it can be worked out that n_3 also casts a vote to $e_{1,1}$. As a result, the first candidate of n_1 – i.e., $e_{1,1}$ – receives two votes, while the second candidate of n_1 receives no vote; and therefore, the true referent entity for n_1 is assigned to be $e_{1,1}$. Similarly, the true referent entities for n_2 and n_3 are found to be $e_{2,1}$ and $e_{3,1}$ respectively.

In case of a tie, the combination approach is applied and the voter that has the highest sum of relatedness scores wins. For example, suppose that the value in [$e_{1,2}$, $e_{2,1}$] is set to 0.6. As are result of this change, both $e_{1,1}$ and $e_{1,2}$ receives one vote. To break the tie, the combination approach is applied to calculate a sum of 1.15 for $e_{1,1}$ and a sum of 1.2 for $e_{1,2}$. As a result, $e_{1,2}$ wins and is selected for n_1.

9.4.2.3 The propagation method

The propagation method resolves ambiguities in an iterative manner, where in each turn two name references are disambiguated by selecting the two entities that produce the highest relatedness score in *A*. Specifically, it proceeds as the following:

1. Identify the highest relatedness score among all values in *A*. Select the two entities as the true referent for their corresponding name references;
2. Delete the rows and columns that correspond to other candidate entities of the two already-resolved name references;
3. Repeat steps 1 (ignoring the previously identified highest scores) and 2 until all name references are resolved.

Starting with the example in Figure 9.2, the highest relatedness score in the beginning state is 0.7, given by $e_{2,1}$ and $e_{3,1}$. They are selected to be the referent entities for the name references n_2 and n_3 respectively. Then, the rows $e_{2,2}$ and $e_{3,2}$ plus the columns $e_{2,2}$ and $e_{3,2}$ are deleted from the matrix, resulting in a new matrix as shown in Figure 9.3.

	$\boldsymbol{e_{1,1}}$	$\boldsymbol{e_{1,2}}$	$\boldsymbol{e_{2,1}}$	$\boldsymbol{e_{3,1}}$
$\boldsymbol{e_{1,1}}$	-	-	0.5	***0.65***
$\boldsymbol{e_{1,2}}$	-	-	0.15	0.6
$\boldsymbol{e_{2,1}}$	0.5	0.15	-	~~0.7~~
$\boldsymbol{e_{3,1}}$	***0.65***	0.6	~~0.7~~	-

Figure 9.3. Example intermediate matrix produced by the propagation method

In the next iteration, the previously identified highest values {0.7} are ignored, and the next highest value is identified. In this case, it is 0.65 produced by $e_{1,1}$ and $e_{3,1}$. Therefore, $e_{1,1}$ is chosen as the referent entity for the name reference n_1. When a tie is encountered, a score is randomly picked.

9.4.3 Exploiting Unambiguous Name References

It has been noted that often ambiguous name references co-exist with unambiguous names in the context and it has been a common practice to exploit the unambiguous name references in the disambiguation of the ambiguous ones (Malin, 2005; Fernández et al., 2007). To implement this feature, the values in the semantic relatedness matrix A are re-set to give a higher weight to the scores between two candidates if one of them belongs to an unambiguous name reference, i.e., there is only one candidate entity for the name. Let A^+ denote the modified matrix, practically, the values in A^+ is reset as:

$$A^{+}(e_{i,j},e_{i',j'})=\begin{cases} A(e_{i,j},e_{i',j'}), & if \ \ |E(n_i)|=1, \ \ or \ \ |E(n_{i'})|=1 \\ d\cdot A(e_{i,j},e_{i',j'}), & else \end{cases} \qquad \textbf{Equation 9.2}$$

where d is a damping factor $d\in(0, 1)$.

9.5 Evaluation and Discussion

This section describes the experiments for evaluating the proposed NED method and discusses results.

9.5.1 Method and Datasets

Methods for evaluating NED typically take the same forms as WSD. In WSD, the gold standard dataset consists of two parts, a sense tagged text collection and a sense inventory. Each piece of text includes a number of target words to be disambiguated. And the word is ‘sense-tagged’, meaning that it is assigned a reference pointing to one entry defined in the sense inventory. An automatic disambiguation method is evaluated based on precision, i.e., the fraction of the correctly annotated target words according to the gold standard. In NED, the datasets are simply replaced a text collection that contains sense-tagged entity names, each pointing to an entry in the entity inventory.

While there are well-maintained and widely used benchmarking datasets for WSD thanks to the availability of universal and generic sense inventories (e.g., WordNet), standard evaluation datasets for NED are relatively lacking. This is due to the lack of an equivalent comprehensive entity inventory, also that specific tasks often require domain specific entity inventories that are non-generalisable. Recent research in NED has recognised the potential of Wikipedia as a large scale knowledge base for named entities. As a result, a number of standard evaluation datasets have been constructed using Wikipedia as entity inventory and are becoming widely used. For this study, the datasets created by Cucerzan (2007) are used in evaluation for a number of reasons:

- They use Wikipedia as an entity inventory, which fits well with the semantic relatedness measure that also explores knowledge of named entities in Wikipedia;
- They cover a wide range of named entities, such as person names, locations, organisations, events etc.
- They are used as part of the NED datasets published by the TAC evaluation campaigns, and therefore, can be a good representation of the NED task.

Specifically, Cucerzan (2007) created two datasets that are detailed below.

The *NEWS dataset* is created based on 20 news stories. These were selected as the top two stories in the ten MSNBC news categories (Business, U.S. Politics, Entertainment, Health, Sports, Technology and Science, Travel, TV News, U.S. News and World News) published in January 2007. The dataset was pre-processed, only the list of entity names extracted from each story is provided. The number of entities in each story ranges from 10 to 50, and in total 756 name references are extracted. For each name reference, the most suitable Wikipedia article’s title is assigned to be the true referent entity. Some reference names do not have a suitable Wikipedia entry. They receive a ‘null’ annotation and are said to be ‘non-recallable’. A total of 127 names are non-recallable. An example of named entities in a news story is presented in Figure 9.4.

Name reference	Wikipedia entity ID (article title)
Timberlake	Justin Timberlake
Diaz	Cameron Diaz
N’ Sync	’N Sync
Justin Timberlake	Justin Timberlake
Cameron Diaz	Cameron Diaz
Star magazine	Star (magazine)
Star	Star (magazine)
Diaz	Cameron Diaz
Christmas	Christmas
Vail	Vail, Colorado
Colo.	Colorado
Timberlake	Justin Timberlake
Memphis	Memphis, Tennessee
N’ Sync	’N Sync
Saturday Night Live	Saturday Night Live
Diaz	Cameron Diaz
Timberlake	Justin Timberlake
Kids’ Choice Awards	Nickelodeon Kids’ Choice Awards
Timberlake	Justin Timberlake
Veronica Finn	Veronica Finn
Britney Spears	Britney Spears
Spears	Britney Spears
Innosense	Innosense (band)
Lou Pearlman	Lou Pearlman

Figure 9.4. Example document in the NEWS dataset

The *WIKI dataset* is based on 350 randomly selected Wikipedia articles that describe named entities. It contains a total of 5,812 name references. The referent entities for these name references are simply the hyperlinked articles created by the Wikipedia contributors. However, many of these (681) pointed to empty pages that are created to invite future contributions, or referenced out-dated articles and no longer available. The input documents are formatted in the same way as the NEWS dataset.

9.5.1.1 Re-creation of gold standards

Cucerzan (2007) created the gold standards based on an earlier version of Wikipedia which was no longer available at the time of this study. Due to the continued updating nature of Wikipedia, the content and structure of Wikipedia pages have substantially changed, invalidating a fair fraction of the data. As a result, the gold standard must be regenerated.

In this process, the same 2007 version of Wikipedia used in Chapter 8 was used to re-create the gold standards for the two datasets. Specifically, for the NEWS dataset, the gold standard answer for each name reference was searched in Wikipedia to retrieve a matching article. If no article can be found, the name reference is deleted. If the resulting article is unambiguous, it is retained in the gold standard; otherwise, the disambiguation page is manually analysed and a new answer is defined.

For the WIKI dataset, the 350 article titles were searched in Wikipedia and the gold standards were created in the same way based on the hyperlinks found in these articles. Articles containing only 1 name reference for disambiguation were discarded. Also, names that link to non-existent pages, disambiguation pages, specific locations within a target page, or management-purpose pages were discarded.

The new datasets are summarised in Table 9.1. The NEWS dataset is smaller than the original dataset by Cucerzan, while the WIKI dataset is larger. The numbers of ambiguous name references in the original datasets were unknown in Cucerzan (2007).

Data	# Docs	# Names	# Ambiguous names
NEWS	20	598	312
WIKI	332	6099	1917

Table 9.1. Experimental datasets statistics

The number of candidate entities for ambiguous names was also studied. The context retrieval method described in Section 9.4.1.2 was applied to each name reference and the number of candidate entities retrieved for was recorded. Among all ambiguous name references (names that have at least 2 candidate entities), the minimum number of candidates is 2, while the maximum is 156 (for ‘London’). Details are shown in Figure 9.5.

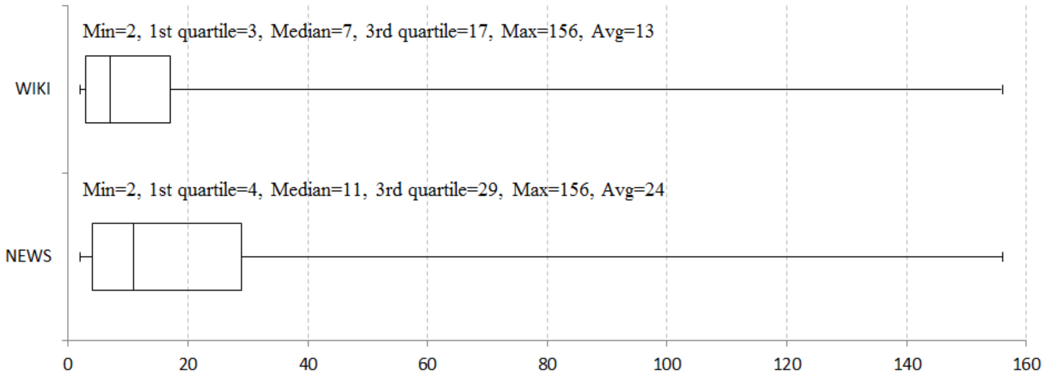


Figure 9.5. Statistics of candidate entities for ambiguous name references in each dataset

9.5.2 Algorithm Settings

As discussed before, the semantic relatedness measure is configured to use only Wikipedia as the background knowledge base. Two feature settings have been introduced in Chapter 8 (Section 8.5.3): *wk4*, four types of features extracted from Wikipedia; and *wk1*, a single type of feature that concatenates all features extracted from Wikipedia. It has been found that (Section 8.5.4.2) the single-feature setting favours relatedness, while the four-feature setting appear to favour measuring similarity on some datasets. Since it is hypothesized that disambiguation depends on semantic relatedness between name references, the *wk1* feature setting is used. Nevertheless, the *wk4* feature setting is also tested to empirically validate this hypothesis.

The agreement maximisation algorithms were applied to both the original semantic relatedness matrix *A*, and the matrix A^+ modified to give higher weights to relatedness scores with unambiguous name references. The damping factor *d* is arbitrarily set to 0.5.

The same baseline system in Cucerzan (2007) is used. The baseline simply picks the first entity listed on a disambiguation page for a name reference, or the first search result if the inverted Wikipedia index is used.

9.5.3 Results

The proposed method is tested on both datasets. Table 9.2 shows the accuracy obtained with the *wk1* feature setting when the entire datasets are considered (NEWS-all and WIKI-all), as well as when only the ambiguous name references are considered (NEWS-ambiguous and WIKI-ambiguous). *Comb, vote,* and *prop* denote respectively the *combination, voting* and *propagation* method applied to the original relatedness matrix *A*; while comb+, vote+ and prop+ denote respectively the combination, voting and propagation method applied to the modified relatedness matrix A^+.

	Method (semantic relatedness features = wk1)						
	Baseline	**Comb**	**Comb+**	**Vote**	**Vote+**	**Prop**	**Prop+**
NEWS-all	58.70	83.78	**88.80**	85.45	86.12	74.25	81.27
NEWS-ambiguous	20.83	68.91	78.53	72.11	73.40	50.64	64.10
WIKI-all	80.10	87.11	**88.60**	88.16	88.26	82.14	86.11
WIKI-ambiguous	39.54	59.15	63.91	62.49	62.81	43.51	56.13

Table 9.2. Disambiguation accuracy obtained when using the wk1 setting for the underlying semantic relatedness measure

Table 9.3 shows the accuracy obtained with the *wk4* feature setting for each method.

	Method (semantic relatedness features = wk4)						
	Baseline	**Comb**	**Comb+**	**Vote**	**Vote+**	**Prop**	**Prop+**
NEWS-all	58.70	78.93	84.95	76.76	77.76	68.23	73.91
NEWS-ambiguous	20.83	59.62	71.15	55.45	57.37	39.10	50.00
WIKI-all	80.10	84.14	85.70	85.44	85.47	80.16	82.52
WIKI-ambiguous	39.54	49.82	54.67	53.89	54.00	37.35	44.76

Table 9.3. Disambiguation accuracy obtained when using the wk4 setting for the underlying semantic relatedness measure

Furthermore, Table 9.4 shows the results reported by Cucerzan (2007) on the original datasets. Note that due to the change in the datasets and gold standard, the results are not directly comparable with those in Table 9.2 and Table 9.3.

Original datasets	**Baseline**	**Cucerzan (2007) best**
NEWS-all	51.7	91.4
NEWS-ambiguous	N/A	N/A
WIKI-all	86.2	88.3
WIKI-ambiguous	N/A	N/A

Table 9.4. The original results reported in Cucerzan (2007)

9.5.4 Discussion

Considering figures in Table 9.2, all the three variations of the proposed NED method largely outperformed the baseline on both datasets. The *combination* (*comb, comb*+) method appears to be the most effective, since it has obtained the majority of the highest figures. The *voting* (*vote*, *vote*+) method generally achieved comparable results, while the *propagation* (*prop*, *prop*+) method has produced the lowest accuracies among the three in all occasions. This is likely due to the fact that it picks the highest relatedness score at a time, which effectively only considers two name references while ignoring other co-occurring names in the context. To some extent, this violates the basic principle of disambiguation based on the context of a name reference.

Table 9.2 also shows that disambiguation can always benefit from unambiguous names in the context. When the semantic relatedness scores are discriminated such that scores with an unambiguous name's candidate entity are given higher weights, the accuracies of each algorithm have been further improved. This strategy appears to be particularly effective for the combination and propagation methods, where up to 13% of improvement was noted.

Consistent findings can be obtained from Table 9.3: generally, the combination method achieved the best results in most cases; the voting method achieved comparable results while the propagation method produced the lowest accuracy. Exploiting unambiguous names in disambiguation has always led to further improvement. Compared against Table 9.2, each method obtained lower accuracy when four types of features (*wk4*) were used for measuring semantic relatedness instead of one concatenated feature type (*wk1*). This could be a signal that suggests that disambiguation depends more on semantic relatedness rather than similarity, since previous experiments (Chapter 8) have proved that merging different feature types favours measuring general relatedness rather than similarity.

Since the datasets and gold standards are different from the originals in Cucerzan (2007), figures in Table 9.4 are not directly comparable. However, they can still serve as a general reference of state-of-the-art. The best performance obtained by this study is 88.80 on the NEWS dataset (NEWS-all) and 88.60 on the WIKI dataset (WIKI-all), which are generally comparable with Cucerzan's 91.4 on NEWS and 88.3 on WIKI. Nevertheless, as discussed before, this study adopts a simplified strategy to context modelling and agreement maximisation and therefore, offers a lighter approach than Cucerzan's method. Compared against standard WSD tasks using Figure 9.5, the task explored in this study is considered to be harder. In a typical WSD task, the average number of candidate senses per word is often around 7 (Agirre and Edmonds, 2007; Marine and Wu, 2007; Ng, 1997). In this study however, the disambiguation algorithm had to deal with much more candidate entities for each ambiguous names: an average of 13 for the WIKI dataset and 24 for the NEWS dataset. For the WIKI dataset, over 50% of ambiguous names have at least 7 candidates; while over 25% have 17 or more. For the NEWS dataset, 50% of the ambiguous names have 11 or more candidates while over 25% have 29 or more.

Error analyses have shown that many incorrect predictions were due to the use of co-references in documents. Often, some of them were disambiguated correctly while others were not. For example, in one document of the NEWS dataset, the 'Crimson Tide' is used to refer to the 'University of Alabama'. However, it was impossible for the candidate/context retrieval procedure of the proposed method to select 'University of Alabama' since it is not found on the pages corresponding to the search phrase. Also in another document, 'Richard Nixon' (the former U.S. President) was properly disambiguated while 'Nixon' was assigned the wrong referent due to a larger number of candidates for the latter name. It is found that in some cases, the heuristics for selecting candidate entities were also found to be over-simplified and can exclude legitimate candidates due to, e.g., spelling variations, and use of different names. For example, the candidate 'Galilee' was not selected for the search phrase 'Galil'; and the candidate 'American Democratic

Party' was not selected for 'The Democrats'. This may be rectified by employing string similarity in candidate selection. On the other hand, those errors could also be due to the over-simplified modelling of context in NED; since in such cases, document context may provide additional useful clues for disambiguation. Alternatively, a co-reference resolution pre-process based on document context can group name references that are likely to refer to the same entity in the same context, which NED can benefit from.

Overall, the experiment has shown the effectiveness of the proposed NED method and validated the hypothesis. The datasets are considered to be a good representation of the task, which can be harder than the typical WSD tasks due to the larger number of candidate entities to be disambiguated for a name reference. Furthermore, this experiment can also be considered as an *in-vivo* evaluation of the semantic relatedness measure. The results have shown from a different perspective that, the underlying semantic relatedness measure can accurately predict the semantic relatedness between named entities by exploiting diverse knowledge in Wikipedia, thus contributing to the NED task.

9.6 Conclusion

This chapter has addressed Named Entity Disambiguation, a task closely related to Named Entity Recognition. While NER extracts entity name mentions and classifies them into pre-defined semantic categories, the mentions can be ambiguous and can refer to multiple real world entities. NED aims to resolve the ambiguity and associate each name mention with one and unique entity that is pre-defined in a knowledge base. Theoretically, both address the similar goal of 'learning' of named entities, from different but complementary levels – one recognises the boundary and semantic type; the other recognises the instance. Practically, it is often necessary to further process the output of NER by a disambiguation process in order for them to be useful by many applications.

The literature on NED and the related task of WSD has been summarised, with their limitations discussed. A method of NED based on the hypothesis of 'agreement maximisation' is then proposed. In brief, the hypothesis states that entities that co-occur in a coherent text discourse are usually semantically related. Their ambiguous name references can be resolved based on the principle of maximising the semantic relatedness among their candidate entities. Following this, the semantic relatedness measure introduced in Chapter 8 is adapted for named entities. A number of methods are proposed to select the true entity for each name reference, while ensuring 'agreement maximisation'. The method is then evaluated on two standard benchmarking datasets for NED, and has achieved very competitive results.

The result of this study further confirmed the effectiveness of the semantic relatedness measure proposed in Chapter 8. It has also shown the strength of Wikipedia as a knowledge base in NED tasks. Among the three different realisations of ‘agreement maximisation’, the *combination* method is found to be most effective; the *voting* method achieves generally comparable performance, while the *propagation* method is the least effective. When unambiguous name references are given higher weights in computing semantic relatedness, the disambiguation algorithms can benefit further.

A couple of research directions will be explored in the future. First, new algorithms will be studied and proposed for ‘agreement maximisation’. For example, in the propagation method, instead of selecting the highest relatedness score each time, the algorithm can be revised to consider the highest sum of scores with unambiguous name references, and then incrementally disambiguate one ambiguous name at a time. This process effectively propagates the influence of unambiguous name references through the incremental disambiguation process. Second, recent research focus begins to shift to the use of linked data, which serve as a massive knowledge base of named entities, the size and scope of which are unprecedented. The Text Retrieval Conference (TREC) introduced an entity track in 2011 (TREC Entity Track, 2011), where the goal was to find entities related to a query from the linked data. This has sparked new interests of exploiting linked data in NE related tasks. In the future, research will be conducted to investigate the methods of mining and exploiting background knowledge of named entities from the data for NED.

Part V – Conclusion

10 Conclusion

PREFACE

This chapter concludes this thesis. Section 1 summarises the previous chapters with discussions of research questions, main contributions and research outcomes; Section 2 discusses limitations of this research and future research directions; Section 3 provides a closing statement to this thesis.

10.1 Summary

This research addressed three research questions related to Named Entity Recognition, an important task in Information Extraction and often the enabling technique to many text mining applications. The literature has been carefully reviewed and limitations of existing studies have been discussed. This research then proposed new methods to address these limitations. This section summarises this work.

10.1.1Problem Definition and Research Questions

NER is a fundamental task in IE and has been a focal research area over the decades. While traditionally NER recognises the *references* – mentions, names – of entities in unstructured texts and their *semantic categories*, this thesis combines NER with a further recognition step – recognising the *real entity* of a reference, a process that requires resolving ambiguities in name references.

The thesis started with an introduction to NER in Chapter 2, where the classic NER task is formalised and the background literature are introduced. The research questions concerning NER were described in Chapter 3, where three core research questions to be addressed by this thesis are discussed in details: training data annotation, gazetteer generation, and resolving ambiguities.

Methods of NER nowadays are largely based on learning based methods. Among these, the mainstream technique is supervised learning method, which requires an essential input – training data. Training data, in the context of NER, are documents with example named entity annotations created by humans. However, it is a widely recognised issue that training data are domain and task specific, resulting in limited transferability and portability of generated learning models. The standard practice for training data annotation is an expensive process, often involving significant investment in personnel, cost, and time. The process is inefficient, and the annotations can sometimes be ineffective. For this reason, this thesis viewed training data annotation an important and the first research question to be addressed: *an efficient and effective approach to training data annotation improves the extensibility and portability of supervised NER methods*.

Another crucial resource to NER is gazetteer, which is reference list of named entities or terms that provides background knowledge to an NER learner. They are found to be important in improving the accuracy of NER learner, particularly in specialised domains where the intrinsic complexity of the terminology and language makes the task more difficult. In the context of NER, gazetteers can be typed or untyped; the former refers to lists of named entities or terms whose semantic types are known a-priori and often relevant to the semantic types of named entities to be recognised in the task; the latter refers to general ways of grouping relevant terms such that the correspondence between the grouping and named entity types are to be learnt automatically by the learner. Gazetteers are not always available in any domains and manually compiling such resources is also an expensive process. As a result, this thesis viewed automatic gazetteer generation another major research question in NER: *addressing this research question will further improve the learning accuracy of NER.*

Due to the polysemous nature of natural language, words can be ambiguous since they can be used to refer to multiple senses. This is also common in named entities: the same name references can be used to refer to multiple entities. Using some standard NER datasets, it has been shown that ambiguous name references can be a prevailing issue in the output of NER. Traditionally, resolving ambiguous name references is dealt by the task of Named Entity Disambiguation. This thesis however, viewed NER and NED two complementary tasks that address the 'recognition' of named entities at different levels: while NER recognises name references and their semantic categories, NED recognises the real reference entity of a reference. Practically, many tasks built on top of NER output either require a compulsory disambiguation process or can benefit from such a process. Therefore, this thesis considered resolving ambiguities as another major research question concerning NER. *Addressing this research question takes further the 'recognition' process and enables the NER output to be ultimately useful to a wide range of tasks and applications*.

10.1.2 Research Main Contributions

This research carried out a series of studies each designed to address the three research questions outlined above. This has resulted in a number of contributions to the research of NER and related fields:

- An approach to training data annotation based on the hypothesis of 'annotator suitability' (H1, Chapter 5)

Starting with a named entity annotation task, the method splits the task into sub-tasks each concerning annotating one type of named entities. It then studies annotators' suitability for annotating each type, and distributes the workload among the most suitable annotators for each type. The motivation is that different annotators will have varying knowledge of different types of named entities, which determines the annotator's suitability for a sub-task. Unsuitable annotators are more likely to cause annotation discrepancies and should be isolated from a sub-task. Empirically, the suitability of annotators are analysed using a series of mini-annotation cycles in which the inter-annotator-agreement on each type of named entities is studied for each pair of annotators and the machine learning accuracy based on the annotations are analysed. Suitable annotators are selected based on these results following some heuristics. The generality of the proposed approach has also been discussed and in theory this approach can also be adapted to other annotation tasks.

- An approach to typed gazetteer expansion using Wikipedia (Chapter 6)

This approach has been proposed for expanding existing gazetteers of pre-defined types. Given an initial gazetteer, the method hypothesizes that Wikipedia contains knowledge that describe the existing gazetteers and also related named entities of the same type (H2.1). It explores various content and structural elements of Wikipedia to extract hypernyms of the seed named entities, and then harvests similar entities that share the same hypernyms with the seed named entities. Another contribution of this work is that it proves the capability of Wikipedia in domain-specific NLP applications.

- An approach to alternative untyped gazetteer generation based on topicality of words (Chapter 7)

Chapter 7 introduces an unsupervised method for generating alternative, untyped gazetteers for NER. It hypothesizes that topic-oriented words specific to a document are often indicative of named entities in the document (H2.2), and exploits this by building document-specific gazetteers based on the topicality of words measured specific to a document context. Compared to the previous work based on similar grounds, this study proposed different approaches to assessing topicality and gazetteer generation. It also contributed a comparative analysis with state-of-the-art to uncover the relation between word topicality and named entities.

- A comprehensive review of state-of-the-art in lexical semantic relatedness (Chapter 8)

In the exploration of semantic relatedness measures for NED, it was noticed that a plethora of methods has been introduced in different domains over the last decades. However, there is lack of a comprehensive review in this area and it was difficult to obtain a thorough understanding of the literature without substantial effort. This also caused very similar methods to be introduced in different contexts, which can cost expensive research effort. Therefore, a comprehensive literature review was carried out to consolidate research from different domains, and to summarise and connect different methods for measuring lexical semantic relatedness. Further lessons were also drawn on the research and application of lexical semantic relatedness measures. This is believed to be a valuable reference for both researchers and practitioners in this area.

- A novel lexical semantic relatedness measure that combines multiple knowledge resources (Chapter 8)

The literature review showed that lexical semantic relatedness measures typically employ a single source of background information, which has different focuses and can have different weakness and strengths. This motivates the idea of combining multiple resources in a single measure of semantic relatedness, based on the hypothetical complementary nature of such resources (H3.2). The proposed approach adopts a simple feature vector based relatedness method, where the central idea is to build a joint feature vector representation for concepts using knowledge extracted from different resources. This is the first semantic relatedness method that combines multiple resources in a single measure. It also contributed an empirical analysis of the strength and weakness of several knowledge bases (i.e., WordNet, Wikipedia, and Wiktionary) for different tasks (i.e., common words, named entities, and domain specific terminology).

- An approach to NED based on the hypothesis of ‘agreement maximisation’, measured in terms of lexical semantic relatedness (Chapter 9)

This approach hypothesizes that ambiguous entity names can be resolved by maximising the agreement among the data held for candidate entities in the same discourse (H3.1). The agreement between entities is measured using the lexical semantic relatedness method adapted from Chapter 8. Three strategies have been proposed for achieving the maximisation of the semantic relatedness among candidate entities; from which a disambiguation solution is derived.

10.1.3 Relating Research Outcomes to Research Hypothesis

Each study carried out in this research is based on a different research hypothesis. Each hypothesis has been discussed with literature support, as well as empirically justified with well-designed experiments.

- H1 Training data annotation: the hypothesis of ‘annotator suitability’ and suitability-based annotation selection is justified with a real-life named entity annotation use case study. The use case concerns a task of annotating three types of named entities in the archaeology domain, a good representation of the problem due to its increasing demand for document annotations and the scientific nature of the field. The use case has shown that, the suitability of annotators can be identified based on varying levels of inter-annotator-agreement on different entity types, as well as different machine learning accuracy obtained on the annotations created by different annotators. When unsuitable annotators are eliminated from a task, the quality of annotations is substantially improved compared with annotations created using the standard document annotation approach. This is demonstrated by higher machine learning accuracy using the data annotated by the proposed approach. Furthermore, distributing workload among suitable annotators also significantly improved the quantity of annotations by the same effort by up to five times.
- H2.1 Typed-oriented gazetteer expansion: the method for automatic expansion of existing gazetteers is empirically tested on three gazetteers of the archaeology domain. Using Wikipedia, the initial gazetteers were doubled (even tripled) in size in a single iteration. They were then applied to NER as an indirect evaluation, which showed that the expanded gazetteers further improved learning accuracy of NER between 1 and 3 points in F-measure. This confirms the quality of the expanded gazetteers as well as the validity of the hypothesis.
- H2.2 Alternative gazetteer generation: the method for generating untyped gazetteers was empirically evaluated on five datasets concerning three domains, and compared against state-of-the-art. Experiments showed that, the method contributed to consistent improvement in learning accuracy of NER (between 0.9 and 3.9 points in F-measure) and it is particularly effective in domain specific contexts (up to 3.9 points improvement in F-measure). The scope of the experiments also verified the generality of the hypothesis. Further analyses showed that compared against state-of the-art, the proposed method better captures the relation between topicality of words and named entities, which contributed to better performance.

- H3.1 Resolving ambiguities: the method for NED was empirically evaluated on part of the standard datasets used by major evaluation campaigns in the relevant field. The results showed that, the disambiguation outcome based on the 'agreement maximisation' consistently outperformed a baseline. The maximum improvements in accuracy on the two testing dataset were 30.1 (NEWS) and 8.5 (WIKI) respectively. The best performing strategy also obtained comparable results with state-of-the-art, while offering a more efficient methodology. It is believed that these are strong evidence supporting the hypothesis based on which the method is introduced.
- H3.2 Lexical semantic relatedness: the proposed lexical semantic relatedness measure combines knowledge from multiple resources, and was evaluated on an extensive set of datasets, specialising in common words, named entities and domain specific terminologies. The results showed that, by combining multiple knowledge bases, the accuracy of measuring semantic relatedness can be further improved. This has led to higher accuracy than state-of-the-art on the common words datasets, and very competitive results on the domain specific datasets. In the first case, Wikipedia was found to be less representative of the data while WordNet and Wiktionary were better knowledge bases; in the second case, Wikipedia has better coverage of domain specific terminologies while WordNet and Wiktionary have very limited coverage. Both cases confirmed the benefits of using multiple resources in lexical semantic relatedness. Although contradictory finding was obtained on the named entity dataset, it is believed to be caused by the extremely poor coverage of named entities in WordNet and Wiktionary, which were unsuitable for the task in the first place.

10.2 Future Work

A number of limitations have been identified and future work is proposed. These have been discussed in previous chapters, and are summarised in this section.

10.2.1On Individual Studies

Training data annotation (Chapter 5): The proposed approach to training data annotation will be improved in two ways. First, the study has focused on only inter-annotator-agreement while ignored intra-annotator-agreement, which can also affect the quality of annotations and the capability of annotators in a task. Future work aims to incorporate intra-annotator-agreement in the assessment of annotator's suitability. Second, currently the interpretation of analysis results is superficial and principles of annotator selection are informal. Future work aims to formalise the selection principles, possibly via enumerating 'suitability' such that it can be measured mathematically and interpreted more easily.

Typed gazetteer expansion (Chapter 6): The research will be further extended in three directions. First, the method will be modified to address scalability, such that it can cope with small input gazetteers, reducing the need for user provided input. Second, additional Wikipedia content and structures will be explored, such as list and table structures that often group similar entities. Third, the long term research objective is to explore a combination of online resources, particularly linked data that naturally integrates various resources in a uniform format.

Untyped alternative gazetteer generation (Chapter 7): Research in this topic will focus on developing new measures of topicality and new methods of deriving gazetteers based on topically. Firstly, since topic-oriented words can be related to keywords, keyword extraction methods may be adapted for this task. They will be empirically compared against IR-based relevance measures in this task, and novel measures of topicality will be investigated. Secondly, other methods of exploiting the non-linear distributional patterns of topic-oriented words over named entities will also be explored.

Lexical semantic relatedness (Chapter 8): Future work will be carried out from the following directions: first, the feature representation proposed under the current method was found to be sparse and can be ineffective in some cases. Therefore, methods of improving the feature representation will be sought. Second, the effects of the context matching and feature mapping strategies under the current method were unclear. Further studies will be carried out to uncover the differences between these strategies, as well as exploring new methods of combining multiple resources in measuring lexical semantic relatedness.

Named Entity Disambiguation (Chapter 9): The future work for NED will focus on developing new algorithms of 'agreement maximisation' based on lexical semantic relatedness. Furthermore, methods of mining and exploiting background knowledge of named entities from linked data for NED will be explored.

10.2.2 On the Overall Research Direction

It can be said at this point that one major focus of this research has been the exploration of background knowledge for NLP tasks. Chapters 6 and 7 have focused on automatic generation of gazetteers, a major source of background knowledge for NER; while chapters 8 and 9 have focused on exploiting background knowledge resources for lexical semantic relatedness and NED. This research has learnt valuable lessons on the use of various background knowledge resources – particularly Web-based resources – for NLP, and proved that they can be effectively used to support different tasks.

Recent development in Web 2.0 and the semantic Web has contributed to the emergence of a new source of background knowledge – the linked data. Linked data describes a recommended practice for exposing, sharing and connecting data using URIs and RDF (linkeddata.org). It originated from the DBpedia project (Bizer et al., 2009), the initial goal of which was to extract structured content from Wikipedia, and make them available as relational RDF triples on the Web such that semantic queries can be performed to gain deep access to Wikipedia resources. For example, it allows complex queries such as 'show me all Americans born after 1940'. The concept was to create a uniform method of linking, representing and querying information on the Web. This has been so well received that linked data has gained significant growth since its birth in 2009. As of March 2012, it has connected 123 datasets, with a total of over 19 billion triples (linkeddata.org). Linked data covers a wide spectrum of domains, such as newswire, biomedicine, music, geography etc.

It is believed that linked data offers great potential as a source of background knowledge for NLP tasks. Specifically:

- It has good domain representation and coverage. Based on Wikipedia, linked data already covers a large number of subjects. Furthermore, it extracts and provides deep content that were otherwise unavailable directly from Wikipedia articles. It also links a significant number of domain specific datasets. For example, the project Bio2RDF (Belleau et al., 2008) has interlinked major datasets in the life science domain and connected them to the linked data cloud, enabling access to domain-specific knowledge in the same open format. BBC Music (bbc.co.uk/music) is a dataset that publishes over 10 million facts (triples) in the music domain.
- It is growing very fast. According to the statistics from linkedata.org, the number of datasets has doubled every year since it was first created. It could be just a matter of time until many domains become well-represented.
- It provides a uniform method of accessing enormous amount of data. Unlike other extensively studied background knowledge resources, such as WordNet, Wikipedia, Wiktionary, or domain specific resources such as the UMLS, linked data provides a

uniform data representation format and access protocol to vast, diverse data sources. For application developers, this is a substantial benefit since they do not need to build implementations tailored to specific datasets due to the intrinsic differences in the underlying data structures of knowledge resources.

Such benefits have been widely recognised, as witnessed by increasing effort on publishing and connecting more data resources to linked data over the years. However, the number of research and applications on using linked data is still limited. Part of the reason could be related to the unbounded nature of linked data: on the one hand, the wild openness of linked data enables access to unlimited information from unlimited domains; on the other hand, a task is typically bounded by contexts. As a result, it is essential to identify only the relevant and limited part of linked data to support a task (Gangemi and Presutti, 2010). Research in this direction has only just taken off in the recent years.

Nevertheless, encouraging results have been obtained in both the application and research using linked data. Kobilarov et al. (2009) used DBpedia resource URI's as a controlled vocabulary to semantify and interlink BBC news articles to create a richly connected network of articles that improves browsing, navigation, and users' reading experiences with BBC websites. Becker and Bizer (2008) used geospatial data in DBpedia to build a recommender system for mobile phone platforms. The system requires a connection to DBpedia and GPS to work. The GPS constantly tracks the user's current geospatial coordinates, and the mobile phone is able to query DBpedia triple store for nearby place of interests by searching for geospatial coordinates within a radius of certain distance. Mulwad et al. (2010) proposed to use linked data to interpret tables and their semantics, and extract entities and relations based on the table structures. The Text Retrieval Conference (TREC) introduced an entity track in 2011 (TREC Entity Track, 2011), where the goal was to find entities related to a query from the linked data. This has encouraged a large number of studies on using linked data to support Information Retrieval or Extraction tasks in general.

The future research will continue in this direction, towards studying and exploiting linked data in Information Extraction tasks. Emphasis will be particularly placed on developing methods of effectively identifying and using relevant background knowledge from linked data to support different tasks.

10.3 Closing Statement

This research has investigated three inter-related research questions concerning NER: training data annotation, which enables supervised NER methods; automatic gazetteer generation, which enables improving the learning accuracy of NER; and resolving ambiguous named entity references, which adds a further step of 'recognition' and enables the NER output to be useful to a wide range of tasks. A number of studies have been carried out, contributing to a series of methods and findings that address each of the research questions. As in any research, there is always room for improvement. This chapter has summarised the work that has been done already; and listed the research outcomes, limitations and directions for future work. Outcomes of this work have shown that Web-based background knowledge resources are thriving nowadays and they can be exploited effectively to support various NLP tasks and therefore, further work in this direction – particular with linked data – will be pursued.

Bibliography

AbdelRahman, S., Elarnaoty, M., Magdy, M. and Fahmy, M. 2010. Integrated Machine Learning Techniques for Arabic Named Entity Recognition. *International Journal of Computer Science Issues*, Vol. 7(4), pp.27-36.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M. and Soroa, A. 2009a. A Study on Similarity and Relatedness using Distributional and WordNet-based Approaches. In *Proceedings of Human Language Technologies (HLT): The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado.

Agirre, E., Chang, A., Jurafsky, D., Manning, C., Spitkovsky, V. and Yeh, E. 2009b. Stanford-UBC at TAC-KBP. In *Proceedings of the TAC2009 Knowledge Base Population Track at the Text Analysis Conference*.

Agirre, E. and Edmonds, P., eds. 2007. *Word Sense Disambiguation: Algorithms and Applications*. 1st ed. Springer.

Agirre, E. and Martinez, D. 2000. Exploring Automatic Word Sense Disambiguation with Decision Lists and the Web. In *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken, Germany, pp.11-19.

AHDS. 1995. *Arts and Humanities Data Service*. [Online] Available at: http://www.ahds.ac.uk/ [Accessed 14 Dec 2011].

Ahmed, K., Gillam, L. and Tostevin, L. 1999. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In *Proceedings of the 8th Text Retrieval Conference (TREC-8).*

Al-Mubaid, H. and Nguyen, H. 2006. A Cluster-based Approach for Semantic Similarity in the Biomedical Domain. In *Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society*., pp.2713-17.

Altmann, U. 2005. Representation of Medical Informatics in the Wikipedia and its Perspectives. *Stud Health Technol Inform*, Vol. 116, pp.755-60.

Alvarez, M. and Lim, S. 2007. A Graph Modelling of Semantic Similarity between Words. In *Proceedings of the International Conference on Semantic Computing*. Irvine, CA, pp.355-62.

Ando, R. 2004. Semantic Lexicon Construction Learning from Unlabeled Data via Spectral Analysis. In *Proceedings of HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004).*

Ando, R. and Zhang, T. 2005. A High-Performance Semi-Supervised Learning Method for Text Chunking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*. Ann Arbor, MI, pp.1-9.

Appelt, D., Hobbs, R., Bear, J., Israel, D., Kaymeyama, M., Kehler, A., Martin, D., Myers, K. and Tyson, M. 1995. SRI International FASTUS system MUC-6 Test Results and Analysis. In *Proceedings of the Sixth Message Understanding Conference*., pp.237-48.

Archaeotools. 2007. *Archaeotools: Data mining, facetted classification and E-archaeology*. [Online] Available at: http://archaeologydataservice.ac.uk/research/archaeotools [Accessed 13 Mar 2012].

Arnold, A., Nallapati, R. and Cohen, W. 2008. Exploiting Feature Hierarchy for Transfer Learning in Named Entity. *Journal of Computational Linguistics*, (June), pp.245-53.

Babych, B. and Hartley, A. 2003. Improving Machine Translation Quality with Automatic named Entity Recognition. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*., pp.1-8.

Banerjee, S. and Pedersen, T. 2003. Extended Gloss Overlap as a Measure of Semantic Relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. Stockholm - Sweden, pp.805-10.

Banerjee, S. and Pederson, T. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*., pp.136-45.

Batet, M., Sánchez, D. and Valls, A. 2010. An Ontology-based Measure to Compute Semantic Similarity in Biomedicine. *Journal of Biomedical Informatics*, Vol. 44(1), pp.118-25.

bbc.co.uk/music. n.d. *BBC Music*. [Online] Available at: http://www.bbc.co.uk/music [Accessed 20 Mar 2012].

Becker, C. and Bizer, C. 2008. DBpedia Mobile: A Location-Aware Semantic Web Client. In *Proceedings of the Semantic Web Challenge*.

Belleau, F., Nolin, M., Tourigny, N., Rigault, P. and Morissette, J. 2008. Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge System. *Journal of Biomedical Informatics*, Vol. 41(5), pp.706-16.

Berners-Lee, T., James, H. and L., O. 2001. The Semantic Web. *Scientific American Magazine*, Vol. 284(5), pp.34-43.

Bhattacharya, A., Bhowmick, A. and Singh, A. 2010. Finding Top-k Similar Pairs of Objects Annotated with Terms from an Ontology. In *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management*. Heidelberg, Germany.

Bishop, C. and Lasserre, J. 2007. Generative or Discriminative? Getting the Best of Both Worlds. *BAYESIAN STATISTICS*, (8), pp.3-24.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellmann, S. 2009. DBpedia – a Crystallization Point for the Web of Data. *Journal of Web Semantics*, Vol. 7(3), pp.154-65.

Blanco, L., Bronzi, M., Crescenzi, V., Merialdo, P. and Papotti, P. 2010. Redundancy-driven Web data extraction and integration. In *proceedings of the 13th International Workshop on the Web and Databases (WebDB 2010)*. Indianapolis, Indiana.

Blitzer, J. 2008. *Domain Adaptaion of Natural Language Processing Systems*. PhD Thesis. University of Pennsylvania.

Blum, A. and Mitchell, T. 1998. Combining Labeled and Unlabelled Data with Co-training. In *Proceedings of the eleventh annual conference on Computational learning theory COLT98.*, pp.92-100.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, Vol. 32(1), pp.D267-70.

Bollegala, D., Matsuo, Y. and Ishizuka, M. 2007. An Integrated Approach to Measuring Semantic Similarity between Words using Information available on the Web. In *Proceedings the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*. Rochester, NY, pp.340-47.

Bookstein, A. and Swanson, D. 1974. Probabilistic Modelsfor Automatic Indexing. *Journal of the American Society for Information Science*, Vol. 25(5), p.312–318.

Boser, B., Guyon, M. and Vapnik, V. 1992. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT92)*. Pittsburgh, PA, p.144–152.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A. 2007. UniProtKB/ Swiss-Prot. *Methods in Molecular Biology*, Vol. 406, pp.89-112.

Brants, T. 2000. Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece, pp.1-5.

Brookes, B. 1968. The Measure of Information Retrieval Effectivenss Proposed by Swets. *Journal o fDocumentation*, Vol. 24(1), pp.41-54.

Budanitsky, A. and Hirst, G. 2006. Evaluating WordNet-based Measures of Semantic Distance. *Computational Linguistics*, Vol. 32(1), pp.13-47.

Bunescu, R. and Pasca, M. 2006. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL06)*. Trento, Italy, pp.9-16.

Burges, C. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, Vol. 2(2), p.121–167.

Byrne, K. 2007. Nested Named Entity Recognition in Historical Archive Text. In *Proceedings of the International Conference on Semantic Computing*. Irvine, CA, pp.589-96.

Cafarella, M., Downey, D., Soderland, S. and Etzioni, O. 2005. Knowitnow: Fast, Scalable Information Extraction from the Web. In *Proceedings of the Human Language Technology Conference (HLT-EMNLP05)*. Vancouver, B.C., Canada, pp.563-70.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. 2004. The Gene Ontology Annotation (GOA) Database: Sharing Knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, Vol. 32(1), pp.262-66.

Caputo, A., Basile, P. and Semera, G. 2009. Boosting a Semantic Search Engine by Named Entities. In *Proceedings of the 18th International Symposium on Foundations of Intelligent Systems (ISMIS)*. Prague, Czech Republic, pp.241-50.

Carletta, J. 1996. Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics*, Vol. 22(2), pp.249-54.

Carvalho, R., Chapman, S. and Ciravegna, F. 2008. Extracting Semantic Meaning from Photographic Annotations using a Hybrid Approach. In *International Workshop on "Metadata Mining for Image Understanding" (MMIU), the 3rd International Conference on Computer Vision Theory and Applications (VISAPP)*. Funchal, Madeira, Portugal.

Chen, H., Lin, M. and Wei, Y. 2006. Novel Association Measures using Web Search with Double Checking. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistic*. Sydney, Australia, pp.1009-16.

Chen, Y. and Martin, J. 2007. Towards Robust Unsupervised Personal Name Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech, p.190–198.

Cherry, J., Adler, C., Ball, C., Chervitz, S., Dwight, S., Hester, E., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. and Botstein, D. 1998. SGD: Saccharomyces Genome Database. *Nucleic Acids Research,* Vol. 26(1), pp.73-79.

Chieu, H. and Ng, H. 2003. Named Entity Recognition with a Maximum Entropy Approach. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL2003)*. Edmonton, Canada, pp.160-63.

Chinchor, N. 1998. Overview of MUC-7/MET-2. In *Proceedings of the 7th Message Understanding Conference*.

Chrupała, G. and Klakow, D. 2010. A Named Entity Labeler for German: Exploiting Wikipedia and Distributional Clusters. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta, Malta, pp.19-21.

Chung, E., Hwang, Y. and Jang, M. 2003. Korean Named Entity Recognition using HMM and CoTraining Model. In *Proceedings of the 6th International workshop on Information Retrieval with Asian Languages*., pp.161-67.

Church, K. and Gale, W. 1995a. Inverse Document Frequency (IDF): A Measure of Deviation from Poisson. In *Proceedings of the 3rd Workshop on Very Large Corpora*. Cambridge, Massachusetts, USA, pp.121-30.

Church, K. and Gale, W. 1995b. Poisson mixtures. *Natural Language Engineering*, Vol. 1(2), pp.163-90.

Cilibrasi, R. and Vitanyi, P. 2007. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19(3), pp.370-83.

Cimiano, P. 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. 1st ed. Springer.

Cimiano, P. and Völker, J. 2005. Towards Large-scale, Open-domain and Ontology-based Named Entity Classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*. Borovets, Bulgaria, pp.166-72.

Ciravegna, F., Chapman, S., Dingli, A. and Wilks, Y. 2004. Learning to Harvest Information for the Semantic Web. In *Proceedings of the 1st European Semantic Web Symposium*. Heraklion, Greece, pp.10-12.

Ciravegna, F., Lavelli, A. and Satta, G. 2000. Bringing Information Extraction out of the Labs: the Pinocchio Environment. In *Proceedings of the 14th European Conference on Artificial Intelligence*. Berlin, Germany.

Clauson, K., Polen, H., Boulos, M. and Dzenowagis, J. 2008. Scope, Completeness, and Accuracy of Drug Information in Wikipedia. *Ann Pharmacother*, Vol. 42(12), pp.1814-21.

Clifton, C., Cooley, R. and Rennie, J. 1999. TopCat: Data Mining for Topic Identification in a Text Corpus. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16(8), pp.949-64.

Collier, N., Nobata, C. and Tsujii, J. 2000. Extracting the Names of Genes and Gene Products with a Hidden Markov Model. In *Proceedings of the 18th conference on Computational linguistics*. Saarbrücken, Germany, pp.201-07.

Collins, A. and Loftus, E. 1975. A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, Vol. 82(6), pp.407-28.

Collins, M. and Singer, Y. 1999. Unsupervised Models for Named Entity Classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Hongkong, China, pp.100-10.

Colosimo, M., Morgan, A., Yeh, A., Colombe, J. and L., H. 2005. Data Preparation and Internannotator Agreement: BioCreAtIvE Task 1B. *BMC Bioinformatics*, Vol. 6(1), p.S12.

Couto, F., Silva, M. and Coutinho, P. 2005. Semantic Similarity over the Gene Ontology: Family Correlation and Selecting Disjunctive Ancestors. In *Proceedings of the ACM Conference in Information and Knowledge Management (CIKM)*. Bremen, Germany.

Cucchiarini, C. and Strik, H. 2003. Automatic Phonetic Transcription An Overview. In *Proceedings of the 15th International Congress of Phonetic Sciences*. Barcelona, Spain, pp.347-50.

Cucerzan, S. 2007. Large-scale Named Entity Disambiguation based on Wikipedia Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic, pp.708-16.

Curran, J. and Moens, M. 2002. Improvements in Automatic Thesaurus Extraction. In *Proceedings of the ACL2002 workshop on Unsupervised lexical acquisition*. Philadelphia, Pennsylvania, pp.59-66.

Da Silva, J., Kozareva, Z. and Noncheva, V. 2004. Extracting Named Entities: a Statistical Approach. In *Proceedings of the XIème Conférence sur le Traitement des Langues Naturelles (TALN)*. Avril, Fez, Marroco, pp.347-51.

Dagan, I. and Church, K. 1994. Termight: Identifying and Translating Technical Terminology. In *Proceedings of ACL Conference on Applied Natural Language Processing*., pp.34-40.

Dalvi, N., Kumar, R. and Soliman, M. 2011. Automatic Wrappers for Large Scale Web Extraction. *Very Large Databases Endowment*, Vol. 4(4), pp.219-30.

De Sitter, A. and Daelemans, W. 2003. Information Extraction via Double Classification. In *Proceedings of the ECML/PKDD 2003 Workshop on Adaptive Text Extraction and Mining (ATEM 2003)*.

Degtyarenko, K., Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. 2007. ChEBI: a Database and Ontology for Chemical Entities of Niological Interest. *Nucleic Acids Research*, Vol. 36, pp.344-50.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S. and Weischedel, R. 2004. The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal, pp.837-40.

Downey, D., Broadhead, M. and Etzioni, O. 2007. Locating Complex Named Entities in Web Text. In *Proceedings of the 2007 International Joint Conference on Artificial Intelligence*. Hyderabad, India.

Dredze, M., McNamee, P., Rao, D., Gerber, A. and Finin, T. 2010. Entity Disambiguation for Knowledge Base Population. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*. Beijing, China.

Duan, H. and Zheng, Y. 2011. A Study on Features of the CRFs-based ChineseNamed Entity Recognition. *International Journal of Advanced Intelligence*, Vol. 3(2), pp.287-94.

Ekbal, A. and Bandyopadhyay, S. 2010. Named Entity Recognition using Support Vector Machine: a Language Independent Approach. *International Journal of Electrical, Computer,and Systems Engineering*, Vol. 4(2).

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D. and Yates, A. 2004. Web-Scale Information Extraction in Knowitall:Preliminary Results. In *Proceedings of the 13th International World Wide Web Conference*. New York, NY, USA, pp.100-10.

Evans, R. 2003. A Framework for Named Entity Recognition in the Open Domain. In *Proceedings of Recent Advances in Natural Language Processing (RANLP2003)*. Borovetz, Bulgaria, pp.137-44.

Farkas, R., Szarvas, G. and Kocsor, A. 2006. Named Entity Recognition for Hungarian Using Various Machine Learning Algorithms. *Acta Cybernetica*, Vol. 17(3), pp.633-46.

Fellbaum, C. 1998. *WordNet an Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fernández, N., Blázquez, J., Sánchez, L. and Bernardi, A. 2007. IdentityRank: Named Entity Disambiguation in the Context of the NEWS Project. In *4th European Semantic Web Conference*. Innsbruck, Austria, pp.640-54.

Ferro, L., Mani, I., Sundheim, B. and Wilson, G. 2000. *TIDES Temporal Annotation Guidelines. Draft Version 1.0.*. MITRE Technical Report MTR 00W0000094.

Finkel, J. and Manning, C. 2009. Nested Named Entity Recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, Singapore, pp.141-50.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, Vol. 20(1), p.116–131.

Finn, A. 2006. *A Multi-Level Boundary Classification Approach to Information Extraction*. PhD Thesis. Dublin: University College Dublin.

Fisher, S., Dunlop, A., Roark, B., Chen, Y. and Burmeister, J. 2009. OHSU Summarization and Entity Linking Systems. In *Proceedings of the TAC2009 Knowledge Base Population Track at the Text Analysis Conference*.

Fort, K., Ehrmann, M. and Nazarenko, A. 2009. Towards a Methodology for Named Entities Annotation. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009*. Singapore, Singapore.

Freeman, M., Ramshaw, L., Bosche, E., Gabbard, R., Kratkiewicz, G., Ward, N. and Weischedel, R. 2011. Extreme Extraction – Machine reading in a week. In *Proceedings of* *the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, UK, pp.1437-46.

Freitag, D. 1998. Machine Learning for Information Extraction in Informal Domains. In *Ph.D. thesis*. Carnegie Mellon University.

Freitag, D. 2004. Trained Named Entity Recognition Using Distributional Clusters. In *Proceedings of the 2004 conference on Empirical Methods of Natural Language Processing*. Barcelona, Spain, pp.262-69.

Friedrich, C., Revillion, T., Hofmann, M. and Fluck, J. 2006. Biomedical and Chemical Named Entity Recognition with Conditional Random Fields: The Advantage of Dictionary Features. In *Proceedings of the Second International Symposium on Semantic Mining in Biomedicine*. Jena, Germany.

Gabrilovich, E. and Markovitch, S. 2006. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence*. Boston, MA, pp.1301-06.

Gabrilovich, E. and Markovitch, S. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. Hyderabad, India, pp.1606-11.

Gangemi, A. and Presutti, V. 2010. Towards a Pattern Science for the Semantic Web. *Journal of Semantic Web, Interoperability, Usability, Applicability*, Vol. 1(1-2).

Ganti, V., Konig, A. and Vernica, R. 2008. Entity Categorization Over Large Document Collections. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. Las Vegas, Nevada, pp.274-82.

Gentleman, R. 2005. *Visualizing and Distances Using GO*. [Online] Available at: http://bioconductor.org/packages/2.0/bioc/vignettes/GOstats/inst/doc/GOvis.pdf [Accessed 11 March 2011].

Giuliano, C. and Gliozzo, A. 2008. Instance-Based Ontology Population Exploiting Named-Entity Substitution. In *Proceedings of the 22nd International Conference on Computational Linguistics*. Manchester, UK, pp.265-72.

Giuliano, C., Lavelli, A. and Romano, L. 2006. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In *proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Trento, Italy.

Giuliano, C., Lavelli, A. and Romano, L. 2007. Relation Extraction and the Influence of Automatic Named Entity Recognition. *ACM Transactions on Speech and Language Processing (TSLP)*, Vol. 5(1), pp.1-26.

Gouws, S., Rooyen, G. and Engelbrecht, H. 2010. Measuring Conceptual Similarity by Spreading Activation over Wikipedia's Hyperlink Structure. In *Proceedings of the 2nd*

Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, in COLING'10. Beijing, China, pp.46-54.

Gracia, J. and Mena, E. 2008. Web-Based Measure of Semantic Relatedness. In *Proceeding of the 9th international conference on Web Information Systems Engineering*. Auckland, New Zealand, pp.136-50.

Grishman, R. and Sundheim, B. 1996. Message Understanding Conference - 6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*. Copenhagen, Denmark, pp.466-71.

Gu, B., Dahl, V. and Popowich, F. 2007. Recognizing Biomedical Named Entities in the Absence of Human Annotated Corpora. In *Proceedings of the international conference on Natural Language Processing and Knowledge Engineering*. Beijing, China, pp.74-81.

Gupta, S. and Bhattacharyya, P. 2010. Think Globally, Apply Locally: Using Distributional Characteristics for Hindi Named Entity Identification. In *Proceeedings of the ACL2010 Named Entities Workshop*. Uppsala, Sweden, pp.116-25.

Gurevych, I. 2005. Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP)*. Jeju Island, Korea, p.767–778.

Gut, U. and Bayerl, P. 2004. Measuring the Reliability of Manual Annotations of Speech Corpora. In *Proceedings of the Second International Conference on Speech Prosody (SP2004)*. Nara, Japan, pp.565-68.

Halavais, A. 2008. An Analysis of Topical Coverage of Wikipedia. *Journal of Computer-Mediated Communication*, Vol. 13(2), pp.429-40.

Han, H., Giles, L., Zha, H., Li, C. and Tsioutsiouliklis, K. 2004. Two Supervised Learning Approaches for Name Disambiguation in Author Citations. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. Tucson, AZ, USA, pp.296-305.

Han, H., Zha, H. and Giles, L. 2003. A Model-based K-means Algorithm for Name Disambiguation. In *In Second International Semantic Web Conference (ISWC03), Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*. Sanibel Island, Florida.

Han, X. and Zhao, J. 2009a. Named Entity Disambiguation by Leveraging Wikipedia Semantic Knowledge. In *Proceeding of the 18th ACM conference on Information and knowledge management*. Hongkong, China, pp.215-24.

Han, X. and Zhao, J. 2009b. NLPR_KBP in TAC 2009 KBP Track: A Two-StageMethod to Entity Linking. In *Proceedings of the TAC2009 Knowledge Base Population Track at the Text Analysis Conference*.

Han, X. and Zhao, J. 2010. Structural Semantic Relatedness: a Knowledge-based Method to Named Entity Disambiguation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pp.50-59.

Harman, D. and Liberman, M. 1993. *TIPSTER Volume 1*. Data. Philadelphia: Linguistic Data Consortium Linguistic Data Consortium.

Harrington, B. 2010. A Semantic Network Approach to Measuring Relatedness. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China, pp.356-64.

Harter, S. 1975. A Probabilistic Approach to Automatic Keyword Indexing: Part I. On the Distribution of Specialty Words in a Technical Literature. *Journal of the American Society for Information Science*, Vol. 26(4), pp.197-206.

Hassan, S. and Mihalcea, R. 2009. Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, Singapore, pp.1192-201.

Hassel, M. 2003. Exploitation of Named Entities in Automatic Text Summarization for Swedish. In *Proceedings of the 14th Nordic Conference on Computational Linguistics*. Reykjavik, Iceland, p.9.

Hassell, J., Aleman-meza, B. and Arpinar, B. 2006. Ontology-Driven Automatic Entity Disambiguation in Unstructured Text. *Information Systems Journal*, Vol. 4273(6), pp.44-57.

Haveliwala, T. 2002. Topic-sensitive PageRank. In *Proceedings of the 11th World Wide Web Conference (WWW)*. Honolulu, Hawaii, USA, pp.517-26.

Hearst, M. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*. Nantes, France, pp.539-45.

Hirst, M. 1998. Automated Discovery of WordNet Relations. In C. Fellbaum, ed. *WordNet: An Electronic Lexical Database*. MIT Press. pp.131-53.

Hirst, G. and St-Onge, D. 1998. Lexical Chains as Representation of Context for the Detection and Correction Malapropisms. In C. Fellbaum, ed. *WordNet: An Electronic Lexical Database and Some of Its Applications*. Cambridge, MA: MIT Press. p.305–332.

Holloway, T., Bozicevic, M. and Börner, K. 2007. Analyzing and Visualizing the Semantic Coverage of Wikipedia and its Authors. *Journal of Complexity, Special issue on Understanding Complex Systems*, Vol. 12(3), pp.30-40.

Honnibal, N. and Dale, R. 2009. DAMSEL: The DSTO/Macquarie System for Entity-Linking. In *Proceedings of the TAC2009 Knowledge Base Population Track at the Text Analysis Conference*.

Howlett, S. and Curran, J. 2008. Automatic Acquisition of Training Data for Statistical Parsers. In *Proceedings of the Australasian Language Technology Association Workshop*. Hobart, Australia, pp.37-45.

Hripcsak, G. and Rothschild, A. 2005. Agreement, the F-measure and Reliability in Information Retrieval. *Journal of the American Medical In-formatics Association*, Vol. 12(3).

Hripcsak, G. and Wilcox, A. 2002. Reference Standards, Judges, and Comparison Subjects: Roles for Experts in Evaluating System Performance. *Journal of the American Medical Informatics Association*, Vol. 9(1), pp.1-15.

Huang, J., Wang, C., Yang, C., Chiu, M. and Yee, G. 2005. Applying Word Sense Disambiguation to Question Answering Systemfor E-Learning. In *Proceedings of the 19th International Conference on Advanced Information Networking and Applications*. Taipei, Taiwan, pp.157-62.

Hughes, T. and Ramage, D. 2007. Lexical Semantic Relatedness with Random Graph Walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic, pp.581-89.

Hulth, A. 2003. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Sapporo, Japan, pp.216-23.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H. and Wilks, Y. 1998. University Of Sheffield: Description Of The Lasie-II System As Used for MUC-7. In *Proceedings of the 7th Message Understanding Conference*.

Ireson, N., Ciravegna, F., Califf, M., Freitag, D., Kushmerick, N. and Lavelli, A. 2005. Evaluating Machine Learning for Information Extraction. In *Proceedings of the 22nd international conference on Machine learning*. Bonn, Germany, pp.345-52.

Iria, J. 2009a. Automating Knowledge Capture in the Aerospace Domain. In *Proceedings of the 5th ACM International Conference on Knowledge Capture (K-CAP)*. Redondo Beach, California, US.

Iria, J. 2009b. *T-Rex*. [Online] Available at: http://t-rex.sourceforge.net/, [Accessed 9 Dec 2011].

Iria, J., Ireson, N. and Ciravegna, F. 2006. An Experimental Study on Boundary Classification Algorithms for Information Extraction using SVM. In *Proceedings of the EACL Workshop on Adaptive Text Extraction and Mining*. Trento, Italy.

Isozaki, H. and Kazawa, H. 2002. Efficient Support Vectors Classifiers for Named Entity Recognition. In *Proceedings of the 19th International Conference on Computational Linguistics*. Taipei, Taiwan, pp.1-7.

Jarmasz, M. and Szpakowicz, S. 2003. Roget’s Thesaurus and Semantic Similarity. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*. Borovets, Bulgaria, pp.111-20.

Jeffrey, S., Richards, J., Ciravegna, F., Chapman, S. and Zhang, Z. 2009. The Archaeotools project: Faceted Classification and Natural Language Processing in an Archaeological Context. *Special Theme Issues of the Philosophical Transactions of the Royal Society A,"Crossing Boundaries: Computational Science, E-Science and Global E-Infrastructures"*, Vol. 367(1897), pp.2507-19.

Jiang, J. and Conrath, D. 1997. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the International conference on research in Computational Linguistics*., pp.19-33.

Jiang, W., Guan, Y. and Wang, X. 2006. Improving Feature Extraction in Named Entity Recognition on Maximum Entropy Model. In *Proceedings of the 5th International Conference on Machine Learning and Cybernetics*. Dalian, China.

Jiang, J. and Zhai, C. 2007. Instance Weighting for Domain Adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, pp.264-71.

Joachims, T. 1999. Making large-Scale SVM Learning Practical. In B. Sch ölkopf, C. Burges and A. Smola, eds. *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

Ju, Z., Wang, J. and Zhu, F. 2011. Named Entity Recognition From Biomedical Text Using SVM. In *Proceedings of the 5th International Conference on Bioinformatics and Biomedical Engineering*. Wuhan, China, pp.1-4.

Kaufmann, M., Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H. and Wilks, Y. 1995. University of Sheffield: Description of the LaSIE System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference*.

Kazama, J., Makino, T. and Ohta, Y. 2002. Tuning Support Vector Machines for Biomedical Named Entity Recognition. In *Proceedings of the ACL2002 Workshop on Natural Language Processing in the Biomedical Domain*. Philadelphia, PA, USA, pp.1-8.

Kazama, J. and Torisawa, K. 2007a. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech, pp.698-707.

Kazama, J. and Torisawa, K. 2007b. A New Perceptron Algorithm for Sequence Labeling with Non-local Features. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and ComputationalNatural Language Learning*. Prague, Czech, pp.315-24.

Kazama, J. and Torisawa, K. 2008. Inducing Gazetteers for Named Entity Recognition by Large-Scale Clustering of Dependency Relations. In *Proceedings of the 46th Annual*

Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT). Columbus, Ohio, pp.407-15.

Keerthi, S. and Sundararajan, S. 2007. *CRF versus SVM-struct for Sequence Labeling*. Technical Report. Yahoo Research.

Keok, L. and Ng, H. 2002. An Empirical Evaluation of Knowledge Sources and Learning Algorithmsfor Word Sense Disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in NaturalLanguage Processing (EMNLP)*. Philadelphia, PA, USA, pp.41-48.

Kilgarrif, A. 2007. Googleology is Bad Science. *Computational Linguistics*, Vol. 33(1), pp.147-51.

Kim, J., Ohta, T. and Tsujii, J. 2008. Corpus Annotations for Mining Biomedical Events from Literature. *BMC Bioinformatics,* Vol. 9(10).

Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y. and Collier, N. 2004. Introduction to the Bio-Entity Recognition Task at JNLPBA. In *Proceedings of the COLING2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*. Geneva, Switzerland.

Kliegr, T., Chandramouli, K., Nemrava, J., Svatek, V. and Izquierdo, E. 2008. Combining Image Captions and Visual Analysis for Image Concept Classification. In *Proceedings of the 9th International Workshop on Multimedia Data Mining, at the 2008 International Conferenc on Knowledge Discovery and Data Mining*. Las Vegas, Nevada, USA.

Knopp, J. 2011. Extending a Multilingual Lexical Resource by Bootstrapping Named Entity Classification using Wikipedia's Category System. In *Proceedings of the 5th International Workshop On Cross Lingual Information Access, at IJCNLP2011*. Chiang Mai, Thailand, pp.35-43.

Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C. and Lee, R. 2009. Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web*. Heraklion, Greece.

Kohler, S., Schulz, M., Krawitz, P., Bauer, S., Dolken, S., Ott., C., Mundlos, C., Horn, C., Horn, D., Mundlos, S. and Robinson, P. 2009. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *The American Journal of Human Genetics*, Vol. 85(4), pp.457-64.

Kozareva, Z., Bonev, B. and Montoyo, A. 2005. Self-training and Co-training applied to Spanish Named Entity Recognition. In *Proceedings of the 4th Mexican International Conference on Artificial Intelligence*. Monterrey, Mexico, pp.770-79.

Kozima, H. and Furugori, T. 1993. Similarity between Words Computed by Spreading Activation an English Dictionary. In *Proceedings of the 6th Conference of the European*

Chapter of the Association for Computational Linguistics. Utrecht, The Netherlands, pp.232-39.

Krishnarao, A., Gahlot, H., Srinet, A. and Kushwaha, S. 2009. A Comparative Study of Named Entity Recognition for Hindi Using Sequential Learning Algorithms. In *IEEE International Advance Computing Conference*. Patiala, India, pp.1164 - 1169.

Kucera, H., Francis, N., Carroll, J. and Twaddell, W. 1967. *Computational Analysis of Present Day American English*. Brown University Press.

Lam, Y. 2010. *Comparing Na ïve Bayes Classifiers with Support Vector Machines for Predicting Protein Subcellular Location Using Text Features*. Master Thesis. Kingston, Ontario, Canada: Queen's University.

Laws, F. and Sch ütze, H. 2008. Stopping Criteria for Active Learning of Named Entity Recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics*. Manchester, UK, pp.465-72.

Leacock, C. and Chodorow, M. 1998. Combining Local Context and WordNet Similarity for Word Sense Identification. In C. Fellbaum, ed. *In WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MT Press. pp.265-83.

Lee, C., Hwang, Y. and Jang, M. 2007. Fine-grained Named Entity Recognition and Relation Extraction for Question Answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR'07*. Amsterdam, The Netherlands, pp.799-800.

Lee, K., Hwang, Y., Kim, S. and Rim, H. 2004. Biomedical Named Entity Recognition using Two-phase Model based on SVMs. *Journal of Biomedical Informatics*, Vol. 37(6), pp.436-47.

Lee, J., Kim, M. and Lee, Y. 1993. Information Retrieval based on Conceptual Distance in IS-A Hierarchies. *Journal of Documentation*, Vol. 49(2), p.188–207.

Lehmann, J., Monahan, S., Nezda, L., Jung, A. and Shi, Y. 2010. LCC Approaches to Knowledge Base Population at TAC 2010. In *Proceedings of the TAC2010 Knowledge Base Population Track at the Text Analysis Conference*.

Lei, Z. and Dai, Y. 2006. Assessing Protein Similarity with Gene Ontology and its Use in Sub-nuclear Localization Prediction. *Journal of BMC Bioinformatics*, Vol. 7(491).

Lesk, M. 1986. Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th annual international conference on Systems documentation*. Toronto, Canada, pp.24-26.

Liao, W. and Veeramachaneni, S. 2009. A Simple Semi-supervised Algorithm for Named Entity Recognition. In *Proceedings of the NAACL-HLT2009 Workshop on Semi-supervised Learning for Natural Language Processing*. Boulder, Colorado, USA, pp.58-65.

Li, Y., Bandar, Z. and McLean, D. 2003. An Approach for Measuring Semantic Similarity between Words using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15(4), pp.871-82.

Li, Y., Bontcheva, K. and Cunningham, H. 2005. SVM Based Learning System For Information Extraction. In J. Winkler, M. Niranjan and N. Lawerence, eds. *Deterministic and Statistical Methods in Machine Learning*. Springer Verlag. pp.319-39.

Li, Y., Lin, H. and Yang, Z. 2009. Incorporating Rich Background Knowledge for Gene Named Entity Classification and Recognition. *Journal of BMC Bioinformatics*, Vol. 10(223).

Lin, D. 1998a. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*. Montreal, Quebec, Canada, pp.768-74.

Lin, D. 1998b. An Information-theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine learning (ICML'98)*. Madison, Wisconsin, USA, pp.296-304.

linkeddata.org. n.d. *Linked Data - Connect Distributed Data across the Web*. [Online] Available at: http://linkeddata.org/ [Accessed 20 Mar 2012].

Lin, D. and Pantel, P. 2002. Discovering Word Senses from Text. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alta, Canada, pp.613-19.

Lin, Y., Tsai, T., Chou, W., Wu, K., Sung, T. and Hs, W. 2004. A Maximum Entropy Approach to Biomedical Named Entity Recognition. In *Proceedings of BIOKDD'04, 4th Workshop on data mining in bioinformatics (with SIGKDD'04)*. Seattle, WA, USA, pp.56-61.

Li, D., Savova, G. and Kipper-Schuler, K. 2008. Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts. In *Proceedings of the ACL2008 Workshop on Current Trends in Biomedical Natural Language Processing*. Columbus, OH, USA, pp.94-95.

Liu, H. and Chen, Y. 2010. Computing Semantic Relatedness between Named Entities using Wikipedia. In *International Conference on Artificial Intelligence and Computational Intelligence*., pp.388-92.

Liu, X., Zhou, Y. and Zheng, R. 2007. Measuring Semantic Similarity in WordNet. In *Proceedings of the 6th International Conference on Machine Learning and Cybernetics*. Hongkong, China, pp.3431-35.

Li, B., Wang, J., Feltus, F., Zhou, J. and Luo, F. 2010. Effectively Integrating Information Content and Structural Relationship to Improve the GO-based Similarity Measure between Proteins. In *Proceedings of the International Conference on Bioinformatics & Computational Biology (BioCOMP)*., pp.166-72.

Lovász, L. 1993. Random Walks on Graphs: A Survey. *Combinatorics Paul Erdos is Eighty*, Vol. 2(1), pp.1-46.

Ma, X. 2009. *Improving Named Entity Recognition with Co-training and Unlabeled Bilingual Data*. PhD Thesis. University of Pennsylvania.

Magnini, B., Negri, M., Prevete, R. and Tanev, H. 2002. A Wordnet-based Approach to Named Entity Recognition. In *International Conference On Computational Linguistics COLING2002 on SEMANET: building and using semantic networks*. Taipei, Taiwan, pp.1-7.

Maguitman, A., Menczer, F., Roinestad, H. and Vespignan, A. 2005. Algorithmic Detection of Semantic Similarity. In *Proceedings of the 14th International World Wide Web Conference*. Chiba, Japan, pp.107-16.

Malin, B. 2005. Unsupervised Name Disambiguation via Social Network Similarity. In *Workshop on Link Analysis, Counterterrorism, and Security at the SIAM International Conference on Data Mining.* Newport Beach, California, USA, pp.93-102.

Mann, G. and Yarowsky, D. 2003. Unsupervised Personal Name Disambiguation. In *Proceedings of the 7th conference on Natural language learning at HLT-NAACL2003 (CoNLL'03)*. Edmonton, Canada, pp.33-40.

Marcus, M., Santorini, B. and Marcinkiewicz, M. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Journal of Computational Linguistics - Special issue on using large corpora*, Vol. 19(2), p.313–330.

Marine, C. and Wu, D. 2007. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation*. Skövde, Sweden, pp.43-52.

Matsuo, Y., Sakaki., T., Uchiyama, K. and Ishizuka, M. 2006. Graph-based Word Clustering using a Web Search Engine. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Sydney, Australia, pp.542-50.

Mayfield, J., McNamee, P. and Piatko, C. 2003. Named Entity Recognition using Hundreds of Thousands of Features. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. Edmonton, Canada, pp.184-87.

McNamee, P. 2010. HLTCOE Efforts in Entity Linking at TAC KBP 2010. In *Proceedings of the TAC2010 Knowledge Base Population Track at the Text Analysis Conference*.

Medical Subject Headings. 1999. *Medical Subject Headings*. [Online] Available at: http://www.nlm.nih.gov/mesh/ [Accessed 15 Mar 2012].

Mei, Q., Fang, H. and Zha, C. 2007. A Study of Poisson Query Generation Model forInformation Retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'07)*. Amsterdam, The Netherlands, pp.319-26.

Meyer, C. and Gurevych, I. 2010. How Web Communities Analyse Human Language: Word Senses in Wiktionary. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*. Raleigh, NC, USA, pp.1-8.

Mihalcea, R. and Moldovan, D. 1999. A Method for Word Sense Disambiguation of Unrestricted Text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. College Park, Maryland, USA, pp.152-58.

Mihalcea, R. and Tarau, P. 2004. TextRank: Bringing Order into Texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain.

Mikheev, A., Moens, M. and Grover, C. 1999. Named Entity Recognition without Gazetteers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*. Bergen, Norway, pp.1-8.

Miller, G. and Charles, W. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, Vol. 6(1), pp.1-28.

Miller, S., Guinness, J. and Zamanian, A. 2004. Name Tagging with Word Clusters and Discriminative Training. In *Proceedings of 2004 Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting*. Boston, USA, pp.337-42.

Milne, D., Medelyan, O. and Witten, I. 2006. Mining Domain-specific Thesauri from Wikipedia: a Case Study. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. Hongkong, China, pp.442-48.

Milne, D. and Witten, I. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of the AAAI2008 Workshop on Wikipedia and Artificial Intelligence*. Chicago, Illinois, USA, pp.25-30.

Minkov, E., Cohen, W. and Ng, A. 2006. Contextual Search and Name Disambiguation in Email using Graphs. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieva*. Seattle, WA, USA, pp.27-34.

Minkov, E., Wang, R. and Cohen, W. 2005. Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, B.C., Canada, pp.443-45.

Mizzaro, S. 1997. Relevance: The Whole History. *Journal of the American Society for Information Science*, Vol. 48(9), pp.810-32.

Mohit, B. and Hwa, R. 2005. Syntax-based Semi-supervised Named Entity Tagging. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Ann Arbor, Michigan, USA, pp.57-60.

Mooney, R. 1996. Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning. In *Proceedings of the 1996 Conference on Empirical Methods in NaturalLanguage Processing (EMNLP)*., pp.82-91.

Morante, R., Asch, V. and Daelemans, W. 2009. A Memory-based Learning Approach to Event Extraction in Biomedical Texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*. Boulder, Colorado, USA, pp.59-67.

Morgan, A. and Hirschman, L. 2003. Gene Name Extraction Using FlyBase Resources. In *In ACL2003 Workshop on Language Processing in Biomedicine*. Sapporo, Japan, pp.1-8.

Morris, J. and Hirst, G. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Journal of Computational Linguistics*, Vol. 17(1), pp.21-48.

Morris, J. and Hirst, G. 2004. Non-classical Lexical Semantic Relations. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*. Boston, USA, pp.46-51.

Mulwad, V., Finin, T., Syed, Z. and Joshi, A. 2010. Using Lnked Data to Interpret Tables. In *Proceedings of the ISWC2010 1st International Workshop on Consuming Linked Data*. Shanghai, China.

Murphy, T., McIntosh, T. and Curran, J. 2006. Named Entity Recognition for Astronomy Literature. In *Australian Language Technology Workshop*. Sydney, Australia, pp.59-66.

Nadeau, D. 2007a. A Survey on Named Entity Recognition and Classification. *Linguisticae Investigationes*, Vol. 30(1), pp.1-26.

Nadeau, D. 2007b. *PhD Thesis: Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. PhD Thesis. University of Ottawa.

Nakashole, N., Theobald, M. and Weikum, G. 2011. Scalable Knowledge Harvesting with High Precision and High Recall. In *Proceedings of the 4th International ACM Conference on Web Search and data Mining*. Hongkong, China, pp.227-36.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., ShuKai, H., Tzu-Yi, K., Magistry, P. and Chu-Ren, H. 2009. Wiktionary and NLP: Improving Synonymy Networks. In *Proceedings of the ACL-IJCNLP2009 Workshop on the People's Web Meets NLP*. Singapore, Singapore, pp.19-27.

Navigli, R. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, Vol. 41(2), pp.10:1-10:69.

Nenadić, G., Rice, S., Spasić, I., Ananiadou, S. and Stapley, B. 2003. Selecting Text Features for Gene Name Classification: from Documents to Terms. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*. Sapporo, Japan, pp.121-28.

Ng, H. 1997. Getting Serious about Word Sense Disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* Madrid, Spain, pp.1-7.

Niu, W., Li, W., Ding, J. and Srihari, R. 2003. Bootstrapping for Named Entity Tagging using Concept-based Seeds. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*. Edmonton, Canada.

Nobata, C., Collier, N. and Tsujii, J. 2000. Comparison between Tagged Corpora for the Named Entity Task. In *Proceedings of the ACL2000 workshop on Comparing corpora*. Hongkong, China, pp.20-26.

Nuray-Turan, R., Kalashnikov, D. and Mehrotra, S. 2007. Self-tuning in Graph-based Reference Disambiguation. In *Proceedings of the 12th international conference on Database systems for advanced applications*. Bangkok, Thailand, pp.325-36.

Ohta, T., Tateisi, Y. and Kim, J. 2002. The GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. In *Proceedings of the 2nd international conference on Human Language Technology Research*. San Diego, California, USA, pp.82-86.

Olsson, F. 2008. *Bootstrapping Named Entity Annotation by Means of Active Machine Learning: A Method for Creating Corpora*. PhD Thesis. University of Gothenburg.

Othman, R., Deris, S. and Illias, R. 2007. A Genetic Similarity Algorithm for Searching the Gene Ontology Terms and Annotating Anonymous Protein Sequences. *Journal of Biomedical Informatics*, Vol. 41(1), pp.529-38.

Page, L., Brin, S., Motwani, R. and Winograd, T. 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford: Stanford InfoLab Stanford InfoLab.

Pandian, S., Geetha, T. and Krishna. 2007. Named Entity Recognition in Tamil using Context-cues and the E-M Algorithm. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Hyderabad, India, pp.1951-58.

Pan, S. and McKeown, K. 1999. Word Informativeness and Automatic Pitch Accent Modeling. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large CorporaE PROCESSING ANDVERY LARGE CORPORA*. College Park, MD, USA, pp.148-57.

Pantel, P., Crestan, E., Borkovsky, A., Popescu, A. and Vyas, V. 2009. Web-scale Distributional Similarity and Entity Set Expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, Singapore, pp.938-47.

Pan, S. and Yang, Q. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22(10), pp.1345-59.

Papineni, K. 2001. Why Inverse Document Frequency. In *Proceedings of the North American Chapter ofthe Association for Computational Linguistics*. Pittsburgh, PA, USA, pp.25-32.

Pasca, M. 2004. Acquisition of Categorized Named Entities for Web Search. In *Proceedings of the 13th ACM international conference on Information and knowledge management*. Washington DC, USA, pp.137-45.

Patwardhan, S. and Pedersen, T. 2006. Using WordNet based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL2006 Workshop on Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together in EAC*. Trento, Italy, pp.1-8.

Pedersen, T. 1998. *Learning Probabilistic Models of Word Sense Disambiguation*. Ph.D. dissertation. Dallas: Southern Methodist University Southern Methodist University.

Pedersen, T., Banerjee, S. and Patwardhan, S. 2005a. *Maximizing Semantic Relatedness to Perform Word Sense Disambiguation*. Research Report. Minneapolis, MN.: University of Minnesota Supercomputing Institute University of Minnesota Supercomputing Institute.

Pedersen, T., Pakhomov, S., Patwardhan, S. and Chute, C. 2006. Measures of Semantic Similarity and Relatedness in the Biomedical Domain. *Journal of Biomedical Informatics*, Vol. 40(3), pp.288-99.

Pedersen, T., Purandare, A. and Kulkarni, A. 2005b. Name Discrimination by Clustering Similar Contexts. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City, Mexico, pp.220-31.

Pekar, V. and Staab, S. 2002. Taxonomy Learning: Factoring the Structure of a Taxonomy into a Semantic Classification Decision. In *Proceedings of the 19th International Conference on Computational Linguistics*. Taipei, Taiwan, pp.786-92.

Peng, Y., He, D. and Mao, M. 2006. Geographic Named Entity Disambiguation with Automatic Profile Generation. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (2006)*. Hongkong, China, p.522–525.

Penn Treebank Project. 1998. *Penn Treebank POS*. [Online] Available at: http://www.cis.upenn.edu/~treebank/ [Accessed 13 Dec 2011].

Pesquita, C., Faria, D., Falc ão, A., Lord, P. and Couto, F. 2009. Semantic Similarity in Biomedical Ontologies. *PLoS Computational Biology*, Vol. 5(7).

Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K. and Duch, W. 2007. A Shared Task Involving Multi-label Classification of Clinical Free Text. In *Proceedings of the ACL2007 Workshop on BioNLP*. Prague, Czech, p.97–104.

Petrakis, E., Varelas, G., Hliaoutakis, A. and Raftopoulou, P. 2006. Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Different Ontologies. In *Proceedings of the 4th Workshop on Multimedia Semantics (WMS'06)*. Chania, Crete, Greece, pp.44-52.

Pinto, D., Tovar, M., Vilariño, D., Beltrán, B. and Somodevilla, J. 2009. BUAP_1: A Naïve Approach to the Entity Linking Task. In *Proceedings of the TAC2009 Knowledge Base Population Track in the Text Analysis Conference*.

Pirro, G. 2009. A Semantic Similarity Metric Combining Features and Intrinsic Information Content. *Data and Knowledge Engineering*, Vol. 68(11), pp.1289-308.

Ponomareva, N., Pla, F., Molina, A. and Rosso, P. 2007. Biomedical Named Entity Recognition: a Poor Knowledge HMM-based Approach. In *Proceedings of the 12th international conference on Applications of Natural Language to Information Systems (NLDB)*. Paris, France, pp.382-87.

Ponzetto, S. and Strube, M. 2011. Taxonomy Induction based on a Collaboratively Built Knowledge Repository. *Journal of Artificial Intelligence*, Vol. 175(9-10), pp.1737-58.

Porter, M. 1980. An Algorithm for Suffix Stripping. *Program*, Vol. 14(3), pp.130-37.

Pozo, A., Pazos, F. and Valencia, A. 2008. Defining Functional Distances over Gene Ontology. *Journal of BMC Bioinformatics*, Vol. 9(50).

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J. and Salakoski, T. 2007. BioInfer: a Corpus for Information Extraction in the Biomedical Domain. *Journal of BMC Bioinformatics*, Vol. 8(50).

Rada, R., Mili, H., Bicknell, E. and Blettner, M. 1989. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 19(1), pp.17-30.

Radinskty, K., Agichtein, E., Gabrilovich, E. and Markovitch, S. 2011. A Word at a Time: Computing Word Relatedness using Temporal Semantic Analysis. In *Proceedings of the 20th International World Wide Web Conference*. Hyderabad, India, pp.337-46.

Ratinov, L. and Roth, D. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL)*. Boulder, CO, USA, pp.147--155.

Rennie, J. and Jaakkola, T. 2005. Using Term Informativeness for Named Entity Detection. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. Salvador, Brazil, pp.353-60.

Resnik, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the International Joint Conference on AI (IJCAI)*. Montreal, Quebec, Canada, pp.448-53.

Riensche, R., Baddeley, B., Sanfilippo, A., Posse, C. and Gopalan, B. 2007. XOA: Web-Enabled Cross-Ontological Analytics. In *IEEE Congress on Services*. Salt Lake City, UT, USA, pp.99-105.

Riloff, E. and Jones, R. 1999. Learning Dictionaries for Information Extraction by Multi-level Bootstrapping. In *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*. Orlando, Florida, USA, pp.474-79.

Roberts, A., Gaizasukas, R., Hepple, M. and Guo, Y. 2008. Combining Terminology Resources and Statistical Methods for Entity Recognition. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*. Marrakech, pp.2974-80.

Rodr ǵuez, M. and Egenhofer, M. 2003. Determining Semantic Similarity among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15(2), pp.442-56.

Rose, T., Stevenson, M. and Whitehead, M. 2002. The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Las Palmas, Canary Islands, Spains, pp.29-31.

Rubenstein, H. and Goodenough, J. 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, Vol. 8(10), pp.627-33.

Ruiz-Casado, M., Alfonseca, E. and Castells, P. 2005. Using Context-window Overlapping in Synonym Discovery and Ontology Extension. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Borovets, Bulgaria.

Sahami, M. and Heilman, T. 2006. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In *Proceedings of the 15th international conference on World Wide Web*. Edinburgh, UK, pp.377-86.

Saha, S., Sarkar, S. and Mitra, P. 2009. Feature Selection Techniques for Maximum Entropy based Biomedical Named Entity Recognition. *Journal of Biomedical Informatics*, Vol. 42(5), pp.905-01.

Sang, E. and Meulder, F. 2003. Introduzion to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the 2003 Conference on Computational Natural Language Learning*. Edmonton, Canada, pp.142-47.

Saracevic, T. 1991. Individual Differences in Organizing, Searching, and Retrieving Information. In *Proceedings of the 54th Annual American Society for Information Science (ASIS) meeting*. Washington DC, USA, pp.82-86.

Sarawagi, S. 2007. Information Extraction. *Foundations and Trends in Databases*, Vol. 1(3), pp.261-377.

Sasaki, Y., Rea, B. and Ananiadou, S. 2007. Multi-topic Aspects in Clinical Text Classification. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*. Fremont, CA, USA, pp.62-67.

Sasaki, Y., Tsuruoka, Y., McNaught, J. and Ananiadou, S. 2008. How to Make the Most of NE Dictionaries in Statistical NER. *BMC Bioinformatics*, Vol. 9(11), p.S5.

Schickel-Zuber, V. and Faltings, B. 2007. OSS: A Semantic Similarity Function based on Hierarchical Ontologies. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Hyderabad, India, pp.551-56.

Schlicker, A., Domingues, F., Rahnenführer, J. and Lengauer, T. 2006. A New Measure for Functional Similarity of Gene Products based on Gene Ontology. *Journal of BMC Bioinformatics*, Vol. 7(302).

Schütze, H. 1998. Automatic Word Sense Discrimination. *Journal of Computational Linguistics - Special issue on word sense disambiguation archive*, Vol. 24(1), pp.97-124.

Seco, N., Veale, T. and Hayes, J. 2004. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In *Proceedings of the 16th European conference on Artificial Intelligence (ECAI)*. Valencia, Spain, pp.1089-90.

Seki, K. and Mostafa, J. 2003. A Probabilistic Model for Identifying Protein Names and Their Name Boundaries. In *Proceedings of the IEEE Computer Society Conference on Bioinformatics*. Stanford, CA, USA, pp.251-58.

Sevilla, J., Segura, V., Podhorski, A., Guruceaga, E., Mato, J., Martinez-Cruz, L., Corrales, F. and Rubio, A. 2005. Correlation between Gene Expression and GO Semantic Similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 2(4), pp.330-38.

Sheng, M., Chen, H., Yu, T. and Feng, Y. 2010. Linked Data based Semantic Similarity and Data Mining. In *IEEE International Conference on Information Reuse and Integration (IRI)*. Las Vegas, USA, pp.104-08.

Shen, D., Zhang, J., Su, J., Zhou, G. and Tan, C. 2004. Multi-criteria-based Active Learning for Named Entity Recognition. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Barcelona, Spain, pp.589-96.

Singh, A. 2011. Named Entity Recognition for South and South East Asian Languages:Taking Stock. In *Proceedings of the Workshop on NER for South and South East Asian Languages, in the 3rd International Joint Conferenceon Natural Language Processing (IJCNLP)*. Hyderabad, India, pp.5-16.

Smith, D. 2002. Detecting and Browsing Events in Unstructured Text. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. Tampere, Finland, pp.73-80.

Smith, A. and Osborne, M. 2006. Using Gazetteers in Discriminative Information Extraction. In *Proceedings of the 10th Conference on Computational Natural Language Learning*. New York City, USA, pp.10-18.

SNOMED CT. 2002. *Systematized Nomenclature of Medicine - Clinical Terms*. [Online] Available at: http://www.connectingforhealth.nhs.uk/systemsandservices/data/uktc/snomed [Accessed 15 Mar 2012].

Spark Jones, K. 1973. Index Term Weighting. *Information Storage and Retrieval*, Vol. 9(11), pp.619-33.

Speer, N., Spieth, C. and Zell, A. 2004. A Memetic Clustering Algorithm for the Unctional Partition of Genes based on the Gene Ontology. In *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. La Jolla, CA, USA, p.25.

Srihari, R. and Peterson, E. 2008. Named Entity Recognition for Improving Retrieval and Translation. In *Proceedings of the 11th International Conference on Asian Digital Libraries*. Bali, Indonesia, pp.404-05.

Srinivasan, H., Chen, J. and Srihari, R. 2009. Cross Document Person Name Disambiguation using Entity Profiles. In *Proceedings of the TAC2009 Knowledge Base Population Track at the Text Analysis Conference*.

STAR. 2007. *STAR - Semantic Technologies for Archaeological Resources*. [Online] Available at: http://hypermedia.research.glam.ac.uk/kos/STAR/ [Accessed 13 Mar 2012].

Steven, A. 2002. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. San Francisco, USA, pp.360-67.

Strube, M. and Ponzetto, S. 2006. WikiRelate! Computing Semantic Relatedness using Wikipedia. In *Proceedings of the 21st national conference on Artificial intelligence*. Boston, USA, pp.1419-24.

Sussna, M. 1993. Word Sense Disambiguation for Free-text Indexing using a Massive Semantic Network. In *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management*. New Orleans, Louisiana, USA, pp.67-74.

TAC KBP Track. 2009. *TAC 2009 Knowledge Base Population Track*. [Online] Available at: http://apl.jhu.edu/~paulmac/kbp.html [Accessed 14 Mar 2012].

TAC KBP Track. 2010. *TAC 2010 Knowledge Base Population (KBP2010) Track*. [Online] Available at: http://nlp.cs.qc.cuny.edu/kbp/2010/ [Accessed 14 Mar 2012].

Takeuchi, K. and Collier, N. 2002. Use of Support Vector Machines in Extended Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning*. Taipei, Taiwan, pp.119-25.

Talukdar, P., Brants, T., Liberman, M. and Pereira, F. 2006. A Context Pattern Induction Method for Named Entity Extraction. In *Proceedings of the 10th Conference on Computational Natural Language Learning*. New York City, USA, pp.141-48.

Tanabe, L., Xie, N., Thom, L., Matten, W. and Wilbur, W. 2005. GENETAG: a Tagged Corpus for Gene/Protein Named Entity Recognition. *Journal of BMC bioinformatics*, Vol. 6(1).

Thahir, M., Estevam, H. and Mitchell, T. 2011. Discovering Relations between Noun Categories. In *Poceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Edinburgh, UK, pp.1447-55.

The British National Corpus. 2007. *The British National Corpus, version 3 (BNC XML Edition)*. [Online] (3) Available at: http://www.natcorp.ox.ac.uk/ [Accessed 14 Mar 2012].

The Gene Ontology Consortium. 2000. Gene Ontology: Tool for the Unification of Biology. *Nature Genet*, Vol. 25, pp.25-29.

The Yapex Corpus. 2005. *Proteinhalt i text*. [Online] Available at: http://www.sics.se/humle/projects/prothalt/ [Accessed 14 Mar 2012].

Thelan, M. and Riloff, E. 2002. A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Context. In *Proceedings of the conference on Empirical methods in natural language processing*. Philadelphia, USA, pp.214-21.

Toral, A. and Munoz, R. 2006. A Proposal to Automatically Build and Maintain Gazetteers for Named Entity Recognition by using Wikipedia. In *Proceedings of the Workshop on New Text, in the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Trento, Italy, pp.56-61.

TREC Entity Track. 2011. *TREC Entity Track Searching for Entities and Properties of Entities*. [Online] Available at: http://ilps.science.uva.nl/trec-entity/ [Accessed 9 Mar 2012].

TREC-8 QA Data. 2002. *TREC-8 (1999) QA Data*. [Online] Available at: http://trec.nist.gov/data/qa/t8_qadata.html [Accessed 30 September 2011].

Tsai, R., Wu, S., Chou, W., Lin, Y., He, D., Hsiang, J., Sung, T. and Hsu, W. 2006. Various Criteria in the Evaluation of Biomedical Named Entity Recognition. *Journal of BMC Bioinformatics*, Vol. 7(92).

Tsatsaronis, G., Varlamis, I. and Vazirgiannis, M. 2010. Text Relatedness based on a Word Thesaurus. *Journal of Artificial Intelligence Research*, Vol. 37(1), pp.1-39.

Tsochantaridis, I., Joachims, T. and Hofmann, T. 2005. Large Margin Methods for Structured and Interdependent Output Variables. *Journal of MachineLearning Research*, Vol. 6(Sep), pp.1453-84.

Turdakov, D. and Velikhov, P. 2008. Semantic Relatedness Metric for Wikipedia Concepts Based on Link Analysis and its Application to Word Sense Disambiguation. In *Proceedings of the Spring Young Researcher's Colloquium On Database and Information Systems*. St. Petersburg, Russia.

Tversky, A. 1977. Features of Similarity. *Psychological Review*, Vol. 84(2), p.327–352.

Usami, Y., Cho, H., Okazaki, N. and Tsujii, J. 2011. Automatic Acquisition of Huge Training Datafor Bio-Medical Named Entity Recognition. In *Poceedings of the ACL2011 workshop on BioNLP*. Portland, USA, pp.65-77.

Uzuner, O. 2008. Second i2b2 Workshop on Natural Language Processing Challenges for Clinical Records. In *Proceedings of the AMIA Annual Symposium*., pp.1252-53.

Vlachos, A. and Gasperin, C. 2006. Bootstrapping and Evaluating Named Entity Recognition in the Biomedical Domain. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*. New York City, USA, pp.138-45.

Wacholder, N., Ravin, Y. and Choi, M. 1997. Disambiguation of Proper Names in Text. In *Proceedings of the 5th conference on Applied natural language processing (1997)*. Washington DC, USA, pp.202-08.

Wang, Y. 2006. Annotating and Recognising Named Entities in Clinical Notes. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*. Singapore, Singapore, pp.18-26.

Wang, J., Du, Z., Payattakool, R., Yu, P. and Chen, C. 2007. A New Method to Measure the Semantic Similarity of GO Terms. *Journal BMC Bioinformatics*, Vol. 23(10), pp.1274-81.

Wang, T. and Hirst, G. 2011. Refining the Notions of Depth and Density in WordNet-based Semantic Similarity Measures. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, UK, pp.1003-11.

Wang, A., Hoang, C. and Kan, M. 2010. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, pp.1-21.

Wan, X., Zhong, L., Huang, X., Ma, T., Jia, H., Wu, Y. and Xiao, J. 2011. Named Entity Recognition in Chinese News Comments on the Web. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand, pp.856-64.

Weeds, J. 2003. *Measures and Applications of Lexical Distributional Similarity*. PhD thesis. East Sussex, UK: University of Sussex.

WePS. 2007. *WePS: Searching Information about Entities in the Web*. [Online] Available at: http://nlp.uned.es/weps/index.php [Accessed 16 Mar 2012].

Wikipedia. 2011. *Support Vector Machine*. [Online] Available at: http://en.wikipedia.org/wiki/Support_vector_machine [Accessed 5 Dec 2011].

Wilbur, W., Rzhetsky, A. and Shatkay, H. 2006. New Directions in Biomedical Text Annotation: Definitions, Guidelines and Corpus Construction. *Journal of BMC Bioinformatics*, Vol. 7(356).

Wojtinnek, P. and Pulman, S. 2011. Semantic Relatedness from Automatically Generated Semantic Networks. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS'11)*. Koblenz, Germany, pp.390-94.

Wong, B. and Kit, C. 2011. Comparative Evaluation of Term Informativeness Measures in Machine Translation Evaluation Metrics. In *MT Summit 2011*. Xiamen, China, pp.537-44.

Woods, J., Sneiderman, C., Hameed, K., Ackerman, M. and Hatton, C. 2006. Using UMLS Metathesaurus Concepts to Describe Medical Images: Dermatology Vocabulary. *Journal of Computers in Biology and Medicine*, Vol. 36(1), pp.89-100.

Wu, Y., Fan, T., Lee, Y. and Yen, S. 2006b. Extracting Named Entities using Support Vector Machines. *Lecture Notes in Bioinformatics(LNBI): Knowledge Discovery in Life ScienceLiterature*, Vol. 3886, pp.91-103.

Wu, Z. and Palmer, M. 1994. Verbs Semantics and Lexical Selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Las Cruces, New Mexico, USA, pp.133-38.

Wu, H., Su, Z., Mao, F., Olman, V. and Xu, Y. 2005. Prediction of Functional Modules based on Comparative Genome Analysis and Gene Ontology Application. *Nucleic Acids Research*, Vol. 33(9), p.2822–2837.

Wu, X., Zhu, L., Guo, J., Zhang, D. and Lin, K. 2006a. Prediction of Yeast Protein Protein Interaction Network: Insights from the Gene Ontology and Annotations. *Nucleic Acids Research*, Vol. 34(7), p.2137–2150.

Yang, D. and Powers, D. 2005. Measuring semantic similarity in the taxonomy of WordNet. In *proceedings of the 28th Australasian conference on Computer Science*.

Yazdani, M. and Popescu-Belis, A. 2010. A Random Walk Framework to Compute Textual Semantic Similarity: a Unified Model for Three Benchmark Tasks. In *Proceedings of the 4th IEEE International Conference on Semantic Computing (ICSC)*. Pittsburgh, PA, USA, pp.424-29.

Yeh, E., Ramage, D., Manning, C., Agirre, E. and Soroa, A. 2009. WikiWalk: Random Walks on Wikipedia for Semantic Relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, ACL-IJCNLP'09*. Singapore, Singapore, pp.41-49.

Ye, P., Peyser, B., Pan, X., Boek, J., Spencer, F. and Bader, J. 2005. Gene Function Prediction from Congruent Synthetic Lethal Interactions in Yeast. *Molecular system biology*, Vol. 1(0).

Yu, H., Gao, L., Tu, K. and Guo, Z. 2005. Broadly Predicting Specific Gene Functions with Expression Similarity and Taxonomy Similarity. *Gene*, Vol. 352, pp.75-81.

Zesch, T. and Gurevych, I. 2007. Analysis of the Wikipedia Category Graph for NLP Applications. In *Proceedings of the TextGraphs-2 Workshop at NAACL-HLT'07*. Rochester, NY, USA, pp.1-8.

Zesch, T. and Gurevych, I. 2010a. Wisdom of Crowds versus Wisdom of Linguists: Measuring the Semantic Relatedness of Words. *Journal of Natural Language Engineering*, Vol. 16(1), pp.25-29.

Zesch, T., Müller, C. and Gurevych, I. 2008a. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Language Resources and Evaluation*. Marrakech, Morocco, pp.1646-52.

Zesch, T., Müller, C. and Gurevych, I. 2008b. Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the 23rd national conference on Artificial intelligence (AAAI'08)*. Chicago, Illinois, USA.

Zhang, L., Pan, Y. and Zhang, T. 2004. Focused Named Entity Recognition using Machine Learning. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. Sheffield, UK, pp.281-88.

Zhang, W., Sim, Y., Su, J. and Tan, C. 2010. NUS-I2R: Learning a Combined System for Entity Linking. In *Proceedings of the TAC2010 Knowledge Base Population Track at the Text Analysis Conference*.

Zhang, K., Zi, J. and Wu, L. 2007. New Event Detection based on Indexing-tree and Named Entity. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. Amsterdam, The Netherlands, pp.215-22.

Zhou, G. and Su, J. 2004. Exploring Deep Knowledge Resources in Biomedical Name Recognition. In *Proceedings of the COLING2004 International Workshop on Natural Language Processing in Biomedicine and its Applications*. Geneva, Switzerland, pp.99-102.

Zhou, X., Zhang, X. and Hu, X. 2007. Dragon Toolkit: Incorporating Auto-learned Semantic Knowledge into Large-Scale Text Retrieval and Mining. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. Patras, Greece, pp.197-201.

Ziegler, C., Simon, K. and Lausen, G. 2006. Automatic Computation of Semantic Proximity using Taxonomic Knowledge. In *Proceedings of the 15th ACM international conference on Information and knowledge management*. Arlington, Virginia, USA, pp.465-74.