Exploiting Domain Knowledge for Cross-domain Text Classification in Heterogeneous Data Sources

Andrea Varga

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy at
the University of Sheffield
Department of Computer Science

June 2014

Abstract

With the growing amount of data generated in *large heterogeneous repositories* (such as the Word Wide Web, corporate repositories, citation databases), there is an increased need for the end users to locate relevant information efficiently. *Text Classification* (TC) techniques provide automated means for classifying fragments of text (phrases, paragraphs or documents) into predefined semantic types, allowing an efficient way for organising and analysing such large document collections. Current approaches to TC rely on supervised learning, which perform well on the domains on which the TC system is built, but tend to adapt poorly to different domains.

This thesis presents a body of work for exploring *adaptive TC techniques* across heterogeneous corpora in large repositories with the goal of finding novel ways of bridging the gap across domains. The proposed approaches rely on the exploitation of domain knowledge for the derivation of *stable cross-domain features*. This thesis also investigates novel ways of *estimating the performance of a TC classifier*, by means of domain similarity measures. For this purpose, two novel knowledge-based similarity measures are proposed that capture the usefulness of the selected cross-domain features for cross-domain TC. The evaluation of these approaches and measures is presented on real world datasets against various strong baseline methods and content-based measures used in transfer learning.

This thesis explores how domain knowledge can be used to enhance the representation of documents to address the lexical gap across the domains. Given that the effectiveness of a text classifier largely depends on the availability of annotated data, this thesis explores techniques which can leverage data from social knowledge sources (such as DBpedia and Freebase). Techniques are further presented, which explore the feasibility of exploiting different *semantic graph structures* from knowledge sources in order to create novel cross-domain features and domain similarity metrics. The methodologies presented provide a novel representation of documents, and exploit four wide coverage knowledge sources: DBpedia, Freebase, SNOMED-CT and MeSH.

The contribution of this thesis demonstrates the feasibility of exploiting domain knowledge for adaptive TC and domain similarity, providing an enhanced representation of documents with semantic information about entities, that can indeed reduce the lexical differences between domains.

Publications

The work in this thesis is present within the following publications:

- Varga, A. and Ciravegna, F. 2014. Ontology-driven Cross-domain Document Zone Classification (in preparation for the Journal of Biomedical Informatics)
- Varga, A., Cano, A.E., Rowe, M., Ciravegna, F. and He, Y. 2013. Linked Knowledge Sources for Topic Classification of Tweets: A semantic-graph based approach. In the Journal of Web Semantics, Volume 26, May 2014, Pages 36-57
- Cano, A.E., Rizzo, G., Varga, A., Rowe, M., Stankovic, M., Dadzie, A. 2014. Making Sense of Microposts (#Microposts2014) Named Entity Extraction & Linking Challenge. In proceedings of WWW Companion Volume, at the International World Wide Web Conference (WWW' 14)
- Cano, A.E., Rowe, M., Varga, A., Stankovic, M., Dadzie, A. 2013. Making Sense of Microposts (#MSM2013) Concept Extraction Challenge. In proceedings of Concept Extraction Challenge at #MSM2013, at the International World Wide Web Conference (WWW' 13)
- Varga, A., Cano, A.E., Ciravegna, F. and He, Y. 2013. On the Study of Reducing the Lexical Differences between Social Knowledge Source and Twitter for Topic Classification. In Studia Universitatis Babes-Bolyai, Informatica Journal
- Cano, A.E., Varga, A., Rowe, M., Ciravegna, F. and He, Y. 2013. Harnessing Linked Knowledge Sources for Topic Classification in Social Media. In proceedings of the ACM Hypertext and Social Media Conference (ACM HT' 13)
- Varga, A., Cano, A.E. and Ciravegna, F. 2012. Exploring the Similarity between Social Knowledge Sources and Twitter for Cross-domain Topic Classification. In proceedings of Knowledge Extraction and Consolidation from Social Media Workshop (KECSM2012), at the International Semantic Web Conference (ISWC 2012)
- Varga, A., Preotiuc-Pietro, D. and Ciravegna, F. 2012. Unsupervised Document Zone Identification using Probabilistic Graphical Models. In proceedings of the International Language Resources and Evaluation Conference (LREC 2012)

Other related publications:

- Mazumdar, S., Varga, A., Lanfranchi, V., Petrelli, D. and Ciravegna, F. 2011. A Knowledge Dashboard for Manufacturing Industries. Revised Selected Papers. In proceedings of the Extended Semantic Web Conference (ESWC 2011)
- Cano, A.E., Varga, A., and Ciravegna, F. 2011. Volatile Classification of Point of Interests based on Social Activity Streams. In proceedings of the Social Data on the Web Workshop (SDOW2011), at the International Semantic Web Conference (ISWC 2011)

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Fabio Ciravegna, for his continued support and guidance throughout my PhD research. I am extremely thankful for always showing a sincere interest in my work, for the extensive discussions concerning my research ideas, for his constructive feedback, and for all the help he has given me throughout my doctoral studies. I would also like to express my gratitude to Rolls-Royce PLC., for providing me the opportunity to collaborate with the members of the Samulet research project, and evaluate some of my research findings in an industrial setting. Many thanks to my thesis examiners, Dr. Lucia Specia and Dr. Georgios Paltoglou, for their valuable comments, which helped improve this thesis to a great extent.

I am grateful to all my colleagues and former members of the Organisations, Information and Knowledge Group for their friendship and feedback on my work. My special thanks go to Dr. Daniel Preoţiuc-Pietro and Dr. Elizabeth Cano for discussions regarding the machine learning aspects of my research. The reading group seminars organised by Daniel Preoţiuc-Pietro helped my to understand how graphical models worked and how I could adapt and apply such techniques to document zoning. I am extremely grateful to Elizabeth Cano for stimulating discussions on the application of transfer learning techniques to social media. A million thanks to Dr. Aba-Sah Dadzie, Dr. Iustina Ilisei and Dr. Laura Hasler for taking the time to proofread key chapters of my thesis.

I owe my thanks to Dr. Trevor Cohn, my thesis committee member, for his useful and enthusiastic discussions on transfer learning and graphical models. His expertise in Machine Learning and Natural Language Processing, as well as his critical eye for details, have particularly helped many parts of this work.

I am also thankful to the University of Sheffield for making this thesis possible by providing access to the iceberg, a high performance computing cluster, for running my experiments. It was an excellent resource to execute many experiments in parallel.

I would like to extend my gratitude to my former research advisor Professor Doina Tatar, who first introduced me to NLP and kindled my research interest in this domain. I am also grateful for the collaboration and friendship with the members of the Research Group in Computational Linguistics from Wolverhampton, with whom I took my first NLP steps.

Last and most of all, I am deeply indebted to my parents and sister for their immense love and support.

Chapter 1

Introduction

1.1 Motivation

A vast amount of electronic information is available in an unstructured format. Large corporate enterprises typically keep records of enormous historical data about their products in various company-wide repositories. In the case of the aerospace industry, the lifecycle of a jet engine model, covering up to 50 years of design, maintenance, tests and service data are all documented in textual format, which can easily sum up to several terabytes. Other classic examples for large textual repositories are the biomedical journal repositories published on the Web, serving as important resources for biomedical practitioners aiming to keep abreast with current research. Pubmed[1], the largest biomedical repository, comprises over 22 million articles, having a rapid rate of publishing, which can reach 1 paper per minute[2]. Nowadays, with the rise of social media, large enterprises are also interested in mining information about important events (e.g., emergency landings) concerning their products (e.g., aircrafts or engines) from social media websites. Popular social media platforms such as Twitter[3] and Facebook[4], provide up-to-date information about events happening in the world on a wide range of topics. They also constitute primary sources of information, often distributing information faster and earlier than traditional news sources [Blanchard et al., 2012]. The estimated rate of messages posted (called tweets) in Twitter, for example, can hit half billion of tweets per day[5].

To handle and organise this large unstructured heterogeneous document collection, there is a need for automatic techniques to extract and organise the information and assign semantic meaning to it. Automatic *text classification* (TC) (also called *text categorisation*) is a suitable approach for organising large amounts of data, providing automated means to categorise *documents* (or *text fragments*) into predefined semantic *categories* (or *classes* or *topics*). TC can be performed at different *granularity levels*, depending on the application at hand, ranging from *fine-grained within-document* text classification (e.g., aimed at assigning semantic classes to text fragments, such as paragraphs or phrases) to more *coarse-grained whole-document* text classification (e.g., aimed at assigning semantic classes

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://duncan.hull.name/2010/07/15/fifty-million/
[3] https://twitter.com
[4] https://en-gb.facebook.com
[5] http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-a-billion-tweets-a-day/

to documents). A typical example of the former TC case could be, when an engineer is seeking enhanced knowledge to resolve a technical issue on a particular product (e.g., engine) of the company. This task involves identifying the root causes of the issue that occurred on the product (e.g., cracks generated by a specific cause on a specific part of an engine), as well as, finding problems encountered on other similar product types. Assuming that these product names (e.g., engine or component names) are already extracted from the documents, the goal of a TC system in this scenario is to analyse the documents mentioning these products, and to recognise text fragments (e.g., paragraphs) that provide additional contextual information about these products (e.g., being involved in the investigation, specific methods used to solve the problem, the investigation results and conclusions). In other situations, however, identifying the main topic of the documents could be useful in revealing existing cases of problems which have occurred. For instance, knowing that the topic or class of the document is accident/maintenance/failure/repair/service (e.g., engine fault), could allow to further generalise the problem space, and reach out to documents discussing problems occurred on products (car engine) different to the one under investigation (jet engine). In such situations, thus the role of a TC system is to provide a semantic categorisation of the documents as a whole, based on the topics discussed in them (e.g., fault).

Single domain TC systems are typically based on *supervised machine learning algorithms* that require a large amount of human annotated data that is often time consuming and expensive to obtain. These approaches can achieve good performance, when the source data (on which the system is trained) and target data (on which the system is tested) follow the same underlying word distribution. When this assumption is violated, however, the performance of a TC system can dramatically decrease [Raina et al., 2007].

The scale of these large repositories and their dynamic nature - new documents and text types[6] being created continuously - presents a clear motivation for the application of advanced TC techniques, i.e., *adaptive TC* techniques, which can deal with the *variations in the language*, *vocabulary* and *style* between the domains. These techniques are based on a new learning paradigm called *transfer learning* [Pan and Yang, 2010], which helps in transferring the knowledge acquired from the source data to the target data by designing *cross-domain pivot features* that are stable across domains. These approaches thus typically make use of large amount of data from a source domain to **train** the transfer learning classifier for the target domain. However, producing such annotated data may require domain expert knowledge, which can be costly and time consuming. The success of these approaches furthermore rely only on the **similarity** between the source and target domain data, and the usefulness of the **cross-domain features** employed. The language used in the domains can also pose additional challenges for adaptive TC systems. For example, when dealing with technical domains present in corporate environments, the language used is quite complex and the problem of TC becomes much more difficult [Guo et al., 2006; Wang, 2009]. On the other hand, when dealing with informal social media posts, the frequency of the misspellings, jargons and abbreviations makes TC challenging.

In order to address these challenges, this thesis investigates the use of *domain knowledge sources* for building adaptive text classifiers and designing domain similarity measures for TC. In particular, it investigates the extent to which the lexical representation of domains can be enhanced by leveraging the information present in *domain knowledge sources*.

[6] An example for different text types are technical documents written in different format.

Domain knowledge sources (KS) are important resources which provide a formal representation of domains, containing rich information and knowledge about domain-specific concepts. Over the past decade, significant advancements have taken place in developing domain knowledge sources and their underlying domain ontologies. For instance, in the biomedical domain, the Unified Medical Language System (UMLS)[7] has been developed, which encapsulates many different biomedical sub-domains and provides a mapping between their vocabularies. Furthermore, thanks to the Linked Open Data (LOD) Project[8], a set of multi-domain[9] KSs (e.g., WordNet[10], Wikipedia, Wikibooks[11], Freebase[12]) have been made freely available and interconnected with DBpedia[13] KS.

Two of the main advantages of exploiting these KSs is that they are freely available and they contain rich semantic information about concepts in a wide range of domains. This thesis thus proposes a range of novel techniques that aim at exploiting this semantic information about concepts with the goal of creating a set of stable *cross-domain features* for transfer learning. It furthermore proposes a novel set of *domain similarity measures*, which make use of the enhanced KS-based representation of domain documents.

1.2 Research Questions

The above problem setting motivates the research within this thesis. As mentioned earlier, TC can be performed at multiple granularity levels, depending on the real world scenario of interest. This thesis focuses on transfer learning approaches for adaptive TC systems, considering two different TC settings: *fine-grained within-document* text classification aiming to identify and classify text fragments (e.g., paragraphs) into predefined categories (e.g., zones); and *coarse-grained whole-document* text classification aiming to assign a particular label (e.g., topic) to individual documents. In particular, this thesis examines the usefulness of domain knowledge sources for bridging the distributional gap between domains. The main research question explored in this thesis is the following:

> *How can document classification be performed across multiple domains and text types?*

In light of the presented problem setting, this research question can furthermore be divided into the following questions:

1. *Is it possible to define automated techniques of text classification that are able to port across domains and text types?*
2. *Can labelled data be gathered inexpensively to build adaptive text classifiers?*
3. *Is it possible to define a measure for quantifying the adaptability of a text classifier?*
4. *Is the effectiveness of adaptive methods comparable to in-domain supervised machine learning methods?*

[7] http://www.nlm.nih.gov/research/umls/
[8] http://linkeddata.org
[9] According to http://lod-cloud.net, the "Media", "Life Science", "Geographic" and "Publications" domains represent over 44% of the LOD.
[10] http://semanticweb.cs.vu.nl/lod/wn30/
[11] http://en.wikibooks.org
[12] http://www.freebase.com
[13] http://dbpedia.org

1.3 Claims of this Thesis

Traditional approaches for TC represent the content of the documents using simple shallow techniques [Sebastiani, 2005]. A typical example is the bag-of-words (BoW) model, which represents the documents as a collection of words presented in the documents. Although this simple BoW representation has been successful for many applications, it can introduce additional challenges for transfer learning, given that different domains can exhibit large variations in the language and vocabulary used.

Further, the content of these documents commonly discusses about well-defined concepts within domains (e.g., person, location, engine type), which are also contained in knowledge sources (such as domain ontologies (SNOMED-CT), taxonomies (MeSH), encyclopaedias (DBpedia, Freebase)). These knowledge sources provide a large amount of information in machine readable format about the concepts (e.g., engine or person) within a domain (e.g., aerospace or biomedical domains), together with various *semantic structures* describing the relationships between the concepts (e.g., topical relatedness or concept hierarchies). Using domain knowledge sources, documents could be enhanced with rich information about the concepts, providing a generalisation over the entities discussed in them. This thesis claims that the lexical gap between domains can be reduced by inducing a new conceptual representation for the domain documents, based on the available background knowledge about concepts in knowledge sources. This claim is explicitly made as follows:

- *Domain knowledge sources contain useful semantic structures from which pivot features can be obtained for adaptive text classification*

Existing supervised transfer learning approaches assume that a large amount of annotated data is available in the source domain [Pan and Yang, 2010; Jiang, 2008a]. This annotated source domain data is then used to train a transfer learning classifier for labelling the target domain examples. However, it has long been recognised that creating and maintaining high quality annotations is time consuming and expensive [Ciravegna et al., 2002; Zhang et al., 2010]. Depending on the TC task, the creation of high quality annotations may require multiple domain experts working on the task, which can be laborious and costly. More importantly, given the drastic increase in the size of large repositories, creating these annotations may be infeasible. Data leveraged from domain knowledge sources, on the contrary, provides background information without the burden of annotated data construction. Domain knowledge sources such as DBpedia[14] and Freebase[15] constitute some of the largest repositories published online, containing an abundant amount of data on a large number of topics[16]. This thesis claims that KS data reflects the topics of the documents in the target domain:

- *Data found in domain knowledge sources can be used to train an adaptive text classifier*

Quantifying the similarity and dissimilarity between the source domain data and the target domain data provides an important insight into the applicability and success of a text classifier. It is expected that the closer the two domains are the better the performance of the text classifier is [Pan and Yang, 2010]. Designing such metrics is therefore potentially useful

[14] http://dbpedia.org
[15] http://www.freebase.com
[16] These knowledge sources will be presented in more details in Chapter 6.

when wanting to apply an existing text classification model on a new text type (e.g., blogs); or when wanting to collect new training data for a text classifier. Previous approaches for measuring transfer adaptability (or domain similarity) mostly focus on *content-based lexical similarity measures*, making use of features derived solely from the content of the documents (e.g., employing the simple BoW model). However, by exploiting the rich information about concepts in KSs, a new higher level concept abstraction can be created for domains, which could essentially improve the generalisation between domains. This thesis claims that the enhanced document representation can provide a better estimate on the transfer adaptability of a text classifier:

- *The accuracy of a text classifier can be measured as a function of conceptual representation of the domain documents*

For many Natural Language Processing (NLP) tasks including TC, transfer learning has been found to be a successful technique for building classification methods across multiple domains and text types. One of the main strength of these approaches is that they try to automatically learn the generalisation patterns between domains, aiming to bridge the distributional gaps across domains. However, research has shown that despite of the success of domain similarity measures aimed at estimating the performance of a classifier, to date, it is not very clear whether these approaches always perform better than in-domain machine learning approaches [Pan and Yang, 2010]. Situations when applying transfer learning worsens the performance of the learner are referred to as negative transfer, which is still considered an open issue [Pan and Yang, 2010]. This thesis claims that domain knowledge sources contain the necessary background information to enrich the documents within the domains, which can indeed improve the generalisation between domains:

- *Adaptive text classification techniques exploiting domain knowledge sources are able to achieve comparable results to in-domain machine learning approaches*

1.4 Contributions

This thesis presents a body of work exploring the benefit of using *domain knowledge* for *adaptive text classification* across *multiple domains* and *text types*. As a consequence, novel techniques and domain similarity measures are proposed, each addressing the corresponding text classification task:

- techniques and domain similarity measures for *within-document* TC:
  - *Supervised transfer learning algorithm*: A transfer learning technique is presented for *within-document* text classification. This technique extends the well known *Easy Adapt* [Daumé, 2007] transfer learning approach by incorporating background knowledge into a *within-document* text classifier. The novel contribution of this technique lies in the use of domain knowledge structures for the generation of pivot features. In addition, several augmentation strategies are presented for incorporating these features into a supervised transfer learning classifier.
  - *Novel unsupervised domain similarity measure*: A new domain similarity measure is presented, which functions in a fully unsupervised fashion, requiring no label information about the documents. In order to achieve this, probabilistic graphical models are exploited. These models discover the *within-document* zone segments of the documents by clustering the paragraphs of the documents using only lexical information (words) present in the documents. Following this, the similarity between domains is computed using different corpus-based and KS-based statistical measures between the discovered paragraph clusters.
- techniques and domain similarity measures for *whole-document* TC:
  - *Supervised transfer learning algorithm*: A novel transfer learning algorithm is presented for *whole-document* text classification, which relies on a simple feature augmentation strategy. The novel contribution of this approach lies in the proposal of a set of pivot features from domain knowledge sources and new weighting strategies for these features which take into account the relevance of each feature in the KS. Different feature augmentation strategies are furthermore exploited for incorporating these features into a supervised transfer learning classifier.
  - *Novel domain similarity measure*: Various entropy-based similarity measures are proposed for measuring the adaptability of a *whole-document* text classifier. These measures make use of the enhanced KS-based document representation of the domain documents.

In order to enable the evaluation of the proposed TC techniques in different cross-domain scenarios, several resources have been compiled:

- *A corpus of technical documents in the aerospace domain annotated with the novel document zone annotation schema*
- *A corpus of scientific documents in the biomedical domain annotated with IMRAD (Introduction-Method-Results-Abstract-Discussion) annotation schema*
- *A corpus of KS data (DBpedia and Freebase) annotated with topics related to Emergency Response*
- *A corpus of tweets annotated with topics related to Emergency Response*

The effectiveness of the explored techniques is furthermore validated conducting a detailed evaluation which compare the performance of these techniques over various well-known baseline methods used in transfer learning, such as a classifier built on the target domain

data (which is called *TGT_ONLY*), a classifier built on the source domain data (which is called *SRC_ONLY*), as well as the classifier built on the joint source and target domain data (called *SRC_TGT*):

- *An extensive evaluation of text classification techniques using real world data*

This thesis further provides a comprehensive literature review on transfer learning approaches for text classification.

1.5 Thesis Structure

The remainder of the thesis is divided into three main parts and structured as follows.

1.5.1 Part I - Background

Chapter 2 introduces the task of TC and describes different applications of TC in Natural Language Processing, ranging from *within-document* TC (e.g., document zoning) to *whole-document* TC (e.g., topic classification). A detailed description is then provided of the existing transfer learning approaches and settings for TC, together with the existing evaluation approaches, summarising the related work to TC. A discussion is also presented specifically on the role of domain knowledge sources for TC, and the possibility of exploiting these sources for adaptive TC.

1.5.2 Part II - Methodology

Chapter 3 proposes the use of domain knowledge sources for adaptive TC. It presents the requirements that adaptive TC techniques must fulfil when dealing with large heterogeneous repositories spanning multiple domains and text types. Following this discussion, an overview of the different stages of the approach is also presented, by discussing the processes of content modelling and context generation and incorporating both into an adaptive TC system. The concept of *semantic meta-graphs* derived from domain knowledge sources is also introduced. This serves the basis for creating pivot features for adaptive TC.

Chapter 4 starts by investigating the *unsupervised transfer learning scenario* for the *within-document* TC task, when there are no annotated data available for the source and target domains. For this case, probabilistic graphical models are proposed (the first such approaches for this tasks). These can flexibly model the content of documents and recognise the intra-document segments (zones) in them in a fully unsupervised fashion. Further, these approaches do *not require any domain knowledge* information for modelling the content of the documents, making them more practical for real world applications (e.g., document zoning in the aerospace industry), when such resources are not available, and creating them is a difficult task. These graphical models will also serve an important role for predicting the performance of a *within-document* TC classifier described in Chapter 5. In addition, this chapter also presents a detailed case study in the aerospace domain to investigate whether existing TC classification schemas capture the inta-document (zone) segments of the documents present in this domain, revealing the need for a novel *document zone annotation scheme* for the aerospace domain.

Chapter 5 continues the presentation of *supervised transfer learning approaches* for *within-document* TC, assuming that a large amount of annotated data is available in the source domain, and a small amount of annotated data is available in the target domain. A transfer learning algorithm is presented, which employs a feature augmentation strategy that explicitly models the domain-specific and domain-independent characteristics of the domains. In order to guide the adaptation, different semantic meta-graphs are exploited from biomedical KSs (such as SNOMED-CT and MeSH) and a set of new pivot features are created and incorporated into the text classifier. The feasibility of this approach is evaluated by comparing the proposed adaptive TC model against various strong transfer learning classifiers trained without semantic enrichment. In addition, a novel domain similarity measure is also proposed, which measures the similarity between domains based on the different zone clusters created by a probabilistic graphical model. This measure combines different lexical and semantic information present in these zone clusters, achieving superior results to previous *content-based similarity measures*.

Chapter 6 turns to the presentation of *supervised transfer learning approaches* for *whole-document* TC. In particular, this chapter proposes the use of social knowledge sources (such as DBpedia and Freebase) for building adaptive text classifiers of social media posts. The feasibility of this approach is evaluated by building several adaptive text classifiers, which make use of the data, knowledge and structure of these KSs, comparing their performance against text classifiers built on microposts data only. Firstly, these adaptive TC models are trained on the KSs data. Next, different semantic meta-graphs are exploited from these KSs for creating cross-domain features. For these cross-domain features then novel weightings are introduced, and different techniques proposed for incorporating them into the adaptive text classifiers. A detailed study on predicting the performance of a TC classifier is also conducted. Novel entropy-based measures are proposed, which make use of enhanced KS-based representation of the documents. These measures are also evaluated against state-of-the-art content-based lexical similarity measures.

1.5.3 Part III - Conclusions

Chapter 7 presents the conclusions drawn from this thesis. In particular it discusses how the requirements presented in Chapter 3 have been met, and how the techniques and approaches explored in the previous chapters have contributed to the claims presented in Chapter 1. In addition, possible future directions for adaptive text classification are discussed.

Background

Chapter 2

Background on Text Classification

2.1 Introduction

The multitude of documents present in *large databases*, including organisation archives and social media platforms, provide a heterogeneous environment comprising documents belonging to *multiple domains* and *text types*. Building TC systems in such environment poses challenges, as the lexical variation between domains can deteriorate the performance of the system. This thesis proposes novel TC techniques based on transfer learning, which provide automatic means for categorising and analysing the content of such large collections, accounting at the same for the lexical variations and differences between these documents. In particular, this thesis claims that the use of domain knowledge can be beneficial for reducing the lexical gap between domains, thus improving the performance of a TC classifier across domains.

This chapter presents the theoretical foundation of this thesis, providing background on TC and presenting a comprehensive literature review on transfer learning for different TC tasks. The structure of this chapter is as follows: Section 2.2 reviews the task of TC as defined in the literature, providing examples of different domains and applications to which TC has been applied. Section 2.3 highlights the main differences between in-domain machine learning approaches and transfer learning. Section 2.4 then presents different transfer learning settings and approaches proposed for TC. Following this, Section 2.5 presents a critical analysis of the limitations of current transfer learning approaches for TC. In Section 2.6, a discussion of the usefulness of domain knowledge for TC is given, emphasizing the need to incorporate this information into adaptive TC systems.

2.2 The Task of Text Classification

This section starts by formally describing the task of *text classification* (TC) (also called *text categorisation*, *document categorization*, *document classification* or *topic detection*), introducing the main notations and definitions commonly used in the machine learning [Mitchell, 1997; Theodoridis and Koutroumbas, 2009], and transfer learning literature [Pan and Yang, 2010]. To illustrate the notations, the task of topic classification is considered, which following the definition in Muñoz García et al. [2011], is the task of deciding whether a document belongs to a predefined set of *semantic categories* (or *classes* or *topics*) (e.g., Accident,

Disaster, etc.).

Given a set of documents $X = \{d_1, d_2, \ldots, d_m\}$ and a fixed set of topics $Y = \{t_1, t_2, \ldots, t_n\}$, the task of TC is thus to determine the topic of document $x = d_i$, $t(d_i) \in Y$, where $t(x) : X \rightarrow Y$ is a *classification function* (*target function* or *scoring function*), whose *domain* is $X \subseteq D$ and range (possible *categories*, *classes* or *labels*) is Y. From a probabilistic viewpoint, this means to assign a probability $p(y|x)$ to an instance x of belonging to y. The outcome of the classifier hence is a hypothesis $h : X \rightarrow Y$ which is the approximation of t [Mitchell, 1997]. The optimal classifier (h_{ERM}) is often obtained by employing the Empirical Risk Minimization (ERM) method [Vapnik, 1999], which aims to minimise the expected error (also called risk or loss) of the classifier over the test data[1]. Namely, given a loss function $l(x, y, \widehat{y}) : X \times Y \times Y \rightarrow \Re$ measuring the difference between the actual (y) and predicted $\widehat{y} = h(x)$ value of an instance x, the optimal classifier can be obtained as follows

$$h_{ERM} = \arg\min_{h \in H} l(x_i, y_i, h(x_i))$$

where H denotes the set of all possible classifiers. A typical example of a loss function is $l(x, y, h(x)) = (y - h(x))^2$, which is called least squares approximation [Mitchell, 1997]. Each observed document $x = d_i$ (also called *instance*, *example*, *observation* or *covariate*) then is described by a vector of *features* or *attributes* denoted by $d = (f_1, \ldots, f_k), f_j \in F$, where F denotes the corresponding feature space. In the running example, each instance x contains various *lexical* (e.g., the presence of a word within a dictionary), *semantic* (the presence of a named entity in the document, e.g., Barack Obama) and *syntactic* features (the part-of-speech tag of a word mentioned in the document) [Muñoz García et al., 2011].

From a probabilistic point of view, a *domain* D can also be described as a tuple $D = (F, P(X))$, having two components: the feature space (F), and a marginal probability distribution $P(X)$ [Pan and Yang, 2010]. Given a specific domain D, a *task* can also be defined as a tuple $T = (Y, t(\cdot))$, consisting of two components: the label space (Y), and the classification function ($t(\cdot)$).

Considering the multi-domain TC scenarios studied in this thesis, additional notations are provided for the domains used in the learning process. The *source domain* used to build a TC system is denoted by D_S, and the learning task in the source domain by T_S. Correspondingly, the *target domain*, on which the TC system is evaluated is denoted by D_T, and the learning task in the target domain by T_T. Furthermore, the two domains are considered *related* if there exists any relationship between the feature spaces of the two domains. Although transfer learning allows the leveraging of knowledge from multiple source domains, for simplicity in the following definition only a single source domain is considered from which to learn.

Definition 1 *Given a source domain D_S and learning task in the source domain T_S, a target domain D_T and a learning task in the target domain T_T, the objective of the* ***transfer learning*** *is to improve the learning performance of the classifier in D_T by leveraging the knowledge acquired in D_S, where $D_S \neq D_T$ or $T_S \neq T_T$ [Pan and Yang, 2010].*

The first condition $D_S \neq D_T$ includes the following different situations:

[1] It is also worth mentioning that in other cases one might be interested in quantifying the precision, recall, F-measure, or accuracy (measuring the proportion of instances correctly classified) over the test data.

1. the feature spaces are the same ($F_S = F_T$) and the marginal probability distributions differ ($P_S(X) \neq P_T(X)$) (e.g., considering documents written in different formats)

2. the feature spaces are different ($F_S \neq F_T$) and the marginal probability distributions differ ($P_S(X) \neq P_T(X)$)

3. the feature spaces are different ($F_S \neq F_T$) and the marginal probability distributions are the same ($P_S(X) = P_T(X)$) (*adaptation across different features spaces*)

It also worth mentioning, that there is a slight difference between the above definition and the one given in Pan and Yang [2010]. That is, in Pan and Yang [2010] all the three cases are considered as adaptation across domains, however, in this thesis this case is regarded as *adaptation across different features spaces*.

The second condition $T_S \neq T_T$ includes the following different situations:

1. the label spaces are the same $Y_S = Y_T$, and the conditional probability distributions differ ($P_S(Y|X) \neq P_T(Y|X)$)

2. the label spaces are different $Y_S \neq Y_T$, and the conditional probability distributions differ ($P_S(Y|X) \neq P_T(Y|X)$)

3. the label spaces are different $Y_S \neq Y_T$, and the conditional probability distributions are the same ($P_S(Y|X) = P_T(Y|X)$)

2.2.1 Applications of Text Classification

Having defined the main task of TC, this section now turns to the main applications of TC studied in the NLP community. In particular, this section provides a short summary of the most widely used TC tasks, which will be discussed in the context of transfer learning in Section 2.4. Among these tasks, two special instances of TC are also presented: an instance of *within-document* TC (called document zoning) and an instance of *whole-document* TC (called topic classification), which are both explored in this thesis.

An overview of the discussed tasks is presented in Table 2.1[2]. According to Sebastiani [2005], these tasks can broadly be classified according to the following dimensions:

1. *the unit of text*: including words, text fragments (such as sentences, paragraphs or phrases) and whole documents,

2. *the structure of the classification schema*: ranging from simple *flat structure* (such as the ones used in topic classification) to more sophisticated *hierarchical structures* (e.g., the ACM classification schema),

3. *the nature of the task:* such as *single-label* or *multi-label* TC, and

4. *the nature of the document*: ranging from traditional text type (such as newswire) to informal text genre (e.g., Twitter messages (tweets), cell phone messages (SMS)).

Focusing on the *unit of classification*, the smallest unit of classification, words, are used for instance in word sense disambiguation (WSD). In WSD the goal of the task is to detect the correct sense of an ambiguous word given its occurrence in the text and its meaning

[2]Further example tasks of TC in speech recognition fields can also be found in [Aggarwal and Zhai, 2012].

defined in a knowledge base (e.g., WordNet). For instance, considering the word bank, the aim is to detect whether bank refers to financial institution as in the sentence "he cashed a cheque at the bank" or a hydraulic engineering artifact as in the sentence "he was at the bank of river Thames".

In another TC task, document zoning (also called *within-document* TC), the goal of classification is to assign a semantic class to larger text fragments (called zones) such as phrases, sentences or paragraphs. For example, one of the most widely used applications of zoning is to recognise the different sections of scientific articles and thus classify sentences as belonging to one of the following following zone categories: Introduction, Method, Discussion, Results, etc.

Among the approaches taking the whole document as a classification unit (referred to as *whole-document* TC), text topic classification aims to detect the main topic(s) discussed in a document. For instance, a tweet discussing the impact of Twitter in the Egyptian revolution could be labelled with the following topics: "technology", "politics", and "war conflict".

	Problem	Unit of Text (x)	Semantic Classes
within-document TC	word sense disambiguation (WSD)	word	the word senses
	part-of-speech tagging (POS)	word	POS tags: {N, V, etc.}
	named entity recognition (NER)	word	entity classes: {PER, LOC, ORG, MISC, etc.}
	document zoning (DZ)	text fragment	zone types: {Introduction, Method, Discussion, etc.}
whole-document TC	spam filtering	document	{spam, not spam}
	language identification	document	languages:{en, ro, hu, etc.}
	sentiment mining	document	sentiments: {+,-}
	text topic classification	document	topics: {Disaster, Crime, etc.}

Table 2.1: Example applications of TC in Natural Language Processing. The TC tasks highlighted in italic correspond to the two main TC tasks explored in this thesis.

Regarding the *classification schema*, certain TC tasks have a well established schema associated to the TC task, such as a knowledge base in the case of WSD, or a flat list of POS tags as in POS tagging. For other TC tasks, however, the classification schema largely depends on the application domain at hand. For instance, for document zoning, there has been a large number of different document zone annotation schemas defined for different genre and scenarios. The next section provides a summary of these different schemas.

2.2.1.1 Classification Schemas for Intra-document Text Classification

This section describes different classification schemas proposed for *within-document* document zoning, aiming to capture the information structure of the documents. A summary of the schemas are also given in Table 2.2.

Applied to scientific literature, the main differences between these schemas are the spe-

	Abbreviation	Zone annotation schema	Unit of annotation	Length of the document	Number of classes
Sentence name based annotation schemas	OMRC	Object, Method, Result, Conclusion (Lin et al. (2006))	sentence	abstracts	4
	IMRAD	Introduction, Method, Result, and Discussion (Agarwal and Yu (2009))	sentence	abstracts and full length articles	4
Argumentative zoning	AZ	Argumentative Zoning I (Teufel and Moens (2002))	sentence	full journal articles	7
	AZ II	Argumentative Zoning II (Teufel et al. (2009))	sentence	full journal articles	15
Core Scientific schema	CoreSC	Core Scientific Schema (Liakata et al. (2010))	sentence	full journal articles	18
Meta-knowledge annotation schema	Meta-knowledge	Meta –knowledge annotation schemas (Nawaz et al. (2010))	phrases	full text articles	6
User centric schema		Focus, Polarity, Certainty, Evidence, Trend (Shatkay et al. (2008))	sentences	full text articles	5

Table 2.2: Overview of the existing *within-document* zone annotation schemas.

cific principle (theory or framework) they employ to classify the zones, and the type of granularity they consider, ranging from the most simple 4-way classification schema to the more fine-grained 7-way or even hierarchical classification ones.

The widely used classification schemas therefore are: *section name based classification*, *argumentative zoning (AZ)* aiming to encode the rhetorical structure of the documents, *Core Scientific Concepts (CoreSC)* concept-driven and ontological-driven classification schema and *meta-knowledge classification* schema.

The simplest classification schema is based on classifying the sentences to the most frequent section names which appear in scientific articles. One such schema, originally developed for biomedical abstracts, consists of a 4-way classification schema, aiming to categorise sentences into the *Objective*, *Method*, *Results* and *Conclusion* (OMRC) zone types [Lin et al., 2006; Hirohata et al., 2008]. The goal of these categories are as follows. The *Objective* zone aims to provide the background and describes the goal of the research; the *Method* zone describes the way in which this goal was achieved; the *Result* zone aims to summarise the findings of the research; and finally the *Conclusion* zone typically contains the analysis, discussion and main conclusion of the research. Another example of such schema consists of the *Introduction*, *Methods*, *Result*, and *Discussion* (IMRAD) zone types [Agarwal and Yu, 2009].

The Argumentative Zoning I [Teufel and Moens, 2002] schema aims to model the argumentative and rhetorical structure of the scientific articles. The main motivation behind this schema is that a scientific paper follows the knowledge claims (KC) of the authors. Therefore it aims to recover the rhetorical structure and the relevant stages in the argumentation, providing a 7-way classification schema capturing the paper's main KC in the AIM zone; the generally accepted background knowledge in the BACKGROUND zone; the description of existing KC in the OTHER zone; the description of the aspects of the new KC in the OWN zone; the comparison with the related work in the BASIS and CONTRAST zones; and finally a description of the structure of the paper in the TEXTUAL zone. An extension of this schema called Argumentative Zoning II [Teufel et al., 2009] was also proposed allowing a much fine-grained classification aiming to better capture the relationship between the paper's main KC and previous research.

The Core Scientific (CoreSC) annotation schema [Liakata et al., 2010], on the other hand, looks into a different aspect of the scientific writing, assuming that a paper is the human-readable representation of scientific investigation. Therefore it aims to identify the components of this investigation as expressed in the paper according to its 3-layer annotation schema. The first layer is ontology-motivated and aims to capture the core concepts in the scientific investigation, the second layer aims to capture the properties of the concepts (for e.g. "old", "new"), while the third layer aims to group instances of the same concepts together.

The meta-knowledge annotation schema [Nawaz et al., 2010] was designed for providing context to biomedical events. It aims to capture the rhetorical intent and level of certainty associated to a particular biomedical event providing a multi-dimensional annotation schema: the first dimension called *Knowledge type* aims to describe the general informal content of the event, the second *Certainty Level* aims to capture the confidence in the truth of the event, the third *Source* dimension aims to distinguish between new and previously reported knowledge, the fourth *Lexical Polarity* dimension aims to identify the events which are negated, the fifth *Manner* dimension aim to capture the manner of the event, and the

last *Local Type* indicates whether an event is propositional or not.

Furthermore, Shatkay et al. [2008] proposed a multi-dimensional classification schema aiming to model the different user needs thus allowing a more user-centric retrieval. The proposed classification schema thus aims to capture the characteristics of the statements in the literature among six dimensions: the *focus* dimension aims to capture the type of information conveyed by the statement (e.g. methodology), the *polarity* dimension indicates the polarity of the statement (positive or negative), other dimensions capture the *certainty* of the statement, and the type of the *evidence* supporting the statement, and finally the *trend* dimension indicates an increase or decrease in a specific phenomenon.

An important observation regarding these classification schemas is that although they were proposed separately, based on separate principles or views, an overlap or complementary relationship among the categories of different schemas can often be found. For e.g. Guo et al. [2011a] conducted a comparison evaluation concerning three annotation schemas: section name based, AZ, CoreSC in the context of cancer risk assessment and revealed a subsumption relationship among the categories of these schemas.

2.2.2 Related Natural Language Processing Tasks

There exists a wide range of applications of TC employing different unit of classification and classification schemas.

Further it is also worth noting, that certain applications of TC share some similarities to other NLP tasks, such as *text segmentation* and *discourse processing*. For the sake of completeness, in what follows, the main similarities and differences between these tasks and TC are discussed.

Text segmentation is often regarded as a pre-processing step in many NLP tasks. It involves the partitioning of the text into distinct textual units, such as words, sentences or topics. In addition to previously discussed applications, text segmentation is topic segmentation [Hearst, 1997], in which the goal is to segment the text into coherent topics. In this approach, the text is split into pseudo-sentences containing words of fixed size, and the shift in topics is detected by measuring the similarity between the different pseudo-sentences. As a result, consecutive segments with a high score will be considered topically coherent. The main difference to TC is thus that the extracted textual segments do not have a semantic class associated to them.

Discourse processing [Stede, 2011], on the other hand, attempts to divide the text into meaningful units that are related to one another through discourse relations. One of the most notable discourse theory is the Rhetorical Structure Theory (RST) [Mann and Thompson, 1988], which assumes the existence of a hierarchical structure within the text, providing an explanation of its coherence. Examples of discourse relations defined in RST[3] including Background, Evidence, Elaboration, Condition, Interpretation, Summary, etc. These relations are typically defined over small discourse units, such as closes, ranging from minimally a noun phrase to maximally a sentence. The goal of discourse processing is then to automatically recognise textual units and assign the corresponding rhetorical class to them. In this respect, discourse processing is closely related to document zoning, following the schema of Argumentative Zoning by Teufel and Moens [2002]. In contrast, however,

[3] http://www.sfu.ca/rst/01intro/definitions.html

the AZ schema, as well as most document zoning schemas is not hierarchical, is less fine-grained, and as opposed to the local-RST relationships it models the rhetorical moves from a more global perspective, examining longer discourse units. For instance, considering the sentence "Unfortunately this work does not solve the problem X". This sentence may refer to the shortcomings of previous research described in the related work of the paper, and thus would be labelled as CONTRAST zone; or if mentioned in the future work, than it may describe the weaknesses of the current paper, and would be labelled OWN [Teufel and Moens, 2002].

2.3 Machine Learning Approaches and Main Differences to Transfer Learning

Before moving to the presentation of transfer learning approaches for the various TC tasks discussed, this section provides a short summary of the in-domain machine learning approaches and highlights their differences to transfer learning.

In-domain machine learning approaches are the state-of-the-art solutions for most TC systems. They can be classified into three categories: supervised, unsupervised and semi-supervised approaches depending on the available resources.

Supervised machine learning algorithms rely on large amount of labelled in-domain data, which are often limited or labour-intensive to build in many domains. This major bottleneck of insufficient in-domain labelled data is addressed by the unsupervised and semi-supervised methods, which assume that a large amount of unlabelled in-domain data is cheap to obtain.

The unsupervised methods are based on clustering algorithms that automatically partitions data into groups, so that data in the same groups are relatively similar, while data in different groups are relatively dissimilar.

The semi-supervised (SSL), also called weakly supervised or bootstrapping techniques try to learn from a limited set of labelled in-domain examples (labelled seeds) and a large amount of unlabelled in-domain examples. There are various types of semi-supervised methods, such as self-training [Zhu, 2006], co-training [Blum and Mitchell, 1998], and active learning [Settles, 2010].

In the self-training setting, a classifier is incrementally learnt based on the labelled in-domain seed set and a set of unlabelled in-domain examples that are labelled with the current classifier until the trained classifier reaches a certain level of accuracy on the test set.

In co-training two or more classifiers are trained using the same seed set of labelled in-domain examples, but each classifier trains with an independent set of features. At each iteration the classifiers label few unlabelled examples. The examples that are then labelled with the current classifiers and the ones on which the classifiers agree with most confidence are added to the pool of labelled examples. The classifiers are then retrained and the process iterates until it reaches a certain level of accuracy on the test set.

In active learning all the in-domain examples are labelled by humans, but the limited number of examples to be labelled are carefully selected by the classifier. The key hypothesis is that if a classifier is allowed to choose the data from which it learns, it will perform better with less training.

As described in the previous subsection, transfer learning allows the domains, tasks and

distributions used in training and testing to be different. However, if the source and target domains are the same ($D_S = D_T$), and their learning tasks are the same ($T_S = T_T$), one is facing the in-domain machine learning problem [Pan and Yang, 2010]. Further, there is also a small difference between transfer learning (domain adaptation) and semi-supervised learning. SSL tries to learn a good classifier from a *small* amount of labelled data, while in transfer learning the labelled data is *large* [Jiang, 2008a].

The next sections presents the various transfer learning settings and approaches proposed in the literature, focusing mostly on the employed techniques. The description of base classifier used in the the transfer learning scenarios is however outside the scope of this thesis[4].

2.4 Transfer Learning Approaches for Text Classification

This section discusses the main approaches to TC that fall under the *transfer learning* paradigm. It provides a more general perspective on TC, including coverage of a number of *transfer learning* scenarios and important *transfer learning* techniques.

There have been several transfer learning sub-settings studied in the literature under different names (see Figure 2.1): such as *Domain adaptation* setting (Subsection 2.4.1) where the learning tasks are the same, *Multi-task learning* setting (Subsubsection 2.4.1.4), where the source and the target tasks are different but related and also learnt simultaneously, and *Unsupervised Transfer learning* setting (Subsection 2.4.2), where there are no labelled source and target data available. A combination of transfer learning with active learning, called *Active Transfer learning* (Subsection 2.4.3), has also been recently proposed.

The existing approaches to transfer learning fall into the following categories: *instance-based*, *feature-representation based*, *parameter-based* and *relational-knowledge transfer approaches*. All these separate approaches aim to identify the relevant knowledge from the source domain which can be beneficial for learning in the target domain, thus addressing the first research issue in transfer learning, that of "*what to transfer*".

Instance-based approaches [Shimodaira, 2000; Zadrozny, 2004; Jiang and Zhai, 2007a] assume that there are certain instances in the source domain which are relevant for the target domain. Therefore, these approaches aim to re-weighting these source domain instances in order to maintain the same distribution in the source and target domains.

The feature-representation approaches [Ando and Zhang, 2005; Blitzer et al., 2006; Satpal and Sarawagi, 2007] assume that there is a common feature representation under which the two domains are more similar. In contrast to instance-based methods, these approaches can be more effective in situations when a few features cause the domains to differ. For example, if one has a feature called "Is capitalized word" in the source domain, while in the target domain none of the names are capitalised.

Parameter-based approaches [Chelba and Acero, 2004; Ciaramita and Chapelle, 2010] assume that the source and target domains share some common parameters or prior distributions. These parameters aim to encode the prior knowledge acquired in the source domain

[4]The reader is referred to [Mitchell, 1997] or [Theodoridis and Koutroumbas, 2009] for more details about the base classifiers.

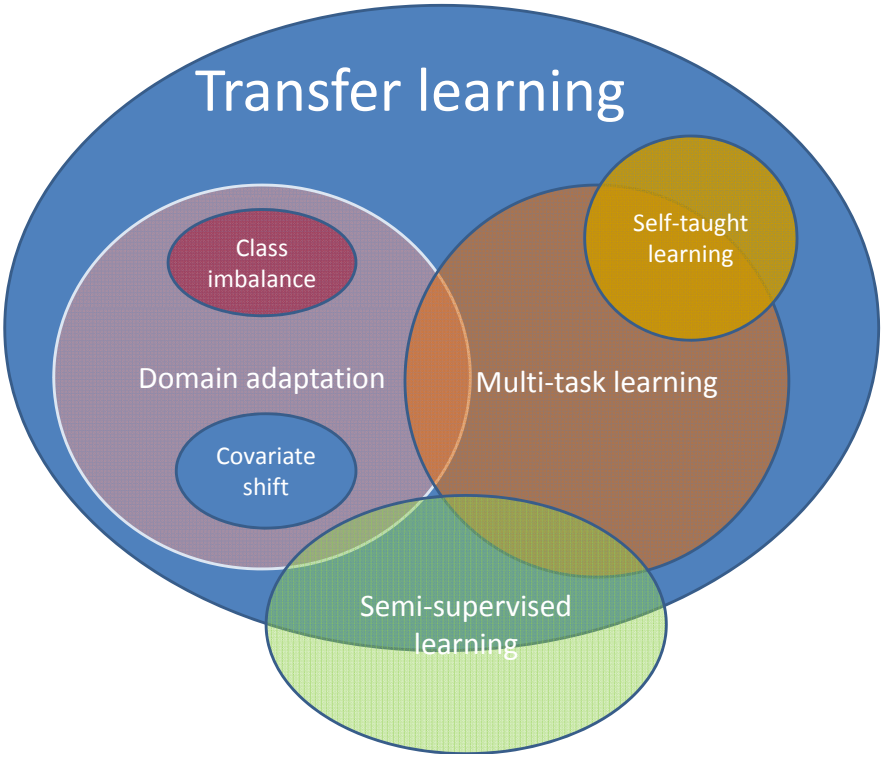

Figure 2.1: Comparison of transfer learning settings.

for improving the learning in the target domain.

Finally, the *relational-knowledge transfer approaches* [Mihalkova, 2009] assume that certain relationships between the instances in the two domains are similar. Therefore these approaches aim at finding a mapping between these relations in the source and target domains.

After the relevant knowledge from the source domain has been identified, the next research issue to consider is "*how to transfer*", which asks how to develop learning algorithms for transferring this knowledge to the target domain.

In the next sections these algorithms are reviewed, which is followed by a discussion about the third main research issue of *"when to transfer"* in Subsection 2.4.4.

2.4.1 Domain Adaptation

The first domain adaptation setting [Jiang, 2008b; Blitzer, 2008] refers to the case commonly present in NLP, where the source and the target tasks are the same, while the source and the target domains are different.

Following the definition of transfer learning, the objective of **domain adaptation** is to improve the learning performance of the classifier in D_T by leveraging the knowledge acquired in D_S, where the domains are different but related ($D_S \neq D_T$) *and* the tasks in the two domains are the same ($T_S = T_T$) [Pan and Yang, 2010]. In addition, a *large amount*

Transfer learning technique	Method name	Labelled/Unlabelled Data				Classifiers			Similarity measure	Background knowledge	Solved task	Corpora	Performance measure
		L_S	U_S	L_T	U_T	Base classifier	Loss	Ensemble method					
Instance-based	Huang et al. (2007)	✓			✓	LMS regression/ SVM with Gaussian kernel					regression and classification	UCI Breast Cancer dataset	error rate
	Bickel et al. (2007)	✓			✓	Logistic Regression with linear kernel & RBF kernel					spam filtering, text classification, landmine detection	Enron and Usenet; Cora dataset,	1-AUC
	Sugiyama et al. (2007), KLIEP	✓			✓	Logistic Regression with Gaussian kernel			KL divergence		regression and classification	UCI auto-mpg dataset	normalized mean square error
	Dai et al. (2007a)	✓			✓	Naive Bayes with EM			KL divergence		text classification	20 Newsgroups, SRAA, Reuters-21578	error rate
	Jiang and Zhai (2007a)	✓		✓	✓	Logistic Regression	ERM				POS tagging, entity type classification, spam filtering	WSJ-;ACE 2005; ECML/PKDD 2006 discovery challenge dataset; PennBioIE corpus	accuracy
	Wu et al. (2009a)	✓			✓	SVM					entity mention detection	ACE 2005	F1
	Ponomareva and Thelwall (2012b)	✓			✓	RANK, OPTIM			SentiWordNet, SO-CAL-dictionaries		sentiment classification	Amazon product review	F1

Table 2.3: Overview of TC models proposed for the *Domain Adaptation* setting. Columns correspond to method proposed (Method name), transfer learning approach applied (Instance-based, Feature-representation based, Parameter-based), resources used during learning (labelled source (L_S) and target (L_T) data, unlabelled source (U_S) and target (U_T) data), classifier employed in transfer learning (Base classifier, Loss function, Ensemble method), similarity measure used to encode the divergence between the domains (Similarity measure), additional background knowledge used to enhance the learning (Background knowledge), the particular NLP problem addressed (Solved task), and the measure used to evaluate the performance of the classifier (Performance measure). **Abbreviations used**: LMS regression for Least Mean Square regression, RBF kernel for Radial Basis Function kernel, MIRA is an online large-margin linear classifier described in Crammer et al. [2006].RANK is a ranking algorithm proposed in Wu et al. [2009b], OPTIM is an optimisation algorithm proposed in Goldberg and Zhu [2006].

Transfer learning technique	Method name	Labelled/Unlabelled Data				Classifiers			Similarity measure	Background knowledge	Solved task	Corpora	Performance measure
		L_S	U_S	L_T	U_T	Base classifier	Loss	Ensemble method					
Feature-representation based	Ando (2004)	✓	✓		✓	Robust Risk Minimization					entity detection	WSJ, CNS corpus, ACE 2001-2002	F1
	Ando and Zhang (2005)	✓	✓		✓	Robust Risk Minimization	Huber loss				syntactic chunking, named entity chunking	CoNLL2003-CoNLL 2000, WSJ, TREC	F1
	Blitzer et al. (2006), SCL	✓	✓		✓	large-margin linear classifier (MIRA)	Huber loss				POS tagging	WSJ, MEDLINE	accuracy
	Blitzer et al. (2007b), SCL-MI	✓	✓		✓	linear classifier	Huber loss				sentiment classification	Amazon product review	accuracy
	Blitzer et al. (2011)	✓	✓		✓	CRF for POS tagging	Square loss				regression for the stars of product review; POS tagging	Amazon product review; WSJ to Biomedical abstracts from Medline	squared error rate
	Arnold and Cohen (2008)	✓			✓	CRF					extracting protein names	Genia abstracts, Pubmed Central full article	F1
	Guo et al. (2009)	✓	✓		✓	RRM					NER	Wikipedia articles	F-measure
	Nallapati et al. (2010) NER-LDA	✓	✓		✓	CRF				unlabelled data NYT	NER	AC E 2005	F1
	Kadar and Iria (2011) TransferLDA	✓	✓		✓	SVM			KL divergence		document classification	20 newsgroup, SRAA corpus	accuracy

Table 2.4: Overview of TC models proposed for the *Domain Adaptation* setting. Columns correspond to method proposed (Method name), transfer learning approach applied (Instance-based, Feature-representation based, Parameter-based), resources used during learning (labelled source (L_S) and target (L_T) data, unlabelled source (U_S) and target (U_T) data), classifier employed in transfer learning (Base classifier, Loss function, Ensemble method), similarity measure used to encode the divergence between the domains (Similarity measure), additional background knowledge used to enhance the learning (Background knowledge), the particular NLP problem addressed (Solved task), and the measure used to evaluate the performance of the classifier (Performance measure).

Transfer learning techniques	Method name	Labelled/Unlabelled Data				Classifiers			Similarity measure	Background knowledge	Solved task	Corpora	Performance measure
		L_S	U_S	L_T	U_T	Base classifier	Loss	Ensemble method					
Feature-representation based	Satpal and Sarawagi (2007)	✓			✓	CRF					NER	CoNLL 2003, Cora citations, Cora headers, Citeseer citations	F1
	Gupta and Sarawagi (2008)	✓			✓	Markov Random Fields					bibliographic IE	Bibliographic entries from the web	F1
	Ciaramita and Yasemin (2005)	✓			✓	Semi Markov Model with Average Perceptron				SEMCOR	NER	WSJ, CoNLL 2003	F1, error rate
	Mika et al. (2008)	✓			✓	HMM with Average Perceptron				Wikipedia, DBPedia	NER	Wikipedia articles	Accuracy
	Wang et al. (2008)	✓			✓	Co-clustering				Wikipedia	text classification	20 Newsgroups, SRAA	F1
	Xiang et al. (2008)	✓			✓	TSVM				Wikipedia	text classification, sentiment classification	20 Newsgroups, Sentiment Reviews	Accuracy
Parameter-based	Ciaramita and Chapelle (2010)	✓			✓	HMM with Average Perceptron				Gazetter, WordNet	NER	BBN, CoNLL 2003	F1

Table 2.5: Overview of TC models proposed for the *Domain Adaptation* setting. Columns correspond to method proposed (Method name), transfer learning approach applied (Instance-based, Feature-representation based, Parameter-based), resources used during learning (labelled source (L_S) and target (L_T) data, unlabelled source (U_S) and target (U_T) data), classifier employed in transfer learning (Base classifier, Loss function, Ensemble method), similarity measure used to encode the divergence between the domains (Similarity measure), additional background knowledge used to enhance the learning (Background knowledge), the particular NLP problem addressed (Solved task), and the measure used to evaluate the performance of the classifier (Performance measure).

Transfer learning technique	Method name	Labelled/Unlabelled Data				Classifiers			Similarity measure	Background knowledge	Solved task	Corpora	Performance measure
		L_S	U_S	L_T	U_T	Base classifier	Loss	Ensemble method					
Feature-representation based	Jiang and Zhai (2007b)	✓		✓	✓	Logistic Regression	ERM				gene recognition, protein name recognition	BioCreative I challenge, task 1b	precision, recall, F1
	Daume III (2007) EA	✓		✓	✓	Average Perceptron	Hinge loss				NER, shallow parsing, POS tagging	ACE 2005, ACE 2006, CoNLL 2003, PubMed, CNN, Penn Treebank	error rate
	Daume III et al. (2010), EA++	✓		✓	✓	SVM					POS tagging, NER	PubMed,	error rate
	Daume III and Jagarlamudi (2011)	✓		✓	✓						out-of-vocabulary	language pairs: Ger to En, and Fr to En; domains: newswire, medical, movie subtitles, technical documentation	BLUE score
	Dai et al. (2007b), TrAdaBoost	✓		✓	✓	SVM with linear kernel		boosting			text classification	SRAA,Reuters-21578, 20 Newsgroups, UCI mushroom dataset	error rate
Parameter-based	Al-Stouhi and Reddy (2011) Dynamic-TrAdaBoost	✓		✓	✓	TrAdaBoost		boosting			text classification	20 Newsgroups, Abalone, Wine	Accuracy
	Chelba and Acero (2004)	✓		✓	✓	Maximum Entropy					capitalization	CNN, ABC, WSJ	accuracy, error
	Arnold et al. (2008)	✓		✓	✓	CRF					NER	Genia, PubMed	F1
	Finkel and Manning (2009)	✓		✓	✓	CRF					NER, dependency parsing	CoNLL, MUC 6, MUC 7, OntoNotes release 2.0	F1

Table 2.6: Overview of TC models proposed for the *Domain Adaptation* setting. Columns correspond to method proposed (Method name), transfer learning approach applied (Instance-based, Feature-representation based, Parameter-based), resources used during learning (labelled source (L_S) and target (L_T) data, unlabelled source (U_S) and target (U_T) data), classifier employed in transfer learning (Base classifier, Loss function, Ensemble method), similarity measure used to encode the divergence between the domains (Similarity measure), additional background knowledge used to enhance the learning (Background knowledge), the particular NLP problem addressed (Solved task), and the measure used to evaluate the performance of the classifier (Performance measure).

of labelled source data ($L_S >> 0$) is available at training time. When there is no labelled target data ($L_T = 0$) available the problem is generally referred to as *unsupervised domain adaptation* in the literature, while when a *small* amount of labelled target data ($L_T > 0$) is available, the problem is generally referred to as *supervised domain adaptation* [Jiang, 2008a].

2.4.1.1 Unsupervised Domain Adaptation Approaches

In what follows, instance-based, feature-representation based and parameter based approaches successfully applied to unsupervised domain adaptation are presented. A comparison of the methods reviewed is also given in Table 2.3.

Instance-based Approaches

The two major techniques for instance-based methods are *re-sampling* and *instance-weighting* methods. Re-sampling methods aim at re-sampling the source domain instances so that the re-sampled data roughly has the same class distribution as the target data. In contrast, instance-weighting methods differently weight the source and target domain instances in order to maintain the same distributions.

The main motivation behind instance-weighting method is to employ ERM to learn the optimal classifier for the target domain. But, as in domain adaptation the distributions in the source and target domains are different ($P_S(X,Y) \neq P_T(X,Y)$), ERM is not generally consistent anymore according to Shimodaira [2000].

However, as it was pointed out in [Jiang, 2008a; Shimodaira, 2000], the following importance weighting is consistent:

$$h_{ERM} \approx \arg\min_{h \in H} \sum_{i=1}^{n_S} \frac{P_T(x_i, y_i)}{P_S(x_i, y_i)} l(x_i, y_i, h(x_i))$$

Therefore the optimal classifier can be learned by weighting the loss function with $\alpha = \frac{P_T(x,y)}{P_S(x,y)}$ [Jiang, 2008a]. However, there are cases when we don't have any labelled target domain instances, so we are not able to compute the exact value of α.

In order to address this problem, several special cases of domain adaptation problem have been studied: including *class imbalance* [Japkowicz and Stephen, 2002], *covariate shift* [Shimodaira, 2000] and *sample selection bias* [Zadrozny, 2004].

In the *class imbalance problem*, although the distribution of the two domains differ ($P_S(X,Y) \neq P_T(X,Y)$), it is assumed that $P_S(X|Y=y) = P_T(X|Y=y)$ for all $y \in Y$, but $P_S(Y) \neq P_T(Y)$. In this case we only need to weight the instances with $\alpha = \frac{P_T(y)}{P_S(y)}$.

The effect of class imbalance problem on various classifiers, including decision trees (C5.0), neural networks (multi-layer perceptron), and SVM (hard margin SVM), was studied in Japkowicz and Stephen [2002]. They showed that the effect of the class imbalance problem on these classifiers is different and it is influenced by the number of the training examples, the degree of the class imbalance, and the complexity of the target function. Their experiments reveal that the most sensitive classifier is the decision tree classifier, followed by the multi-layer perceptron and finally SVM was shown to be unaffected by the problem.

In the *covariate shift* and *sample selection bias* settings, although the distribution of the two domains differ ($P_S(X,Y) \neq P_T(X,Y)$), it is assumed that $P_S(Y|X=x) = P_T(Y|X=$

x) for all $x \in X$, but $P_S(X) \neq P_T(X)$. Sample selection bias is a special case of covariate shift, where each training instance is sampled from the test distribution according to a boolean selector variable s. When the value of the s is 1, the instance is moved to the training set, otherwise when the value is 0, the instance is moved to the test set.

Under these settings the optimal classifier for the target domain can be learned by weighting each training instance with $\beta = \frac{P_T(x)}{P_S(x)}$. However, estimating the distribution through the examples is difficult, as in general we have high dimensional feature spaces. Various methods to estimate β have been proposed.

Shimodaira [2000] proposed to re-weight the log likelihood of each source domain example with β to minimize the loss of the classifier on the target domain data. This was theoretically shown to lead to the optimal model for the target domain. However, this approach is *not practical* because we don't have a probability distribution over the examples in each domain.

Similarly, Zadrozny [2004] proposed to use a selection ratio as a weight for each source domain example to correct the distributional difference between the domains. She analytically studied the effect of sample selection bias on several classifiers, including *local learners* (Bayesian classifier, logistic regression, hard margin SVM [Joachims, 2000]) and *global learners* (Naive bayes, decision trees, and soft margin SVM [Schölkopf and Smola, 2001]). Their experiments show that global learners are affected by sample selection bias, while local learners are not.

As mentioned above, estimating probability distributions though examples is difficult, especially in high dimensional spaces, therefore, a number of methods have been proposed to estimate the probability ratio *directly*, without first estimating the probability of instances in each of the domains [Huang et al., 2007; Bickel et al., 2007].

Huang et al. [2007] proposed to directly estimate the probability ratio using a non-parametric kernel mean matching (KMM) method. This way, the mean of the weighted source domain instances and target domain instances become close in a reproduced Hilbert space [Dinuzzo and Schölkopf, 2012]. Experimental results on regression and classification problems show that this method outperforms the un-weighted method, and match or exceed the performance of the method proposed in Zadrozny [2004]. Similarly, Bickel et al. [2007] proposed to directly estimate β with the classification model parameters deriving a kernel-logistic regression classifier. This classifier estimates the probability that an instance is from the target domain as against the source domain. In training this classifier, the instances in source domain are treated as negative examples, while instances in target domain are treated as positive examples. Experimental results on classification problems showed that the proposed method together with the kernel mean matching [Huang et al., 2007] and logistic regression classifier perform well.

In addition to sample re-weighting methods, there has been work extending semi-supervised learning for domain adaptation [Dai et al., 2007a; Jiang and Zhai, 2007a; Wu et al., 2009a; Ponomareva and Thelwall, 2012b].

Dai et al. [2007a] proposed an extension of the traditional EM-based Naive Bayes [Nigam et al., 2000a] classifier for domain adaptation. They first set the initial probability parameters using the source domain data, and next they revise them using the target domain data. The KL-divergence measure was used to measure the distributional difference between the two domains and to estimate the trade-off parameters for EM. Empirical results on text classification show that the proposed method outperforms SVM and Naive bayes on binary

text classification problems.

Jiang and Zhai [2007a] proposed an instance weighting framework which minimizes the empirical loss of the classifier not only on the weighted labelled target domain instances but also considering the weighted source domain instances and unlabelled target domain instances. They introduced some weighting parameters and gave several heuristic methods to set them using semi-supervised methods. Therefore, it might be an interesting research direction to study how to estimate the parameters more accurately.

Similarly, Wu et al. [2009a] proposed a bootstrapping approach for domain adaptation which aims to identify the bridge instances from the target domain, which are instances that contain both domain specific and domain independent examples. First they train a source domain classifier on the labelled source domain instances and use this classifier to label the target domain instances. Next they build a target domain classifier on these labelled target domain instances. Then, at each iteration the algorithm selects the most informative examples such that they are classified more confidently by the target classifier and adds them to the source classifier. This procedure repeats until there are no more bridge instances in the target domain, or when the source classifier is more confident about labelling the instances than the target classifier. Experimental results on named entity recognition on the ACE corpora show that this approach outperforms standard bootstrapping and the balanced bootstrapping proposed in Jiang and Zhai [2007a].

Ponomareva and Thelwall [2012b] studied two different graph-based approaches for cross-domain sentiment analysis. The first approach, named RANK [Wu et al., 2009b], uses ranking to assign sentiment scores to the target domain documents, while the second approach, named OPTIM [Goldberg and Zhu, 2006], solves an optimisation problem for labelling documents with sentiments labels. In both cases, the graph is built between the labelled instances of the source domain and unlabelled instances of the target domain, and the weights between the edges are computed according to various document similarity measures. These similarity measures fall into two categories: feature based similarity, consisting of uni-grams and bi-grams weighted by inverse document frequency; and lexicon based similarity employing different sentiment resources (e.g. SentiWordNet and SO-CAL dictionaries) for assigning scores to words and sentences (for e.g. number of positive words versus. number of negative words). Experimental results on benchmark sentiment dataset show promising results, RANK with the combined feature and lexicon based similarity measure consistently outperforming Structural Correspondence Learning (SCL) [Blitzer et al., 2006].

Feature-representation Approaches

The feature-representation based approaches to domain adaptation rely on finding a good feature-representation by discovering features that have similar distribution across domains.

In some works *additional new features* are added to the original feature spaces of the two domains using unlabelled data [Ando, 2004; Ando and Zhang, 2005; Blitzer et al., 2006; 2007b; Blitzer, 2008; Blitzer et al., 2011; Guo et al., 2009; Nallapati et al., 2010; Kadar and Iria, 2011; Arnold and Cohen, 2008], while in other works a *subset of features* are considered [Satpal and Sarawagi, 2007].

Blitzer et al. [2006] proposed Structural Correspondence Learning (SCL) which extends the semi-supervised Alternative Structural Optimization (ASO) from [Ando and Zhang, 2005], which is based on canonical correlation analysis. SCL focuses on finding a common

representation for features across domains using unlabelled source and target instances. Firstly it selects the so-called "pivot features" that occur frequently in the unlabelled data of both domains. Next, linear predictors for those features are learned on all the other features. And finally singular value decomposition (SVD) is performed on the collection of learned linear predictors corresponding to different pivot features. At training stage the original feature spaces of both domains are augmented with these new features. Experimental results on POS tagging show that SCL consistently outperforms both supervised [Ratnaparkhi, 1996] and semi-supervised learning (ASO) methods. SCL was also tested on a sentiment-classification task [Blitzer et al., 2007b]. They showed that by choosing pivot features with high Mutual Information with the source labelled instances can further reduce error on the target domain.

More recently, Blitzer et al. [2011] proposed a coupled learning algorithm which aims to create two domain-specific sub-spaces from the low dimensional shared space created by SCL between the domains. Experimental results on sentiment analysis and POS tagging show, that when the shared space is small this coupled learning algorithm outperforms SCL. On the other hand, when the shared space is large and the coupled space misses part of it, then SCL outperforms the coupled learning algorithm. Although, Blitzer et al. [2007a] showed that the representation created by SCL decreases the distance between the distributions in the two domains, the selection of pivot features is still domain dependent.

Ando [2004] proposed another SCL-like approach to derive new features which are more stable across domains. However, they performed SVD on the feature-token matrix, where the matrix contains all the sentences from the unlabelled source and unlabelled target domain instances.

Arnold and Cohen [2008] proposed to use additional intra-document structural features for the problem of extracting protein names across the different sections of biomedical journal articles, considering the abstracts of the articles as the source domain and the captions and full text of the articles as the target domains. These new features therefore can be created using only unlabelled data, by considering the frequency of each word in the separate sections and the conditional frequency of each word across the sections (for e.g. the probability of the word appearing in the caption of the article, given that it appeared in the abstract), thus allowing to explicitly model the differences between the sections. In addition they proposed to augment the original training set of the source domain with positive and negative examples, called snippets, gathered from the target domain. Experimental results on Genia abstracts and Pubmed Central full articles show promising results outperforming the baseline classifier trained only with lexical features.

Another approach for generating additional features from unlabelled data rely on applying *unsupervised Latent Dirichlet allocation topic model* [Guo et al., 2009; Nallapati et al., 2010; Kadar and Iria, 2011].

Guo et al. [2009] employed Latent Dirichlet Allocation (LDA) topic model to learn the topic assignments among the unlabelled examples in the source and target domains, resulting in new features for adaptation. These learned features are then used to augment the labelled source domain examples and build a supervised classifier to make predictions on the unlabelled target domain examples. Experimental results on NER show that the proposed model significantly outperform the base classifier built without semantic association features.

Similarly, Nallapati et al. [2010] proposed to augment the labelled source domain instances with LDA features, created by another LDA model, called NER-LDA, which models the words as being conditioned over the named entity (NE) labels and topics. Furthermore this NER-LDA model also allows different distributions of NE labels for different topics. The new LDA features therefore consist of the topic id's assigned to each word and to the surrounding words together with their corresponding NE labels. Moreover, the generation of the topics can be considered using additional unlabelled data from New York Times. Experimental results on NER from six different domains from the ACE 2005 corpora show promising results for the NER-LDA model, outperforming the basic CRF classifier trained without these features.

Instead of requiring labelled instances from the source domain, Kadar and Iria [2011] proposed TransferLDA and TransferzLDALF models which both employ a new feature labelling paradigm [Druck et al., 2008a], thus further reducing the annotation time and costs. The first TransferLDA model aims to minimize the distribution gap between the domains by measuring the KL-divergence. While the second TransferzLDALF model further employs an zLDA model to first cluster the features from both domains and then augment the original labelled features from the source domain with the most probable labelled features from the target domain in each cluster. Experimental results on document classification problems using 20 news group and SRAA corpus show that using only a few labelled features, on average 18 labelled features per topic, TransferLDA consistently outperforms the baseline supervised Maximum Entropy classifiers, while TransferzLDA performs comparable with the supervised Transductive SVM (TSVM) classifier.

In contrast to the above approaches, Satpal and Sarawagi [2007] proposed to select a subset of features by assigning a weight to each feature that is equal to the difference in the expected value of the feature in the two domains. Experimental results on entity extraction show significant accuracy gain with varying training and test size, outperforming SCL and SSL.

Gupta and Sarawagi [2008] proposed to use domain independent properties for a bibliographic information extraction task. For example, a property might be 'Does the title appear after the author' in a bibliographic record. The regularities between domains thus can be captured by using the classifier trained on the source domain to jointly annotate all the records in the target domain such that the output labels are regular with respect to the affected set of properties. One drawback of this approach is that the properties are still application specific.

In other works the usage of *background knowledge* has been explored [Ciaramita and Yasemin, 2005; Mika et al., 2008; Wang et al., 2008; Xiang et al., 2010].

Ciaramita and Yasemin [2005] proposed a Semi-Markov model (SMM) trained with average perceptron algorithm for the task of NER. They used an external domain-independent dictionary (called SEMCOR) to improve the generalisation across domains. Namely, additional dictionary features about the extracted entity segments (for e.g."George Duffield") were added to the model. One such feature is the Jacard distance between the dictionary entries and the extracted entity. The length of the entity ($length$("George Duffield") = 2) is considered a separate feature as well. The comparative results show that the SMM model supported by the external dictionary outperforms HMM [Collins, 2002] without external knowledge. It would be therefore interesting to further explore the impact of different dic-

tionaries or other external sources, such as ontologies.

As the semantic annotation for many IE tasks, including NER is limited to few public data sets, and acquiring training data is often expensive, Mika et al. [2008] explored the possibility to generate additional training data using Wikipedia and DBPedia to improve a NER tagger. In their scenario, a NER tagger trained on CoNLL corpora is applied to a collection of Wikipedia articles in order to enrich the metadata information present in the infoboxes. This step also involves mapping the vocabulary of CoNLL tags to the more fine-grained vocabulary of Wikipedia's infobox properties. For instance, in an article describing Pablo Picasso, the place of birth (Spain) annotated with the LOCATION CoNLL tag is mapped to placeOfBirth infobox property. Next these semantic annotations are linked with the structured knowledge from the DBpedia, which is a lightweight ontology consisting of extracted information from Wikipedia infoboxes. They then apply this mapping to the corpus and generate additional training instances for the tagger. One limitation of this approach is that the distribution of entities can be skewed, because the set of training sentences are not a random sample of Wikipedia text.

Wang et al. [2008] proposed a co-clustering based approach, which exploits additionally information from Wikipedia. The approach functions by first extracting the Wikipedia concepts inside the domain documents from both source and target domains. After that a proximity matrix is defined, which captures the closeness between any two terms, according to several relationships defined in Wikipedia: synonymy, hyponymy and associative concepts relationships. This matrix can thus be used to project instances of both domains into a semantic space where the instances are closer to one another. Following this the original lexical feature spaces of both domains are extended with the projected semantic features, and a co-clustering of the instances of both domains are performed according to the approach presented in Dai et al. [2007a]. Experimental results on both Newsgroups and SRAA datasets showed promising results outperforming its counterpart model without semantic enrichment.

Xiang et al. [2010] proposed a novel approach for domain adaptation, which employs semi-supervised learning and knowledge from Wikipedia to guide the adaptation between domains. In the first step of this approach, candidate documents are retrieved from Wikipedia, based on a similarity (relatedness) measure computed between these knowledge source documents and both source and target domain documents. In order to achieve this the LDA topic model is run over the documents of the source domain, the documents of the target domain, and the Wikipedia articles separately. Then the similarity between the domain documents and knowledge source articles are computed by summing over each document and taking the product of the document topic probabilities. As a result a set of related domain documents are created, which are used to iteratively train a Transductive SVM (TSVM) classifier in a semi-supervised fashion. Experimental results on 20 Newsgroup data and Sentiment Reviews show promising results, outperforming various baseline transfer learning approaches such co-clustering based adaptation approach [Dai et al., 2007a] or Transductive SVM [Joachims, 1999].

Parameter-based Approaches

The parameter-based approaches to domain adaptation assume that the source and target domains share some common parameters (e.g., class prior distributions) [Ciaramita and Chapelle, 2010], which can be beneficial for transfer learning.

Ciaramita and Chapelle [2010] proposed an adaptive extension of the perceptron trained Hidden Markov model (HMM) for the task of NER, where the adaptive components are parameters estimated from the unlabelled source and target domain data combined with background knowledge. Assuming that the distribution of source and target domains differ partly due to difference in the *class prior distributions*, at decoding time they adjust the class labels by considering the estimated class frequencies in the source and target domains. The estimation of class frequency is done by using a list of words from an independent third source (gazetteers from GATE[5]), and computing the frequency of each word in the corresponding domain. The gazetters and WordNet are also used to create additional features for the model, such as membership in a gazetter or capturing the most frequent supersense in WordNet. Experimental results on two newswire datasets (BBN-ConNLL) show that the proposed model performs as well as the SCL model when adapting from BBN to CoNLL, and outperforms all the baseline methods (including self-training, SCL) when adapting from CoNLL to BBN. Their results also show that adapting from specific to general (BBN-ConNLL) is harder than in the opposite direction.

2.4.1.2 Supervised Domain Adaptation

This subsection continues by describing the models proposed for supervised domain adaptation, which refers to the situation when there is also a small amount of target domain data available ($L_T > 0$). In the following, feature-representation and parameter-based approaches applied for this setting are presented. A comparison of the methods reviewed is also given in Table 2.4.

Feature-representation Approaches

The feature-representation approaches assume that there exists a representation under which the two domains look more similar. Therefore, with this new representation the performance of the classifier on the target domain is expected to improve significantly [Daumé, 2007; Daumé III et al., 2010; Jiang and Zhai, 2007b].

Daumé [2007] proposed a simple algorithm called EasyAdapt (EA), which tries to overcome the domain difference, by extending the original feature space with a domain-specific copy of the original features for each of the domain. In other words, the original features are extended by features specific to the source domain and features specific to target domain. The augmented source domain instances therefore contain original and source-specific features, while the augmented target domain instances contain original and target-specific features. Next, a supervised classifier is trained on the labelled source and target instances with these new features and used to make prediction on the unlabelled target domain instances. Experimental results on NER, POS tagging, shallow parsing, show that this simple method outperforms several state-of-the-art methods (they are called SRCONLY, TGTONLY, ALL, PRED, LININT) on a range of datasets. SCRONLY refers to the classifier trained on the source domain instances only; TGTONLY refers to the classifier trained on the target domain instances only. ALL trains a classifier on both source and target domain instances. PRED uses the output of the source classifier as features in the target classifier. And LININT linearly interpolates the predictions of the SRCONLY and TGTONLY classifiers. One

[5] http://gate.ac.uk/

drawback of EASYADAPT, however is that it does not make use of the unlabelled target domain data which is usually plentiful in most practical problems.

This problem has been addressed in a recently proposed method, called EASYADAPT++ (EA++). EASYADAPT++ [Daumé III et al., 2010] creates an augmented feature space for the *unlabelled target instances* also, in a way that the source and the target classifiers agree on the unlabelled target instances. Furthermore, the augmented unlabelled instances are added to an augmented training set used to train the supervised classifier. Each unlabelled instance is added with all the possible class labels. For example, for a binary classification problem, two copies are added, one that assigns +1 label to the instance and another one that assigns -1. Empirical results on sequence labelling tasks (NER, POS tagging) show improved accuracy over the EA.

Jiang and Zhai [2007b] proposed a two-stage approach to domain adaptation considering multiple source domains. In the first *generalisation stage*, they identify a set of generalisable features learned from multiple source domains and set the appropriate weights for them. They proposed two ways for the identification of these features, namely an alternating optimisation procedure and domain cross validation, the latter one being observed to be more effective in their experiments. While in the second *adaptation stage* they select the features that are specific to the target domain using *semi-supervised* methods. One limitation of this framework is that it requires at least two source domains. An interesting research direction therefore would be to study whether a single domain and target examples with pseudo labels can also be used to identify the generalisable features.

Parameter-based Approaches

The parameter-based approaches assume that the source and target domains share some common parameters or prior distributions. The encoding of these parameters or priors can be done using labelled source domain instances as shown in [Arnold et al., 2008; Finkel and Manning, 2009].

Arnold et al. [2008] extended the original model proposed in Chelba and Acero [2004], which relies on a Maximum a posterior (MAP) adaptation through Gaussian priors of a Maximum Entropy (MaxEnt) model. That is, the model uses labelled source domain to find optimal weight parameters of a MaxEnt model, and then it set these parameters as a prior on the values of a model trained on the target domain. The novel extension then lies in assumption that the source and target domains share a common tree hierarchy. The leaves in the hierarchy correspond to these weight parameters which are also connected with hyper-parameters to form the tree. This method therefore first applies MAP estimation of the parameters and hyper-parameters using labelled source domain data, and next it set these parameters as Gaussian prior on a CRF model trained on the target domain. Experimental results on NER show that the hierarchical prior model outperforms several baseline methods, including the non hierarchical Chelba and Acero [2004] model.

Similarly, Finkel and Manning [2009] employed a hierarchical Bayesian prior model for domain adaptation, in which each of the domain has its own domain-specific prior, a Gaussian prior on the mean and variance of its weight vector. Also these parameters are connected via a hyper-prior thus forming a hierarchy. The goal of the learning is then to maximise the performance of the classifier over all the domains. Experimental results on named entity recognition and dependency parsing show that this model outperforms Daumé [2007] and

several baseline systems.

2.4.1.3 Combining Multiple Models

There has been some work on using ensemble methods for domain adaptation by employing boosting [Dai et al., 2007b; Al-Stouhi and Reddy, 2011].

Dai et al. [2007b] proposed a transfer learning framework called TrAdaBoost, which extends the Adaboost boosting-based algorithm. TrAdaBoost assumes that because of the distributional difference between the source and target domains, some of the source domain instances will be relevant in learning in the target domain but others will not be relevant and could even be harmful. The key idea, therefore, is to filter out instances that are irrelevant for the target domain by automatically adjusting the weights in each iteration as in Adaboost. Thus, it assigns less weight on the wrongly classified source domain instances, and more weight on the wrongly classified target domain instances, as the objective of the learning is to improve the performance of the final classifier on the target domain. Experimental results on several classification tasks show that TrAdaBoost has better transfer ability than in-domain learning techniques (SVM) and the error rate on the target domain depends on the similarity between the two domain (as measured by KL-divergence). The framework could be extended to deal with multiple different distributions.

Al-Stouhi and Reddy [2011] further presented an extended version of the TrAdaBoost approach, called Dynamic-TrAdaBoost, which aims to address the bias induced by the combined normalization of source and target instances. For this purpose a dynamic instance weighting function is employed, which combines the benefit of AdaBoost and Weighted Majority Algorithm. In particular the weighted majority algorithm being used for updating the source domain weights, and the Adaboost approach for weighting the target domain instances. Empirical results demonstrated the superiority of Dynamic-TrAdaBoost over TrAdaBoost on three benchmark datasets (20 Newsgroups, Abalone, Wine).

2.4.1.4 Multi-task Learning, Self-thought Learning

Unlike domain adaptation, where the learning tasks we wish to perform are the *same*, in multi-task learning [Caruana, 1997] these tasks are different but related and learned *simultaneously*.

Following the definition of transfer learning, the objective of **multi-task learning** is to improve the learning performance of the classifier in D_T by leveraging the knowledge in D_S, where the domains are the same ($D_S = D_T$) *and* the learning tasks are different but related ($T_S \neq T_T$) [Caruana, 1997; Pan and Yang, 2010]). The first condition implies that the feature spaces of the two domains are the same $F_S = F_T$ and there is a single distribution of the instances $P_S(X) = P_T(X)$. The second $T_S \neq T_T$ condition implies that either (i) the label spaces differ ($Y_S \neq Y_T$) or (ii) the conditional probability distributions are different ($P_S(Y|X) \neq P_T(Y|X)$). In addition, a *large amount* of labelled source data ($L_S > 0$) is available at training time.

The intuitive idea behind multi-task learning is that learning the tasks simultaneously improves the learning performance on the target task relative to learning each task independently. Another related approach called **self-taught learning** [Raina et al., 2007] has been studied in the literature, which makes the same assumption that *labels in the source*

Transfer learning technique	Method name	Labelled/Unlabelled Data				Classifiers			Similarity measure	Background knowledge	Solved task	Corpora	Performance measure
		L_S	U_S	L_T	U_T	Base classifier	Loss	Ensemble method					
Feature-representation based	Raina et al. (2007), Self-taught learning		✓	✓		SVM with Fisher kernel					image, audio and text classification	Reuters, Usenet, SRAA, Caltech101 image classification dataset	accuracy
	Daume (2009)	✓	✓	✓	✓	EM					sentiment analysis, landmine detection, text classification	Amazon product review, 20 newsgroup	accuracy
	Al-Stouhi and Reddy (2014)	✓	✓	✓	✓	NMF					text classification	20 newgroups, Reuters, WebKB4	accuracy

Table 2.7: Overview of the TC models proposed for the *Multi-task* and *self-taught* learning settings. Columns correspond to method proposed (Method name), transfer learning approach applied (Instance-based, Feature-representation based, Parameter-based), resources used during learning (labelled source (L_S) and target (L_T) data, unlabelled source (U_S) and target (U_T) data), classifier employed in transfer learning (Base classifier, Loss function, Ensemble method), similarity measure used to encode the divergence between the domains (Similarity measure), additional background knowledge used to enhance the learning (Background knowledge), the particular NLP problem addressed (Solved task), and the measure used to evaluate the performance of the classifier (Performance measure).

and target domains may be different. In the following we review two feature-representation methods proposed for this settings[6]. The methods reviewed are also presented in Table 2.7.

To learn a common feature representation for the two domains, in some work labelled source data is used [Daumé, 2009; Al-Stouhi and Reddy, 2014], while in others only the unlabelled source data is considered [Raina et al., 2007].

Daumé [2009] proposed a non-parametric Bayesian latent hierarchical model for both multi-task and domain adaptation settings. This model creates a tree structure over the data by employing Kingsman's coalescent prior [Kingman, 1982] and estimates the model parameters (such as the structure of the tree and the hyperparameters) using the EM algorithm. The leaves of the tree then corresponds to the domain-specific weight vectors, while the root of the tree corresponds to the global(common) weight vector. Experimental results on sentiment classification, landmine detection and text classification show significant improvement over EA [Daumé, 2007] and other baseline systems.

Al-Stouhi and Reddy [2014] proposed a multi-task clustering framework which combines multi-task learning with non-negative matrix factorisation (NMF) to support cross-domain clustering. For this purpose a multi-task affinity kernel is introduced, which allows to capture the intra-task and inter-task dependencies between domains. As a first step a multi-task graph is created over the instance and feature pairs of both source and target domains, where each edge being associated with a weight. Further various transformations are performed on this graph, which aims to reduce the distance between tasks by decreasing the weights which connects them (thus controlling the inter-task dependencies between domains). Then a symmetric NMF is employed for discovering clusters such that the clustering solution forces intra-task solutions. Experimental results on text classification showed promising results outperforming k-means, normalised N-Cut and standard symmetric NMF.

Raina et al. [2007] proposed a transfer learning algorithm called self-taught learning (STL) which can be used to learn from unlabelled source domain data, having different class labels than the target domain data. For example, in a text classification task of classifying Usenet articles into four predefined category (real auto, real aviation, simulated auto, simulated aviation) Reuters newswire articles were considered as unlabelled source data. First, the algorithm learns a high-level feature representation from the unlabelled source data by solving an optimisation problem. In particular, it learns a set of basis vectors and the corresponding weights for them. Next, it applies this representation to the labelled target data and trains a supervised classifier on the labelled data. Experimental results show that the algorithm significantly outperforms the classifiers built only on raw features and PCA features. However, one limitation of this algorithm is that the basis vectors learned on the source domain data may not be optimal for use in the target domain data.

2.4.2 Unsupervised Transfer Learning

In the previous two sections learning algorithms that require labelled source data were considered. In this section we review methods that don't require any labelled source and target data.

Following the definition of transfer learning, the objective of the **unsupervised transfer learning** is to improve the learning performance of the classifier in D_T by leveraging the

[6]For a summary of other methods proposed for these settings we direct the reader to a recent survey on transfer learning Pan and Yang [2010].

Transfer learning technique	Method name	Labelled/Unlabelled Data				Classifiers			Similarity measure	Background knowledge	Solved task	Corpora	Performance measure
		L_S	U_S	L_T	U_T	Base classifier	Loss	Ensemble method					
Feature-representation based	Huang and Yates (2010) I-HMM		✓		✓	HMM with CRF					POS tagging	WSJ, MEDLINE, Penn Bio IE	Accuracy
	Martín-Wanton et al. (2013)		✓		✓	LDA					tweet clustering	Replab 2012 Monitoring Task	Accuracy
	Tong et al (2013) GPDRTL	✓			✓	Gaussian Process, Spectral regression					text classification, face recognition	20 Newsgroup, wine quality data, face recognition dataset	Accuracy

Table 2.8: Overview of the transfer learning models proposed for the *Unsupervised transfer learning* setting. Columns correspond to method proposed (Method name), transfer learning approach applied (Instance-based, Feature-representation based, Parameter-based), resources used during learning (labelled source (L_S) and target (L_T) data, unlabelled source (U_S) and target (U_T) data), classifier employed in transfer learning (Base classifier, Loss function, Ensemble method), similarity measure used to encode the divergence between the domains (Similarity measure), additional background knowledge used to enhance the learning (Background knowledge), the particular NLP problem addressed (Solved task), and the measure used to evaluate the performance of the classifier (Performance measure). **Abbreviations used**: MI for mutual information.

knowledge acquired in D_S, where the two domains are different but related ($D_S \neq D_T$), and the learning tasks are also different but related ($T_S \neq T_T$), and Y_S and Y_T are not observable [Pan and Yang, 2010].

The majority of the work proposed for unsupervised transfer learning focuses on unsupervised tasks such as clustering [Huang and Yates, 2010; Martín-Wanton et al., 2013] and dimensionality reduction [Tong et al., 2013]. The methods reviewed are presented in Table 2.8.

Huang and Yates [2010] proposed the I-HMM algorithm, which uses unsupervised HMM model to induce clustering of words for POS tagging. The algorithm also creates several copies of the produced clusters, resulting in a multi-dimensional representation for transfer learning. Next these clusters are used to create new features for a supervised classifier (CRF) employed to make prediction on the unlabelled target domain data. Experimental results show that the proposed HMM features are more stable across domains than the purely lexical features. Also the model outperforms several state-of-the-art methods including self-training, SCL. However, one limitation of this method is that the number of layers is not determined automatically.

Similarly, Martín-Wanton et al. [2013] proposed a graphical model to transfer knowledge across domains for topic classification in the context of reputation management involving multiple companies. In this scenario, documents related to a particular company being considered as individual domains, and thus a small collection of documents related to that company is regarded as a target domain. For compiling the source domain data further three different cases have been exploited: considering a large collection of documents about a company of interest, considering a large collection of documents about a different company, and creating a joint set of documents comprising of the previous two collections. The employed graphical model, called TwitterLDA, then aims to cluster the target domain instances into topics while making a distinction between domain specific and domain independent words. Experimental results indicate that the modified TwitterLDA model exploiting all the three different source domain data significantly outperforms the TwitterLDA model using only target domain instances.

Tong et al. [2013] studied the problem of dimensionality reduction for transfer learning. Their approach relies on exploiting the labelled information from multiple source domains to perform dimensionality reduction on the unlabelled target domain instances. For this purpose a Gaussian process model, named Gaussian process for dimensionality reduction in Transfer learning (GPDRTL), is employed, which relies on the transformation of the original dimensionality problem into a regression problem. In order to achieve this, a spectral regression algorithm is first applied, which creates a low dimensional space for the multi source domains and the target domain also. Following this a GP model is constructed on the labelled source domain data exploiting two different regularisation terms. The first such regularisation term aims to capture the similarity between the target domain data and source domain data in a specific dimension of the reduced space using a smoothing function. The second regularisation term on the other hand, aims to capture the relationship in both the data and task between the source and target domains. Experimental results on text classification show that GPDRTL outperforms various state-of-the-art approaches such as Principal component analysis (PCA), transfer dimensionality reduction [Pan et al., 2008], and transfer component analysis [Pan et al., 2009].

2.4.3 Active Transfer Learning

As described in the previous sections (2.3, 2.4) when labelled data is scarce, transfer learning and active learning are separate solutions to obtain labelled data for the target domain, however both have some practical constraints.

Transfer learning can leverage the knowledge acquired from a related domain without incurring labelling cost, but there is no guarantee that the transferred knowledge will help improve learning performance in the target domain. While active learning asks domain experts to label a small set of examples, but there is an implicit cost associated with obtaining the labels.

To avoid domain transfer risk and reduce labelling cost, a combination of the two methods have been recently proposed, called *active transfer learning* [Shi et al., 2008]. The methods reviewed are presented in Table 2.9.

In some work the transfer learner requires an initial pool of labelled target domain data [Shi et al., 2008; Xiao and Guo, 2013], while in other work only labelled source domain data are necessary [Rai et al., 2010].

Shi et al. [2008] proposed an active transfer learning algorithm called AcTraK (Actively Transfer Knowledge) which first applies a traditional active learning algorithm (called Error Reduction Sampling method (ERS) [Roy and Mccallum, 2001]) to select an example from the unlabelled target data. Next, this selected example is labelled by a transfer classifier (SVM) trained on the labelled source data (L_S) and small amount of labelled target data (L_T). When the transfer classifier's labeling confidence is low, domain experts are asked to re-label the example. Experiments on text classification and remote sensing problems show that AcTrak significantly outperforms the TrAdaBoost transfer learning algorithm and the ERS active learning algorithm. However, one drawback of AcTraK is the requirement of a small amount of labelled target domain data (L_T) used to train the transfer classifier.

Xiao and Guo [2013] proposed a framework which combines multi-task learning with active learning in a cost sensitive way, considering the situation when annotated data in both domains are available and further querying labels for the target domain is cheaper than for the source domain (quantified as a cost ratio). As the base active learner, the Cesa-Bianchi approach [Cesa-Bianchi et al., 2003] is employed which uses the perceptron algorithm. First, an initial supervised classifier is built on the annotated instances of both domains. Then the instances of both domains are randomly split into two distinct views, and the active learner is iteratively executed querying labels for them considering two different multi-task strategies: multi-view uncertainty strategy and multi-view disagreement strategy. In the first uncertainty strategy, the cost ratio is used to select the instance to be labelled from either source domain or target domain. For the selected instance then the prediction confidence values of the learner are computed, and in the case the labels for the two views disagree then a new label is queried. In the second multi-view disagreement case, the prediction confidence of both instances and views are first computed and compared. Then, if the disagreement of the views occurs for only one of the instances, then a new label is queried for only that instance, otherwise the predicted labels for each instance is compared, and if the labels are different than labels are queried for both instances, while if the labels are the same (indicating that the two instances contain similar information concerning the label) then a new label is queried only for one instance. Experimental results on sentiment

Transfer learning technique	Method name	Labelled/Unlabelled Data				Classifiers			Similarity measure	Background knowledge	Solved task	Corpora	Performance measure
		L_S	U_S	L_T	U_T	Base classifier	Loss	Ensemble method					
Feature-representation based	Shi et al. (2008), AcTrak	✓		✓	✓	SVM					remote sensing; text classification	Landmine data, 20 Newsgroup	accuracy
	Rai et al. (2010), DS-AODA	✓	✓		✓	Perceptron	Hinge loss		A-distance		sentiment classification	Amazon product reviews	accuracy
	Xiao and Guo (2013)	✓	✓		✓	Perceptron	Hinge loss		A-distance		sentiment classification	Amazon product reviews	accuracy

Table 2.9: Overview of the TC transfer learning models proposed for the *Active transfer learning* setting. Columns correspond to method proposed (Method name), transfer learning approach applied (Instance-based, Feature-representation based, Parameter-based), resources used during learning (labelled source (L_S) and target (L_T) data, unlabelled source (U_S) and target (U_T) data), classifier employed in transfer learning (Base classifier, Loss function, Ensemble method), similarity measure used to encode the divergence between the domains (Similarity measure), additional background knowledge used to enhance the learning (Background knowledge), the particular NLP problem addressed (Solved task), and the measure used to evaluate the performance of the classifier (Performance measure). **Abbreviations used**: MI for mutual information.

analysis using the Amazon products review domain showed promising results outperforming various approaches exploiting labelled strategies from a single domain, as well as an approach using single view uncertainty strategy.

Rai et al. [2010] proposed the Active Online Domain Adaptation (AODA) algorithm, which first learns the optimal weights for a transfer learning classifier (such as SCL or KLIEP) using labelled source (L_S), unlabelled source (U_S), and unlabelled target data (U_T). Next the learned weights are set to an active learner, which employs the Cesa-Bianchi approach [Cesa-Bianchi et al., 2003] to iteratively query labels for the unlabelled target examples. A version of AODA, called Domain-Separator based Active Online Domain Adaptation (DS-AODA) was also proposed, which uses the domain divergence information between the domains to select the examples to be labelled. The domain divergence measure used is called *proxy A-distance*, which is computed similarly to *A-distance* [Ben-David et al., 2006]. Namely, a linear classifier is trained on the unlabelled source and target data by treating each source domain example as negative example, and each target domain example as positive example. And the value of proxy A-distance is then the average per-example hingle-loss of the classifier subtracted from 1. Experiments on sentiment classification show that AODA outperforms several state-of-the-art supervised domain adaptation methods, such as EASYADAPT [Daumé, 2007], TGTONLY in-domain classifier, and DS-AODO leads in further reduction in label complexity. An interesting future direction would be to precisely quantify the amount by which the label-complexity is expected to reduce.

2.4.4 Mistake Bounds and When to Transfer

The transfer learning models presented in previous sections address the research issues of "*what to transfer*" and "*how to transfer*". This section now turns to the discussion of the third research issue of "*when to transfer*", which asks in which situations should transfer learning be applied.

As presented in the previous sections, transfer learning has been successfully applied for many learning problems, however there is no guarantee that it improves the learning performance on the target domain. Situations when applying transfer learning worsens the performance of the learner are referred to as *negative transfer*, which is still considered an open issue [Pan and Yang, 2010].

However, there have been a few theoretical studies of the problem of transfer learning, giving mistake bounds on the target domain, by considering the performance of the model on the source domain and the divergence information between the domains. In some work learning from a *single source domain* [Ben-David et al., 2006] is studied, while in others learning from *multiple source domains*[Blitzer et al., 2007a; Ben-David et al., 2009] is considered.

Ben-David et al. [2006] gave an upper bound on a classifier's error in the target domain in terms of its error in the source domain and a divergence measure between the two domains:

$$\epsilon_T(h) \leq \epsilon_S(h) + \beta + d_A(U_S, U_T) + \lambda,$$

where $\epsilon_T(h)$ is the expected error of the classifier in the target domain, $\epsilon_S(h)$ is the expected error of the classifier in the source domain, $d_A(U_S, U_T)$ is the A-distance measure between the two domains (computed as in [Rai et al., 2010, Section 2.4.3]), λ is the sum of the error

of the optimal classifier in the source and target domains, and β is a constant parameter. Their analysis show that the SCL finds a good feature representation resulting both in small empirical classification error and small A-distance between the domains.

Ben-David et al. [2009] gave an error bound on the target error of a classifier which minimizes the weighted combination of its error in the source and target domains, in situations when there is a large amount of labelled source data available from multiple source domains, and a little or no labelled data from the target domain. They also studied the conditions under which such classifier is expected to perform well on the target data. Their obvious observations were as follows: if the source and target domain data are the same, than the best performance is achieved by uniformly weighting the source and target data. If there is only a small amount of target domain data available, than it might be the case that the source data is not enough to justify it, and in this case the best is to ignore the source data. And finally, if there is enough target data, then no source data is required, and actually using this source data will worsen the performance.

In addition, there have been some empirical studies conducted in the literature, which looked at domain similarity measures which correlate well with the performance of a classifier across domains for a given NLP task, such as POS tagging [Van Asch and Daelemans, 2010] or sentiment analysis [Ponomareva and Thelwall, 2012a].

Van Asch and Daelemans [2010] showed that computing the similarity between the source and target domains can help to predict the performance of a classifier across domains. They conducted experiments on POS tagging using several frequency based similarity measures, which can be computed from unlabelled source and target data (including Renyi, L1, Eucledian, Cosine, KL divergence). They concluded that the best similarity measure to choose for the target task is the one which gives the best linear correlation between the similarity measure and the accuracy of the classifier applied across domains. If such measure is found, we can take annotated data from a related source domain and assume that this yields to best accuracy on the target domain. Using such measure for clustering the domains, may also help to measure transferability across domains.

Ponomareva and Thelwall [2012a] proposed a domain divergence metric for sentiment analysis, which takes into account the contribution of two different factors which can influence the performance of a classifier across domains: *domain similarity* and *domain complexity*. For measuring the domain similarity between domains different information theoretic measure were exploited such as KL divergence, Jenson-Shannon divergence, Jacard coefficient, cosine similarity, and χ^2. While for quantifying the domain complexity the percentage of rare words, word richness (measured as the ration of vocabulary size to the corpus size) and relative entropy (measured as the ratio of corpus entropy to the maximum entropy) were considered. Experimental results in the context of Amazon product reviews revealed χ^2 as being the best domain similarity measure, and the percentage of rare words as being the best domain complexity measure. Further the linear combination of these two domain characteristics showed a high correlation, over 88%, with the accuracy loss of a sentiment classifier across domains.

Another interesting perspective of analysing different domain difference types was studied in Jiang [2008b]. She proposed *four domain difference types* and gave recommendations on the transfer learning techniques to apply based on these types. All the four domain difference types consider the properties of the features in the two domains. The first two

types distinguish between a set of features which are frequent in one domain but infrequent in the other domain. The next two types distinguish between features which are discriminative in one domain but non-discriminative in the other domain. Empirical results reveal that the different domain difference types require different transfer learning techniques. Her findings were as follows: when the domain difference comes from the domain specific characteristics in the target domain, then the best performance is achieved by increasing the contribution of the labelled target domain examples rather than applying instance-weighting or feature-selection methods on the source domain. If the domain difference comes from the domain specific characteristics in the source domain, than if there is a small amount of labelled target domain data, then they can be used to remove the irrelevant examples from the source domain using either instance-weighting or feature selection methods or to increase the contribution of the target domain examples. An interesting future direction would be to propose other measures of domain difference and study whether recommendations of transfer learning techniques can be given based on these measures.

2.5 Limitations of Current Approaches

The previous sections reviewed the current state-of-the-art approaches for adaptive TC. The majority of these approaches employed various content-based lexical pivot features (bag-of-words, POS tags) for transfer learning. That is, these approaches extracted pivot features from the textual content of the domain documents only, and also computed the weights for these features based on some content-based statistics. In the same vein, existing domain similarity measures apply content-based statistical measures to estimate the performance of a text classifier.

Despite of the success of existing approaches, they still suffer from several limitations.

Firstly, the majority of work still relies on labelled source data [Jiang and Zhai, 2007a; Blitzer et al., 2006; Guo et al., 2009] and target data [Dai et al., 2007b; Daumé III et al., 2010; Arnold et al., 2008]. However, it has been shown that creating labelled data is both time consuming and expensive [Ciravegna et al., 2002; Zhang et al., 2010]. Also, the rapid rate of document publication in large repositories makes it infeasible to collect annotated data for every domain. Although some methods leveraged unlabelled data in unsupervised [Dai et al., 2008; Huang and Yates, 2010] and semi-supervised [Dai et al., 2007a; Jiang and Zhai, 2007a] manner, most of these methods were tested on specific TC tasks (such as part-of-speech-tagging, sentiment classification), and it is unclear whether these approaches are generalisable to other TC tasks, such as *document zoning* and *text topic classification*.

Secondly, a rich type of resource for text classification is *background knowledge*. Often, a large amount of information already exists in certain types of formats (e.g., knowledge bases, databases, unlabelled corpora). Research has shown that using *background knowledge* can indeed enhance the adaptation of TC systems across domains (Ciaramita and Yasemin (2005), Ciaramita and Chapelle (2010)). However, these methods are limited to exploiting simple *background knowledge* structures such as gazetters, and dictionaries. A systematic and generalisable methodology is required to harness the diverse and rich background knowledge, which is believed to be largely beneficial to domain adaptation of TC systems.

Thirdly, most of the previous work on transfer learning has focused on relatively small-scale applications using publicly available datasets (e.g., newswire) [Daumé, 2009; Finkel

and Manning, 2009]. The adaptation of TC systems to highly heterogeneous technical domains has been scarcely studied. This problem is much more complex, due to the intrinsic complexity of the language within these documents [Guo et al., 2006; Wang, 2009]. Therefore, it is unclear whether the existing transfer learning techniques are applicable to more complex text classification tasks.

In order to address these limitations, this thesis explores the use of *domain knowledge sources* for both building adaptive TC systems and designing better domain similarity measures for TC. Before proceeding to the presentation of these approaches, however, in the next section a discussion is given on the main roles of domain knowledge in TC.

2.6 The Role of Domain Knowledge in Text Classification

It has been largely agreed that content-based lexical features play an important role in text classification. These features are usually generated based on the terms (words or other larger textual units such as phrases) present in the documents, and their contexts. Research has shown that these features are not always sufficient. Typical examples are short, ill-formed texts, such as social media posts, where the correct interpretation and classification of text is very challenging due to the lack of contextual clues, and presence of abbreviations and misspellings. For instance, the following text "Microsoft and Apple are long-time rivals.", may refer to a text written about sports, politics or the financial situations.

The annotation of such short text by a human reader would be a challenging task, if the reader is not provided with any labelled examples on this topic. In such cases the reader will use his/her background knowledge to interpret the text. Further, if the annotator is unfamiliar with the domain, then s/he would rely on some background knowledge, such as a dictionary or an encyclopaedia to assign an appropriate topic label. For instance, s/he may exploit the Wikipedia encyclopaedia, and find "Microposts" and "Apple" as instances of "Software companies", and thus reveal that this text refers to financial situation between the two companies.

In similar situations, incorporating *domain* (or *background* or *external* or *prior*) knowledge has been found to provide additional contextual information for the documents, and as a result to lead to better TC performance. As the term *domain* knowledge has been vaguely defined in the literature, in this thesis the definition from [Zhang, 2013] is employed, which considers domain knowledge any additional learning evidence used for text classification.

There has been various domain knowledge types explored in the TC literature. In the coming section, the most frequently used domain knowledge types are presented, and a classification of their type is given.

2.6.1 Domain Knowledge Types used in Text Classification

The main domain knowledge types used in TC can be divided into three main classes: *domain-specific lexicons*, *unlabelled data* and *external knowledge sources*.

2.6.1.1 Domain-specific Lexicons

Much work on TC employs domain-specific lexicons in the form of gazetteers or lookup lists. These lexicons often contain a list of representative words for the corresponding document classes.

Sentiment classification approaches typically make use of subjectivity lexicon: for instance, one consisting of words describing positive polarity, and one containing negative polarity words. Barbosa and Feng [2010] implemented a two-stage approach for sentiment detection of tweets. The first stage distinguishes between subjective and non-subjective tweets, while the second step takes the subjective tweets previously identified, and labels them to positive or negative classes. To build the subjectivity classifier, the bag-of-word features are extended with subjectivity information for each word (e.g., weak or strong subjectivity) from a lexicon [Wiebe and Riloff, 2005]. For the polarity classification step multiple off-the-self sentiment extractors are combined, which make also use of their own sentiment lexicons.

Specifically for social media posts, it is also common to employ emoticon lexicons consisting of a list of emoticon signs bearing both positive and negative sentiments. For instance, the lexicon of positive sentiments would contain ":)", and the lexicon of negative sentiments would contain ":(". Go et al. [2009] used a pre-compiled emoticon lexicon to automatically collect tweets with positive and negative sentiment. For this purpose the Twitter API is queried for the different emoticons. Following that, a supervised classifier is built on the labelled instances using both uni-gram and bi-gram features.

For text classification, there is a recent trend of incorporating prior knowledge in the form of lexicons provided by human annotators, consisting of features associated with a particular class. Druck et al. [2008b] proposed a semi-supervised approach for text classification, making use of prior knowledge in form of labelled features. For instance, when building a classifier distinguishing between topics such as baseball and hockey, instead of providing labelled instances to the classifier, this approach requires as input a list of words related to baseball (such as "ball", "baseball", "pitching", hitter etc.) and a list of words related to hockey (such as hockey, period, shots etc.). These labelled features are then used to obtain predictions on the unlabelled instances. For this purpose the generalised expectation criteria is used with the Maximum Entropy discriminative classifier, allowing to put constraints on the classifier's expectations for certain class-word combinations.

Although the exploitation of domain-specific lexicons seems to improve the performance of a text classifier, their application still faces several limitations. First, the vocabularies of lexicons are constantly changing, as new words are continuously created for describing a particular domain. For instance for the domain of internet computing words such as "Pinterest", "Instagram", "Google+", have been introduced in the last four years. Second, the meaning of a particular word can also evolve with time. For example, the word "owl" has been used to refer to a type of nocturnal bird, however with the advent of semantic web, it has been assigned new meaning (Word Ontology Language).

2.6.1.2 Unlabelled Corpora

Another source of valuable background knowledge for TC is a large collection of unlabelled documents. In such situations clustering based approaches are typically employed. These

approaches use classical clustering assumption that two documents which are in the same cluster have the same label.

Liu et al. [2004] proposed an approach which makes use of a large pool of unlabelled examples for TC. This approach relies on the combination of clustering and feature selection. First the documents from the pool are clustered using k-means, resulting in each document being associated with a category. Next, entropy-based feature selection is performed on the clusters to identify a ranked list of representative words. Following this process, the selected words are labelled by human annotators and used to create a list of representative documents for each class, serving as labelled data for TC. After that, the Expectation-Maximization (EM) algorithm with a Naive Bayes classifier [Nigam et al., 2000b] is built on the representative examples and the remaining unlabelled data.

Sindhwani and Melville [2008] proposed a semi-supervised approach for sentiment classification which exploits the information present in both sentiment lexicons and unlabelled data. The information present in the sentiment lexicon is used to label a few documents with sentiment labels. Further, this model constructs a document-word bipartite graph containing both labelled and unlabelled instances, where the values of the edges correspond to the inverse document frequency of terms in a document. On this partially labelled graph, then different regularisation operators are applied which conceptually implement a co-clustering approach, enforcing terms which are in the same cluster dominantly supported on positive (negative) sentiment documents, to be most likely positive (negative) sentiment words.

One of the main advantages of exploiting background knowledge from unlabelled corpora, as compared to domain-specific lexicons, is that unlabelled corpora is typically much easier to obtain and maintain. However, the existing approaches for mining knowledge from unlabelled corpora typically rely on semi-supervised or unsupervised approaches, which may produce inaccurate results, introducing noise information into learning, which can lead to a performance loss for TC.

2.6.1.3 External Domain Knowledge Sources

The third type of domain knowledge exploited in TC is *external domain knowledge sources (KSs)* (also called *knowledge bases*), which typically define information and knowledge about concepts, which is not present in the training and unlabelled data. Such information about concepts is present in KSs via various graph structures surrounding concepts. These graph structures connect concepts to one another using links and edges that represent a certain type of semantic relatedness.

In the literature, there have been several different terms used to refer to KSs: such as *taxonomy*, *ontology* or *semantic network*. In a taxonomy the concepts are grouped into a hierarchical structure according to the is-a (generalisation/superclass) and has-a (specialisation/subclass) relationships. In an ontology, the concepts are organised into hierarchies according to several other semantic relationships, such as synonymy, antonymy or class properties. An ontology can thus be viewed as an enriched taxonomy. A semantic network or semantic graph is regarded as a concept graph in which the concepts are connected by any type of semantic relationships.

Over the past few decades, a large number of KSs have been developed. These knowledge sources can broadly be classified into *domain-specific* knowledge sources (such as the

ones developed for the biomedical domain, e.g., Unified Medical Language System UMLS, SNOMED-CT), and *domain-independent* knowledge sources (which cover a wide range of different domains, e.g., WordNet, Wiktionary, DBpedia, etc.).

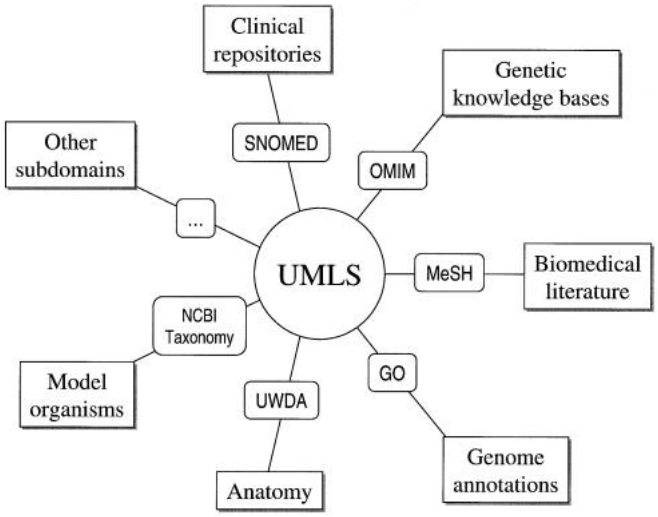

Figure 2.2: A fragment of the UMLS domain-specific ontology from Bodenreider [2004].

Among the domain-specific knowledge sources, an example of the most widely used *biomedical ontology* in NLP is UMLS. This ontology organises the biomedical concepts and knowledge into hierarchies and associated relations. There are three main components in UMLS: the *Metathesaurus*, the *Semantic Network* and the *Specialised Lexicon*. These concept have been added and maintained by the US government with the help of domain experts since 1984. The *Metathesaurus* is a taxonomy of biomedical concepts. The *Semantic Network* provides a broad classification of Metathesaurus concepts considering categories such as organisms, biologic functions or chemicals, and further connects these semantic categories through relations (such as hypernymy, meronymy or synonymy relation). The *Specialised Lexicon* is a lexicon of both English common words and biomedical terms. The main benefit of UMLS is that it interlinks several biomedical ontologies, and provides a mapping between their vocabularies. A partial list of these interlinked ontologies is shown in Figure 2.2[7]:

- Medical Subject Headings (MeSH) [Rogers, 1963][8]: is a taxonomy of biomedical terms and concepts, which is used to categorise biomedical literature for indexing purposes. These concepts are manually selected by experts and then used to index the biomedical articles through the Medline[9]/Pubmed[10] article databases.

- Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [IHTSDO, 2010][11]: is a comprehensive taxonomy of clinical healthcare terminology, covering terms related to clinical findings, symptoms, diagnoses, procedures, body structures, organisms, etc.

[7]The full list of interconnected ontologies can be found at http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/source_vocabularies.html
[8]http://www.nlm.nih.gov/mesh/
[9]http://www.nlm.nih.gov/pubs/factsheets/medline.html
[10]https://www.ncbi.nlm.nih.gov/pubmed/
[11]http://www.ihtsdo.org/snomed-ct/

- Gene Ontology (GO) [Ashburner et al., 2000]: is an ontology for the functional annotation of gene-products. It connects each term to other terms within and across sub-domains using various relations (such as synonymy, hypernymy, meronymy, and domain-specific relations such as regulation).

- National Center for Biotechnology Information (NCBI) Taxonomy: is a taxonomy covering terms related to organisms, mostly those described in GenBank[12], a database of nucleotide sequences for organisms.

Concerning the domain-independent knowledge sources, the most widely used KSs in TC are the following:

- Wordnet [Fellbaum, 1998]: is a lexicalised ontology for the English language, which has been manually developed by the Cognitive Science laboratory at Princeton University. This KS serves the role of a dictionary and thesaurus, aiming to group words together into sets of synonyms (called *synsets*), providing a short definition of each word (called *gloss*), and defining several semantic relationships among them (e.g., hypernymy, hyponymy, coordinate terms, synonymy and antonymy).

- Wikipedia[13]: is a KS built in a collaborative manner, providing a network of articles on a wide range of different topics and domains. In this KS, the articles can thus be regarded as the entities and concepts of the domains. There are several properties and useful graph structures defined for an article. Concerning the properties, the first paragraph of the articles provides a *description* (short definition) about the article; and additional *metadata information* (such as the place-of-birth for a person) is provided in an *infobox* template, using a tabular format. With regard to semantic graph structures, the Wikipedia concept graph assigns classes to articles based on their type (e.g., Person, Book, etc.), while the Wikipedia category graph associates articles to category labels (general concepts) capturing the topic of the articles. Both of these semantic graphs form a hierarchical structure, capturing the broader/narrower relationships. In addition, the synonymous relationships among articles are defined using *redirects*, and polysemous article names are grouped together under *disambiguation* pages.

- DBpedia [Auer et al., 2007]: is a KS derived from Wikipedia, consisting of structured information extracted from Wikipedia infoboxes. This extracted information is then used to define properties and relationships among the articles, which are stored using Semantic Web standards in RDF triples, in the form of subject-property-object pairs.

- Freebase [Bollacker et al., 2008]: is another knowledge source built in a collaborative manner, harvesting information from multiple sources such as Wikipedia, ChefMoz, NNDB and MusicBrainz[14], along with data individually contributed by users. Freebase has its own ontology, where the classes form a hierarchical structure according to the generalisation/specialisation relationships. The classification of articles in Freebase is, however, slightly different. For a given Freebase article, a *domain* denotes the topic of the article; a *type* defines a particular kind of entity such as person or location (e.g.,

[12] https://www.ncbi.nlm.nih.gov/genbank/
[13] http://www.wikipedia.org/
[14] http://wiki.freebase.com/wiki/Data_sources

"Barack Obama" is a Person); and *properties* describe an entity (e.g., "Barack Obama" has a "place of birth").

- OpenCyc: is a broad coverage ontology built in the context of the Cyc artificial intelligence project [Lenat, 1995], capturing common sense knowledge of everyday life. The content of the ontology is divided into two main ontological categories: *collections*, corresponding to the classes of the ontology (e.g., people, books, etc.), and *individuals*, which are instances of the collections. This KS thus contains a network of concepts and their relationships (mostly taxonomical relations).

As a great achievement of the Semantic Web community, the above mentioned cross-domain KSs have been interlinked to each other, as well as to other KSs (such as Geonames, Linked MDB) as part of the Linked Open Data (LOD) project[15]. This provides a uniform approach to interlink entities which have a unique dereferentiable URI, a URI which is accessible online. According to the latest LOD cloud[16] statistics, the cross-domain dataset (e.g., encyclopedias mentioned above) constitute 13% of the LOD cloud, and further 44% covers a wide range of domains including "Media", "Life Science", "Geographic" and "Publications" domains.

Knowledge Sources and Knowledge Representation

In the majority of the KSs, the information and knowledge present in a KS is classified into its own consistent ontology, providing means to formally describe the knowledge of their covered domains. A formal definition of the ontology is given as follows:

Definition 2 (Ontology) *An ontology O is a tuple, $O = (Cls, P, R, T, Y, A)$ where*

- `Cls, P, R` *are finite sets whose elements are* `classes(or concepts)`, `properties` *and entity* `resources(facts)` *of the ontology;*
- `T` *is a set of relationships between concepts,* $\texttt{T} \subseteq (\texttt{Cls} \times \texttt{Cls})$ *(e.g., Mother is a kind-of Person)*
- `Y` *is a set of relationships between an ontological element and its instances,* $\texttt{Y} \subseteq (\texttt{R} \times \texttt{Cls}) \cup (\texttt{R} \times \texttt{P}) \cup (\texttt{R} \times \texttt{R})$ *(e.g., "Barack Obama" is a President; "Barack Obama" wasborn on 4/08/1961; "Michelle Obama" is the wife of "Barack Obama")*
- `A` $= \{condition => conclusion(cls_1, \ldots, cls_n), \forall i, c_i \in Cls\}$ *a set of axioms, rules that allow checking the consistency of an ontology and infer new knowledge through some inference mechanism. (e.g., if two daughters have the same mother then they are sisters.)*

An ontology typically consists of *classes* (also called types or categories), e.g., *concepts* that represent entities, which describe a set of objects (for example, "Mother" $\in Cls$, or "President of United States" $\in Cls$); *individuals* (or entity resources) which are instances of classes (e.g., `http://dbpedia.org/resource/Barack_Obama`); *properties* (or attributes) that entities may have (e.g., "birthDate"); *relations* which connect concepts and other ontological elements together, and *axioms*, a set of rules defined over the ontological classes. It is worth noting that the literature often makes a distinction between classes and concepts, considering concepts to be the instances of classes, representing more fine grained information [Butters and Ciravegna, 2010]. For instance, given the DBpedia ontology, one may consider

[15] `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`
[16] http://lod-cloud.net

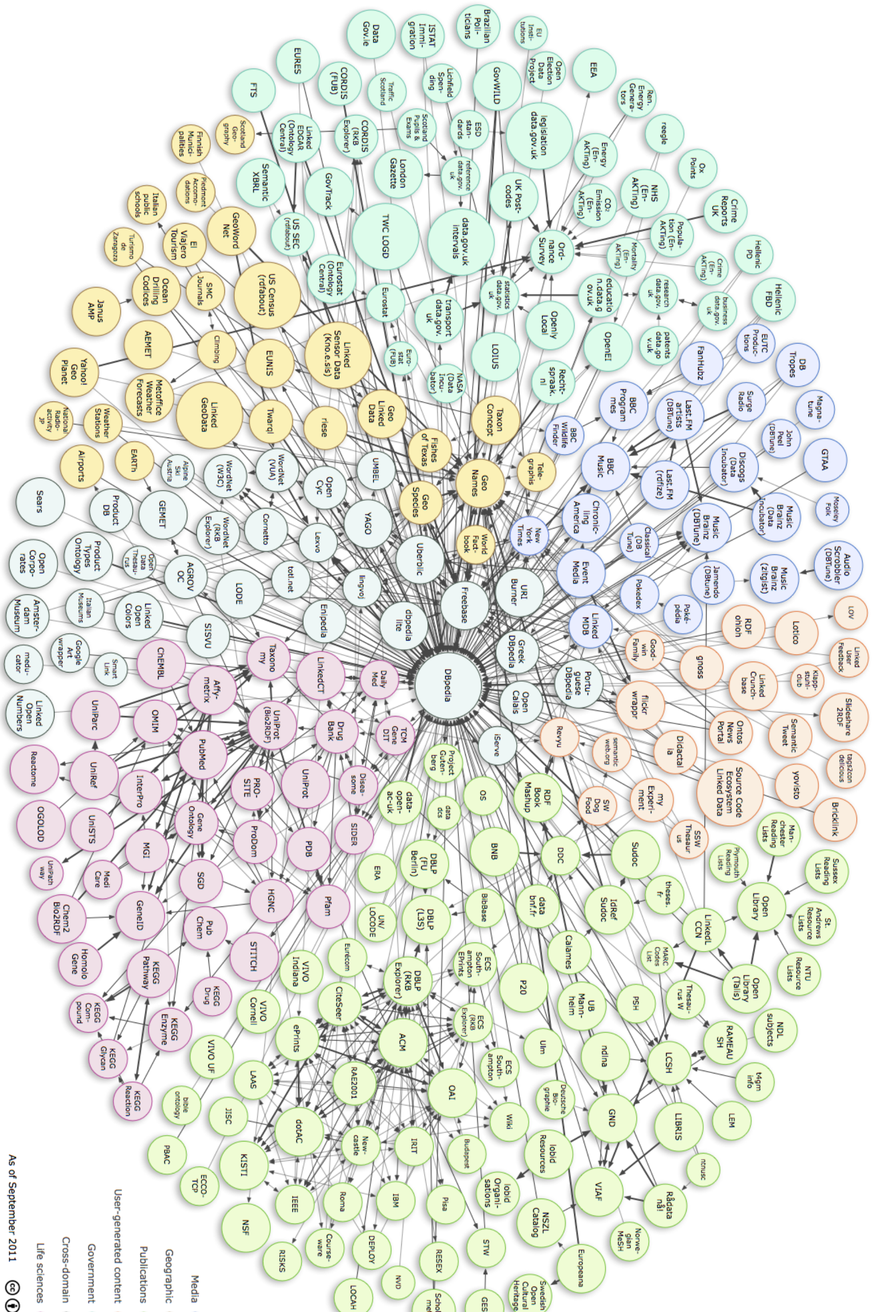

Figure 2.3: Linked Open Data Cloud interlinking various domain-specific and domain-independent ontologies [LOD, 2011].

"President of United States" as a concept, and "Person" as a class. Within the context of this thesis, however, no distinction is made between concepts and classes, and thus these terms are used interchangeably.

In order to describe ontologies, several formal *ontology languages* have been designed. These different languages differ by their level of complexity (and specificity) ranging from thesaurus level to description logics (using e.g., OWL Lite[17], OWL DL, RDF[18]), full first-order logics (e.g., OWL Full,) or higher order logic (e.g Cyc). For describing the instances, RBF triples are typically used following a subject-predicate-object syntax. For instance, having an entity resource in an ontology, and considering its relations to other entities, one can consider the wrote-novel or president-of and attributes, such as population-size or date-of-birth (e.g., Obama wasborn on 4/08/1961).

Domain Knowledge Sources and Text Classification

The broad information coverage of KSs, their structural properties and continual updates have made KSs appealing for TC. In addition to these advantages, certain KSs (mostly the collaborative KS such as Freebase and DBpedia) have the advantage of associating textual data (called definition or abstract) with its entities and concepts.

This wealth amount of domain knowledge has been exploited for the following benefits: improving the coverage of domains [Garla and Brandt, 2012; Li et al., 2009; Saif et al., 2012; Hulpus et al., 2013], enhancing the representation of documents [Saif et al., 2012] or providing a broader classification of document classes based on the KS ontological classes [Hulpus et al., 2013].

Garla and Brandt [2012] explored UMLS for clinical text classification aimed at detecting whether a document discusses a particular disease or not (e.g., Hypertension, Hypercholesterolemia). In order to achieve this a novel feature selection approach is proposed, which aims at representing the documents by their most discriminative features. The proposed feature ranking method is a modified version of the Information Gain measure, which takes into account the taxonomical structure of UMLS (specifications (children)) and hypernymyms of a concept), and the similarity between concepts (as measured by the Lin measure [Lin, 1998]). The intuition underlying this feature selection is that, if a concept is not relevant to a classification task, then the similarity between this concept's children is also not relevant.

Li et al. [2009] exploited WordNet for the topic classification of documents. Given a name of a topic (e.g., "computing"), the first step in this approach consists of performing a word sense disambiguation approach for mapping this name to its correct WordNet sense. Based on the selected word sense, Wordnet is used to retrieve the synonyms, hypernyms, hyponyms, meronyms and holnyms of the word and create a "pseudo document" (expanded feature set) for that topic. Next, a probabilistic labelling approach is applied, which represents the unlabelled documents and the pseudo document using a vector space model, and computes the cosine similarity between these two vectors to decide if a document belongs to a topic or not. Following this a discriminative classifier is built on the labelled documents and applied on the unlabelled documents.

Gabrilovich and Markovitch [2006] exploited Wikipedia for creating an enhanced representation of documents for TC. Their approach relies on mapping the content of the

[17] http://www.w3.org/TR/owl-features/
[18] http://www.w3.org/RDF/

documents to Wikipedia articles (called concepts), and derive additional semantic features about these concepts, complementing the original bag-of-words features. In order to achieve this, an inverted index is built on the Wikipedia articles, in which each word is associated with the Wikipedia articles they appear. Given an input document, then the system combines the vectors associated with each of its words, incorporating two different semantic features: concepts and categories derived from the Wikipedia category graph, weighted by term frequency-inverse document frequency.

Saif et al. [2012] exploited DBpedia for sentiment analysis of tweets. Their approach aimed at addressing the sparsity of tweets' content by enhancing the representation of the documents with additional semantic features. For this purpose, a named entity extractor is first employed which identifies the entities and concepts within the tweets. Then the original bag-of-words representation of the tweets is augmented with additional features about concepts, and the expanded feature set used to train a Naive Bayes model on the labelled instances.

Another successful application of DBpedia has been proposed by Hulpus et al. [2013], addressing the problem of topic classification. Their approach relies on mapping the words inside the documents into DBpedia concepts. In order to achieve this, an off-the-self word sense disambiguation approach is employed. Once the best concept for each words is extracted, several graph centrality measures are applied. These assume that words which co-occur in text are likely to refer to concepts that are close in the DBpedia graph. The benefits of this approach are better corpus coverage and the ability to model very broad topic labels.

2.7 Summary

This chapter presented an introduction to the task of text classification, describing the main instances of text classification explored in this thesis. It also provided a systematic comparison of state-of-the transfer learning approaches for building adaptive text classifiers across domains. The analysis of the literature revealed that these approaches enable the creation of stable features for adaptation, making use of the sole textual content of documents.

Following the critical review of state-of-the-art approaches, this chapter presented a discussion on the available domain knowledge types and their benefits for text classification. It has been noted that knowledge sources (e.g., the biomedical Unified Medical Language System knowledge source or the Linked Open Data cloud knowledge sources) provide a broad coverage of concepts for a wide range of domains, and due their structural properties and continual updates, they are valuable resources for enhancing the representation of documents for text classification.

Inspired by the success of exploiting these knowledge sources for traditional text classification, this thesis proposes the use of domain knowledge sources and Semantic Web technologies to enrich the content of documents and thus guide the adaptation across domains.

This chapter concludes the background discussion regarding the literature on transfer learning and the usefulness of domain knowledge for text classification. The next part of this thesis presents the methodology to investigate the use of domain knowledge sources for building adaptive text classifiers, providing novel ways of performing text classification

across domains.

Methodology

Chapter 3

Requirements and Design for Adaptive Text Classification

3.1 Introduction

The previous chapter described the main problems in building adaptive text classification systems in large repositories spanning multiple domains and text types. One of the major challenges in designing such systems lies in the distributional differences between the training and test data, which arise as a consequence of the differences in *vocabulary*, *language* and *style* between domains. The second main challenge is to deal with the constantly growing size of these large repositories, making it extremely difficult to collect annotated data for every new domain and text type. In order to make sense of the information present in such large heterogeneous environments, therefore, there is a need for automated techniques which can extract and locate relevant information, and also deal with the multitude of domains. *Adaptive text classification* is a suitable approach for making sense of the information present in such large repositories by, at the same time handling, the variations between domains. It aims to classify the documents within these repositories into semantic classes, allowing an effective way to organise and explore the information.

This chapter proposes a *knowledge-driven* approach for adaptive text classification, which exploits the *data, knowledge and structure* within domain knowledge sources to support adaptive text classification approaches across different domains and text types. It first presents the requirements which adaptive TC approaches must fulfil. Following this, a unified framework is presented for TC at multiple granularity levels (corresponding to the two TC tasks studied *within-document* and *whole-document* TC), which distinct transfer learning approaches can investigate in a modular way, -exploiting annotated data from a source (KS) domain to categorise documents (or text fragments) into semantic categories.

The remainder of this chapter is structured as follows: Section 3.2 describes the requirements imposed by the environment which adaptive TC must fulfil. Section 3.3 then describes a unified approach for performing adaptive TC, providing a detailed description of the individual steps of the approach.

3.2 Requirements

As detailed previously in Section 2.4, adaptive TC techniques can be divided into two main classes: *supervised* TC approaches (encompassing domain adaptation and multi-task learning techniques), and *unsupervised* TC approaches. *Supervised* approaches require a sufficient amount of annotated data from a source domain for classifying documents in the target domain - corresponding to the initial prior knowledge about the TC task. In contrast, *unsupervised* approaches do not rely on any annotated data from the source and target domains. In this case the common and domain-specific patterns are identified by discovering groupings of instances in the two domains (using for example *k-means* clustering). Both approaches have their advantages and disadvantages: *supervised* approaches may be more appropriate for related domain pairs, but may suffer from the uncertainties and ambiguities in the supervision (or labels), while *unsupervised* approaches may be applicable for any domain pairs, but require further parameters to be set (e.g., the number of clusters k). Previous research has also demonstrated that for most TC tasks *supervised* approaches achieve superior results to *unsupervised* approaches [Pan and Yang, 2010]. In light of these findings, this research also focuses on the investigation of *supervised* approaches for adaptive TC. However, in order to apply these techniques in large repositories, a set of requirements needs to be identified, which the chosen adaptive TC techniques must fulfil, ensuring efficient and effective TC in these repositories.

3.2.1 Requirements for Adaptive Text Classification

The requirements for the adaptive text classification techniques are as follows:

- *Perform adaptive text classification with a minimal amount of target domain annotated data:*
  For most domains, gathering annotation for every new domain may require domain expert knowledge, which could be an expensive and laborious process. The adaptive TC system thus must be able to classify documents of new domains with minimal supervision, requiring fewer annotations than in-domain machine learning approaches. Following the provision of annotated data, the system must perform in a fully automated fashion without any human intervention, enabling efficient knowledge management in these large repositories.

- *Achieve classification accuracy comparable to supervised machine learning approaches:*
  The promise of adaptive TC systems lies in capturing the distributional differences between domains, with the goal of being easily portable to new domains and text types. In order to achieve effective and efficient text classification in new domains, therefore, these approaches must also perform better than in-domain supervised machine learning approaches, thus providing an efficient and effective alternative to them.

- *Enable the creation of pivot features from knowledge sources:*
  One of the key components contributing to the success of an adaptive text classifier is the pivot features used to bridge the lexical gap between domains. Knowledge sources present an abundant amount of semantic information for a large number of domains, serving as potential sources for the creation of pivot features for transfer

learning. However, in order to ensure that the TC systems can fully exploit the benefit of semantic information within a particular domain KS, the employed KS must be representative of the domain under investigation, providing broad coverage of its concepts. Further, the adaptive text classifier incorporating the created KS pivot features must perform better than in-domain supervised machine learning approaches.

- *Be able to predict the performance of an adaptive text classification:*
  In order to ensure that the application of text classifier is successful on a target domain, it is important to provide automated means for quantifying its performance. For this purpose domain similarity measures must be applied which provide an estimation on the usefulness of KS pivot features for the target domain. These domain similarity measures further must achieve high correlation with the performance of an adaptive TC system, providing strong evidence for its predictive power.

- *Comply with the limitations and constraints posed by real-world application scenarios*:
  The adaptive text classifier must take into account the limitations and constraints of real world scenarios, such as application domains where the formatting of documents is lost. This situation can arise in heterogeneous repositories where multiple document formats are used. Restricting the text classifier to utilise the lexical information (words) from the documents only, ignoring the formatting of the documents, allows the system to integrate with other tools used by largest majority of search engines (e.g., the FAST Enterprise Search Platform[1], which has a very large adoption in corporate environments).

3.3 Overview of Approach

This thesis proposes a unified framework for performing *supervised adaptive TC* across domains. The use of domain knowledge sources is exploited at each stage of the approach in order to provide additional labelled data for learning, to facilitate a better representation of the domains, as well as to create a set of good pivot features for transfer learning.

The main stages of the proposed *knowledge-driven* approach are summarised as follows:

1 *Content Modelling*

2 *Context Modelling*

  2.1 *Concept Enrichment*

  2.2 *Semantic Meta-graph Generation*

3 *Pivot Feature Derivation*

4 *Adaptive Text Classification*

In the first stage, *Content Modelling*, an initial feature representation is constructed for the source and target domains, making use of the content of the domains, e.g., by employing a simple BoW model. The second stage, *Context Modelling*, involves the extraction of domain concept mentions from within the source and target domains, and the generation

[1] www.fastsearch.com

of various *semantic meta-graph models* from domain knowledge sources surrounding these concepts, to provide an enhanced representation of the domains. The third stage, *Pivot Feature Derivation*, exploits these semantic meta-graph models and identifies a set of pivot features for transfer learning. The final step *Adaptive Text Classification* then applies various semantic augmentation strategies for incorporating these pivot features into a transfer learning classifier.

In the following subsections, the various stages of this modular framework are described in more detail.

3.3.1 Content Modelling

This initial step of the framework deals with the formal representation of the domain content, serving as input to the *supervised adaptive text classifier*. In order to enable supervised transfer learning, firstly, annotated data from a source domain needs to be provided. The collection of such annotated data, however, may impose difficulties for certain repositories, for example social media platforms, due to the high topical diversity and rapid publication rate of the documents. In such cases, data from the Linked Open Data (LOD) cloud is leveraged. For instance, DBpedia and Freebase LOD KSs contain a large number of documents (articles) on a range of different domains. These articles are exported and used as annotated source domain data for learning.

Once the annotated data needed for training a supervised transfer learning classifier is compiled, the next step consists of constructing an initial feature representation of the domains comprising various state-of-the-art features used for the studied TC tasks. For instance, for the *within-document content zoning* tasks, lexical features are considered [Hirohata et al., 2008], while for the *topic classification* task, a combination of lexical, syntactic and semantic features are considered [Muñoz García et al., 2011].

3.3.2 Context Modelling

The second step of the framework aims to enrich the lexical representation of the domains, by exploiting the contextual information about concepts present in the documents. This stage comprises two main tasks: *concept enrichment* employing various online concept extraction tools, and *semantic meta-graph construction* exploiting the semantic structures surrounding concepts in KSs.

3.3.2.1 Concept Enrichment

Entity or concept extraction from documents has been long researched in the Natural Language Processing (NLP) and Semantic Web (SW) communities. In the NLP community, various named entity recognition challenges have been organised over the years, such as the Message Understanding Conferences (MUC) [Grishman and Sundheim, 1996], Automatic Content Extraction (ACE)[2], CoNLL[3] challenges for the *news-wire domain*, or the BioCreAtIvE[4] and BioNLP Shared Tasks[5] for the *biomedical domain*, resulting in the development

[2] http://www.itl.nist.gov/iad/mig//tests/ace/
[3] http://ifarm.nl/signll/conll/
[4] http://biocreative.sourceforge.net/
[5] https://sites.google.com/site/bionlpst/

of various entity extraction systems. The majority of these systems aimed to recognise a small set of named entities specific to a domain, e.g., person (PER), location (LOC) or organisation (ORG) in the newswire domain; or gene and protein names in the biomedical domain. The SW community has also contributed to the development of entity extractors, providing web services (OpenCalais[6] and Alchemy[7]), which in addition to extracting the entity mentions in the documents, also assign a unique URI to them according to different LOD KS ontologies. Recent evaluation campaigns, comparing the performance of state-of-the-art entity extractors showed relatively high performance in F-measure, achieving accuracy between 71% and 91% on well formatted news documents [Rizzo and Troncy, 2011], and a slightly lower accuracy on informal social media posts, in the range of 59%-70.6% [Cano et al., 2013; Cano Basave et al., 2014]. As an example, consider the tweet presented in Figure 3.1, highlighting the main entities extracted together with their DBpedia URI obtained using OpenCalais API.

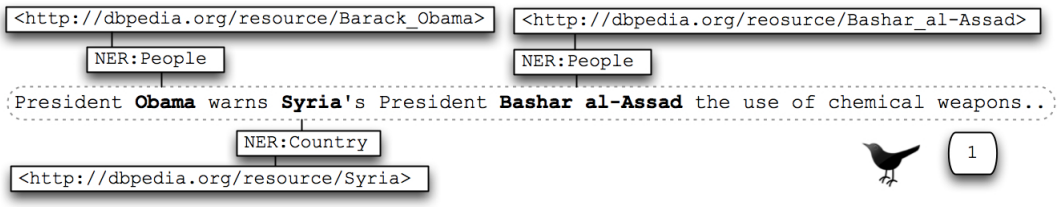

Figure 3.1: Entities and concepts extracted using OpenCalais API.

The corresponding *concept enrichment document model* is described in Definition 3.

Definition 3 (Concept Enrichment) *Given a domain* $D = (F(W), P(X))$*, consisting of a set of documents* $X = \{d_1, \ldots, d_n\}$*, and a domain ontology* $O = (C, P, R, T, Y, A)$*, a concept enriched representation of the domain is* $D_{enriched} = (F(W \cup C), P(X))$*, where* W *stands for the initial lexical features of the domain,* $C = \{c_1, \ldots, c_m\}$ *denotes a set of unique concept types (classes) extracted from the domain, and* c_k *refers to the unique URI assigned from* `R`*.*

3.3.2.2 Semantic Meta-Graph Generation

Following the process of recognition of domain-specific concepts within the documents, *semantic meta-graphs* are created from KSs exploiting the semantic information about concepts. The main role of these semantic meta-graphs is to provide a new conceptual representation of the domains, which can bridge the lexical gap between domains. A formal definition of a semantic meta-graph is given as follows:

Definition 4 (Semantic Meta-Graph) *Given a domain* $D_{enriched} = (F(W \cup C), P(X))$ *enhanced with concepts, and a domain ontology* $O = (C, P, R, T, Y, A)$*, a semantic meta-graph (semantic concept graph) is a sequence of tuples* $\mathtt{G} := (R, P, C, Y)$ *where*

- `R`, `P`, `C` *are finite sets whose elements are* `resources`, `properties` *and* `classes` *;*

[6] http://www.opencalais.com
[7] http://alchemyapi.com/

- $\mathtt{Y}$ *is the ternary relation* $\mathtt{Y} \subseteq \mathtt{R} \times \mathtt{P} \times \mathtt{C}$ *representing a hypergraph with ternary edges. The hypergraph of a Semantic Meta Graph* $\mathtt{Y}$ *is defined as a tripartite graph* $\mathtt{H}(\mathtt{Y}) = \langle \mathtt{V}, \mathtt{D} \rangle$ *where the vertices are* $\mathtt{V} = \mathtt{R} \cup \mathtt{P} \cup \mathtt{C}$*, and the edges are:* $\mathtt{D} = \{\{r, p, c\} \mid (r, p, c) \in Y\}$.

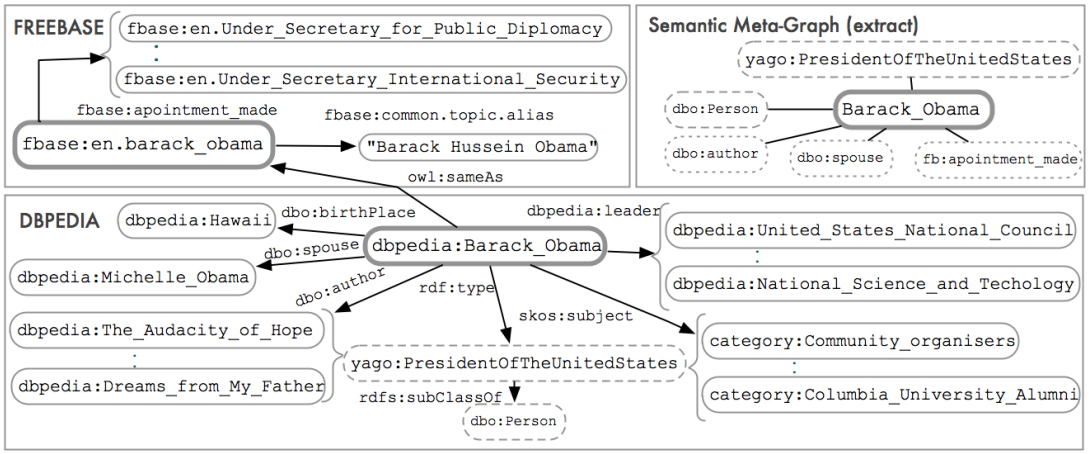

Figure 3.2: Example semantic meta-graph constructed from the DBpedia and Freebase knowledge sources about the entity "Barack Obama".

This definition will be used to define different variants of the semantic graphs explored for the individual TC tasks and domain ontologies in Chapter 5 and Chapter 6.

3.3.3 Pivot Feature Derivation

The next step of the framework makes use of the generated semantic concept graphs to identify a set of pivot features, aiming to improve the generalisation between domains. The adaptation of a TC classifier is thus guided by the knowledge and structure of the KS's surrounding concepts. The main benefits of this approach are that the generalisation (hierarchies) between concepts is explicitly defined in the ontologies, and more importantly, in contrast to previous approaches, the selection of pivot features does not require the computation of any corpus-based statistics (e.g., word frequencies) from the source and target domains.

3.3.4 Text Classification using Semantic Augmentation

The final stage of the framework aims to build a *supervised* adaptive text classifier, which makes use of the large amount of annotated data from the source domain, and a small amount of labelled data from the target domain to classify documents (or text fragments) in the target domain into semantic classes. For modelling the domain-specific and domain-independent characteristics of the domains, the derived pivot features are employed, creating a new induced semantic feature space for the domains. This new semantic space is then used to extend the original feature spaces of the domains to build the supervised adaptive text classifier. In order to perform adaptive TC at multiple granularity (documents, phrases) levels, covering different domains and text genre, different semantic augmentation techniques and KSs are exploited. These techniques are briefly described below.

3.3.4.1 Semantic Augmentation using Feature Duplication

This semantic augmentation technique employs a special feature augmentation strategy called *EasyAdapt* [Daumé, 2007] for explicitly modelling the *general* (domain-independent) and *domain-specific* features of the domains. A modified version of this approach (OntoEA) is proposed, which makes use of the cross-domain pivot features to model the general and domain-specific characteristics of the domains at a conceptual level. In this approach, the feature weights for the pivot features are assigned using simple corpus-based statistics (e.g., TF-IDF).

3.3.4.2 Semantic Augmentation using Knowledge Source Feature Weights

The main idea behind this semantic augmentation strategy is to capture the domain-specific and domain-independent weights for the cross-domain pivot features, by comparing different semantic meta-graphs generated from surrounding features in the KSs. In this case, unlike the OntoEA approach, the feature weights for the pivot features were computed solely from the KSs. This allows to capture the importance of the features from a more global perspective (exploiting the relationships between different sub-graphs in the KSs), which is only dependent on the KS structure, providing a principled way for assigning weights for both seen and unseen features.

3.4 Summary

This chapter presented a *knowledge-driven* approach to *adaptive text classification*. A modular architecture, consisting of three main steps was introduced, enabling the application of different transfer learning techniques for text classification. The first step consists of modelling the content of the domain data, creating the initial base features space for learning. Following this step, different semantic meta-graphs from KSs are exploited to create a range of cross-domain pivot features. The newly created features are then used to expand the original feature spaces of the domains. The proposed approach follows a *supervised transfer learning setting*, utilising a large amount of labelled data in the source domain, and a small amount of annotated data in the target domain to categorise document in the target domain into topics or semantic classes.

After the careful revision of state-of-the-art approaches, this chapter also enumerates the main requirements which the adaptive text classification techniques must fulfil. These requirements were imposed by the heterogeneous nature of the environment, characterised by the continuous growth of documents and text types. In order to address these challenges, Chapter 5 and Chapter 6 will present different adaptive techniques which incorporate background knowledge from KSs. Before that, however, the next chapter proposes a representation of the domains which goes beyond the simple BoW representation.

Chapter 4

Unsupervised Domain Content Modelling for Document Zoning

4.1 Introduction

For many real-world application scenarios and domains (e.g. the aerospace domain), it is very difficult to obtain the annotated data necessary for building high accuracy *supervised TC systems* for new domains. Having insufficient or no prior information about examples belonging to the predefined semantic classes raises the question of learning text classification models in purely unsupervised manner.

This chapter exploits different *unsupervised* approaches for modelling the content of the domains for the task of *within-document* TC (referred to as *document or content zoning*), aiming to recognise the information content (or zones) of *long documents*. The proposed approaches are based on probabilistic text modelling techniques (called *probabilistic graphical models*), which go beyond the typical bag-of-words representation, providing a flexible way to model the documents as a *mixture of multiple zones* (or categories). Further, these models do not rely on any annotated labelled data or domain knowledge information, which makes them practical in many applications.

The main focus of this chapter is to study and evaluate the proposed graphical models for identifying the zones of the documents in different domains. The next chapter (Chapter 5) will investigate the benefit of exploiting these models for quantifying the similarity between domains.

The remainder of the chapter is organised as follows: Section 4.2 presents the state-of-the-art approaches on *unsupervised within-document* zoning. Section 4.3 describes the proposed *graphical models* for modelling the content of the documents. In Section 4.4, the datasets used in the experiments are presented together with the novel document zone classification schema proposed for the aerospace domain. Section 4.5 then presents the experimental results obtained. Possible future extensions are further discussed in Section 4.6.

4.2 Related Work on Unsupervised Document Zoning

There has been little work on applying *unsupervised machine learning approaches* to document zone classification [Barzilay and Lee, 2004; Kagan et al., 2008; Reichart and Korhonen, 2012]. These approaches have mostly been evaluated on specific domains, such as the *newswire domain* [Barzilay and Lee, 2004] or the *biomedical domain* [Kagan et al., 2008; Reichart and Korhonen, 2012].

In the former case, Barzilay and Lee [2004] proposed an unsupervised approach for modelling the content structure of the documents, with the aim of improving two complementary tasks: information ordering and extractive summarisation. They introduced a Hidden Markov Model (HMM) to document zoning with the states corresponding to topics from the documents, and used a state-specific language model to generate the sentences relevant to the topics. They first applied clustering to compute the similarity between sentences as measured by the cosine metric and then they estimated the parameters for the HMM.

For the latter *biomedical domain*, the approaches have mostly tackled a specific *document zone* annotation schema, such as the Argumentative Zoning (AZ) annotation schema.

Kagan et al. [2008] proposed an unsupervised approach, which constructs a Huffman decision tree from the content of the full biomedical journal articles according to the AZ classification schema. Their approach aims to classify each sentence as belonging to one of the seven AZ zone categories by creating a grammatical argumentative zone structure in which the leaves of the tree correspond to the AZ zone types. In order to build the tree, different features are first constructed for each sentence, capturing the probability of mentioning a particular phrase (or expression) referring to one of the zone types, as well as the position of the sentence in a paragraph. Then the final probabilities for each features are computed using the Perron-Frobenius theory, by building a graph between the sentences and features, and computing the Perron eigenvalues, serving as the final argumentative tree.

Reichart and Korhonen [2012] recently proposed an unsupervised graphical model based on Markov Random Fields, a generalised version of the Conditional Random fields model, allowing the representation of a full joint distribution over its variables, to recognise the information structure of full journal articles in biomedicine. Their model allows the incorporation of different document- and corpus-level information, such as within-document discourse patterns (e.g., verb tenses, or position of the sentence within the document) and cross document sentence similarity (according to various features such as POS, location of the sentence within a document, and words that appear as verb objects in the sentence), showing superior results to other clustering algorithms such as k-means clustering and hidden topic markov model topic models [Gruber et al., 2007].

Despite of the success of these models, current approaches still suffer from some limitations. Firstly, these approaches have been evaluated on well written scientific documents. The application of *unsupervised* approaches to *technical domains*, such as the aerospace domain, has also not been addressed yet. These domains can pose additional difficulties for a document zone classifier due to the intrinsic complexity of the language in them [Butters and Ciravegna, 2008]. Secondly, the majority of the approaches [Barzilay and Lee, 2004; Kagan et al., 2008] take into account the order of the sentences when discovering the zones of the documents. This can, however, restrict the possibility of flexibly modelling the content of the documents, and account for the heterogeneous nature (reflected in variations in the

format and style) of the documents in large repositories.

In order to address these limitations, this thesis proposes a generic approach for modelling the content of the domain documents which is based on probabilistic graphical models.

4.3 Probabilistic Graphical Models for Document Zoning

This section describes two probabilistic graphical models for document zoning which are refined versions of the Latent Dirichlet Allocation (LDA) [Blei et al., 2003a] model[1]. Both models can flexibly model the zones categories of the documents by ignoring the order in which the sentences occur in them, thus allowing the zones to repeat after each other. The first model (described in Subsection 4.3.2), called *zoneLDA* aims to cluster the sentences into zone classes using only unlabelled data. An extension of zoneLDA, called *zoneLDAb* (introduced in Subsection 4.3.3), furthermore makes distinction between common words and non-common words within the different zone types when clustering sentences into zone categories.

Before presenting these models in details, the next section (Subsection 4.3.1) starts by formally introducing the task of document zone identification.

4.3.1 Problem Setting

The role of document zoning is to identify certain portions of documents which can play important roles in the text and to assign a semantic class to them. There have been various document zone annotation schemas introduced in the literature for different genre and TC tasks (as discussed in Subsubsection 2.2.1.1).

For the purpose of describing the notations and graphical models, the widely used 5-way Abstract-Introduction-Method-Results-Discussion (IMRAD) classification schema will be employed, where $Z \in \{\text{Introduction, Methods, Results, Abstract, Discussion}\}$ and $|Z| = N_Z = 5$. Example documents in the biomedical and aerospace domains annotated with different zone types are shown in Figure 4.1 and Figure 4.2.

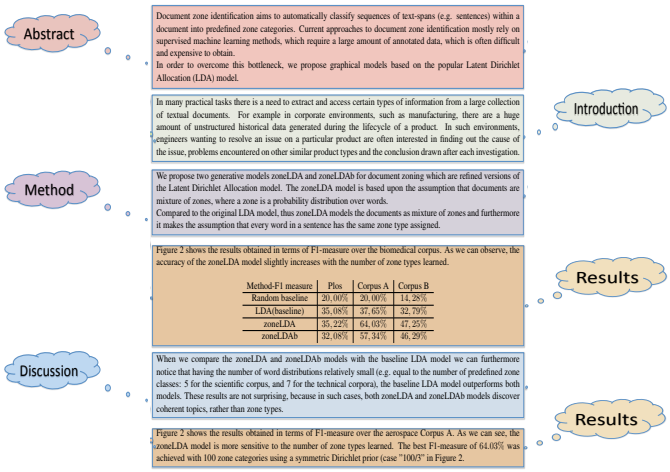

Figure 4.1: Example biomedical document annotated with zones.

[1] An introduction on probabilistic graphical model can be found in Section A.1.

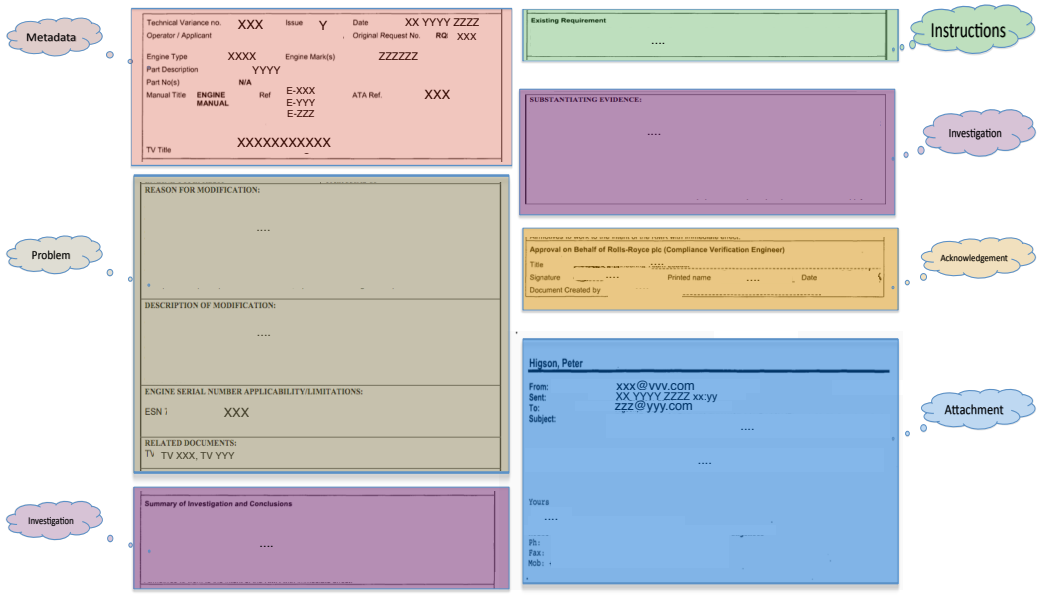

Figure 4.2: Example aerospace document annotated with zones. (Please note that this diagram describes structure, not the content of an actual report).

Formally, let us assume a domain D consisting of a set of documents $D = \{d_1, \ldots, d_{N_d}\}$, where each document d_i in D is decomposed into a sequence of sentences of size N_s denoted by $d_i = \{s_{i,1}, \ldots, s_{i,N_{s_i}}\}$, and each sentence contains a sequence of $N_{s_{i,j},n}$ words (in the more general case a sequence of *n-grams*) $s_{i,j} = \{w_1, \ldots, w_{N_{s_{i,j},n}}\}$, where the words are taken from the vocabulary V.

The goal of document zoning is then to learn the parameters of the models from a set of *unlabelled training* documents $U_S = \{(d_{S_1}), \ldots, (d_{S_n}) | d_i \in D\}$, and perform inference on the *unlabelled test* documents $U_T = \{(d_{T_1}), \ldots, (d_{T_n}) | d_i \in D\}$ by assigning to each sentence $s_{i,j}$ in document D_i a zone category $z \in Z$.

4.3.2 zoneLDA model

This subsection describes the first LDA-based model, *zoneLDA*, depicted in Figure 4.3, which assumes that the domain documents are a mixture of zones, where a zone is a probability distribution over words. In contrast to the original LDA model, zoneLDA thus models the documents as mixture of zones, rather than as a mixture of topics, and furthermore it makes the assumption that every word in a sentence has the same zone type assigned.

The generative process of *zoneLDA* (as shown in Algorithm 1) can be viewed as a procedure describing how documents are written based on the available zone types Z. That is, first, the distribution over the mixture of zones (θ^d) is chosen for the document. Then, for each sentence a zone type is randomly selected from the zone distribution, and the corresponding words from that sentence are generated according to the corresponding word-zone distribution (ϕ^z).

Similar to the original LDA, Gibbs sampling is applied to estimate the posterior distribution of hidden variable z for sentence i in document d:

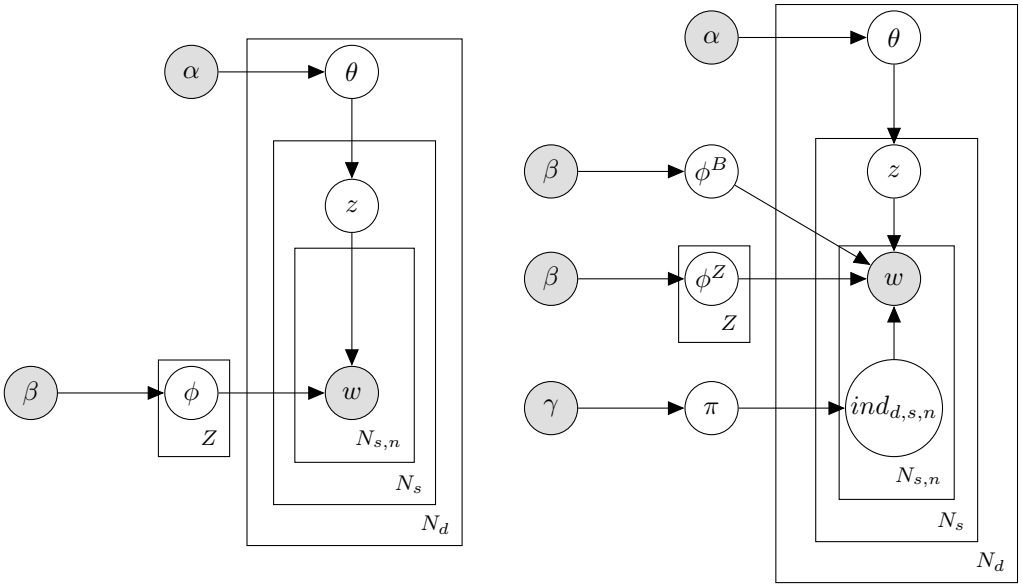

Figure 4.3: Graphical models of zoneLDA (left) and zoneLDAb models (right). The words w are observed, while the per document zone distributions θ and per zone word distributions ϕ are hidden variables.

Algorithm 1 Generative process of zoneLDA. Z denotes the number of zones, N_d denotes the number of documents, N_s denotes the number of sentences, $N_{s,n}$ denotes the number of words in sentence s, α refers to a vector for Dirichlet prior for the document zone distributions, θ^d refers to the document zone distribution for document d, $w_{d,s,n}$ denotes the word at the position n of the sentence s in document d, β refers to the word probability vectors as $Z \times V$ for the Dirichlet prior for each zone.

1: **for all** document $d = \{1, \ldots, N_d\}$ **do**
2: **draw** $\theta^d \sim \text{Dir}(\alpha)$
3: **for all** zone type $z = \{1, \ldots, Z\}$ **do**
4: **draw** $\phi^z \sim \text{Dir}(\beta)$
5: **for all** sentence $z_{d,s}$, where $s \in \{1, \ldots, N_s\}$ **do**
6: **draw** a zone class $z_{d,s} \sim \text{Multinomial}(\theta^d)$
7: **for all** word $w_{d,s,n}$ **do**
8: **draw** $w_{d,s,n} \sim \text{Multinomial}(\phi^{z_{d,s}})$
9: When running the model with the number of zone types(Z) greater than the number of predefined zone classes, perform *k-means* clustering with distributions of words as features to obtain $|N_z|$ of zone categories

$$P(z_{d,i} = k | z_{-d,i}, w) \propto \frac{n^k_{d,-i,\cdot} + \alpha_k}{n_{d,\cdot,\cdot} + Z\alpha_k} \frac{\sum_{v=1}^{V}(n^k_{d,i,v}) + \beta}{n^k_{\cdot,\cdot,v} + V\beta},$$

where $n^k_{d,-i,\cdot}$ denotes the number of sentences assigned to zone k for document d, $n_{d,\cdot,\cdot}$ denotes total number of zone types assigned to document d, $n^k_{\cdot,\cdot,v}$ denotes the number of times word v is assigned to zone k, and $n^k_{d,i,v}$ is the number of times word v from sentence i of document d is assigned to zone k.

The last step in the zoneLDA process is then to reduce the number of zone types (Z) to the number of predefined zone classes (N_Z). In order to achieve this, the k-means clustering algorithm [MacQueen, 1967] is employed over the word distributions of each zone. K-means partitions zones such that each zone belongs to a cluster with the nearest mean. For com-

puting the distance between the zones, k-means uses the Euclidean distance between the distributions.

4.3.3 zoneLDAb model

The second LDA-based model, *zoneLDAb* model (presented in Figure 4.3), is an extended version of the *zoneLDA*. The main extension compared to the *zoneLDA* model is that *zoneLDAb* tries to distinguish between common words and non-common words within the different zone types. For this reason, an additional zone cluster called *background zone* is created. The background zone cluster contains common words (e.g. "use", "determine", "indicate", "cell") present in multiple zone types, allowing the other zone clusters to capture only zone specific words. The intuition behind this approach is that multiple zone categories are likely to introduce noise (e.g. zone types with incoherent words), and thus those words are not discriminative for a zone category.

As described in Algorithm 2, the generative process of zoneLDAb also differs from that of zoneLDA, in that for each sentence a word distribution is chosen either from the background zone distribution or a selected zone distribution. In zoneLDAb thus it is necessary to infer the zone distribution for each document (θ^d), the word distributions for each zone type (θ^z) and the word distributions for the background words (θ^B). Furthermore, the π variable has the role of determining whether a word is a background word or a zone specific word.

Algorithm 2 Generative process of zoneLDAb. Z denotes the number of zones, N_d denotes the number of documents, $N_{s,n}$ denotes the number of words in sentence s, α refers to a vector for Dirichlet prior for the document zone distributions, θ^d refers to the document zone distribution for document d, $w_{d,s,n}$ denotes the word at the position n of the sentence s in document d, β refers to the word probability vectors as $Z \times V$ for the Dirichlet prior for each zone, $ind_{d,n,s}$ indicates whether a word is a background word or not

1: **draw** $\phi^B \sim \text{Dir}(\beta)$, $\pi \sim \text{Dir}(\gamma)$
2: **for all** zone type $z = \{1, \ldots, Z\}$ **do**
3: **draw** $\phi^z \sim \text{Dir}(\beta)$
4: **for all** document $d = \{1, \ldots, N_d\}$ **do**
5: **draw** $\theta^d \sim \text{Dir}(\alpha)$
6: **for all** sentence $z_{d,s}$, where $s \in \{1, \ldots, N_s\}$ **do**
7: **draw** a zone class $z_{d,s} \sim \text{Multinomial}(\theta^d)$
8: **for all** word $w_{d,s,n}$ **do**
9: **draw** indicator $ind_{d,s,n} \sim \text{Multinomial}(\pi)$
10: **draw** word $w_{d,s,n} \sim \text{Multinomial}(\phi^B)$ if $ind_{d,s,n} = 0$ and $w_{d,s,n} \sim \text{Multinomial}(\phi^{z_{d,s}})$ if $ind_{d,s,n} = 1$
11: When running the model with the number of zone types(Z) greater than the number of predefined zone classes, perform *k-means* clustering with distributions of words as features to obtain $|N_z|$ of zone categories

To estimate the posterior distribution of the hidden variable z for sentence i in document d, again Gibbs sampling is employed:

$$P(z_{d,i} = k | z_{-d,i}, w, ind)) \propto \frac{n^k_{d,-i,\cdot} + \alpha_k}{n_{d,\cdot,\cdot} + Z\alpha_k}$$

$$\times \frac{\Gamma(n^k_{\cdot,\cdot,\cdot} + V\beta)}{\Gamma(n^k_{\cdot,\cdot,\cdot} + n^k_{d,i,\cdot} + V\beta)} \prod_{v=1}^{V} \frac{\Gamma(n^k_{\cdot,\cdot,v} + n^k_{d,i,v} + \beta)}{\Gamma(n^k_{\cdot,\cdot,v} + \beta)},$$

where $n^k_{d,-i,\cdot}$ denotes the number of sentences assigned to zone k for document d, $n_{d,\cdot,\cdot}$ denotes total number of zone types assigned to document d, $n^k_{\cdot,\cdot,\cdot}$ denotes the number of times any word is assigned to zone k, $n^k_{d,i,\cdot}$ is the number of times any word from sentence i of document d is assigned to zone k, $n^k_{\cdot,\cdot,v}$ denotes the number of times word v is assigned to zone k, and $n^k_{d,i,v}$ is the number of times word v from sentence i of document d is assigned to zone k.

4.4 Compiling a Gold Standard

In order to assess the effectiveness of the proposed graphical models, two different domains are considered: the *scientific domain* and the *technical domain*. For the former domain, the *biomedical domain* is considered, due to the availability of a large collection of *biomedical journal* articles in on-line database repositories such as Pubmed[2]. For the *technical domain*, the *aerospace domain* is considered, having access to a collection of technical reports collected as part of the SAMULET research project, which provided partial funding for this research[3]. Considering the lack of annotated (multi-domain) corpora publicly available for both domains, the following subsections (Subsection 4.4.1, Subsection 4.4.2) describe the process of compiling and annotating these corpora.

4.4.1 Constructing a Corpus for the Scientific Domain

The biomedical domain has long been researched for *document zoning*, using annotation schemas such as *AZ*, or *IMRAD* [Agarwal and Yu, 2009], however, most of these corpora are small in size (less than 100 documents), or aren't publicly available.

In order to establish a gold standard of large size for the IMRAD annotation schema, the Pubmed repository was crawled, which stores a large amount of biomedical journal articles indexed by different biomedical journals. From this collection, the PLoS Pathogens journal[4] was selected, due to its wide coverage spanning multiple years.

From this journal, a total of 1,106 articles were selected, being published between January 2006 and June 2011[5]. The selection criteria was that each article had to contain all five zone categories of IMRAD.

In the pre-processing stage, further, some of the zone categories were also ignored, for e.g "References", "Supporting Information", "Synopsis", etc., as they do not belong to the IMRAD categorisation schema. In addition, other zone names which would give away valuable information such as figures, the text in tables or captions were also eliminated.

Additional pre-processing steps aiming to reduce the sparsity of the data include: removal of numbers, words made out of special characters, citations, references; stemming (using Porter stemming); elimination of sentences containing less than 5 words and words which occur in less then 10 documents; removal of stop words and one-character words. The final corpus, thus consists of only stemmed content words that are not particularly document-specific, a typical procedure when training topic models. Following the execution

[2]http://www.ncbi.nlm.nih.gov/pubmed

[3]SAMULET research project was funded by Rolls-Royce and the Technology Strategy Board.

[4]http://www.ncbi.nlm.nih.gov/pmc/journals/349/

[5]The IDs of the Plos journals articles can be found at https://sites.google.com/site/missandreavarga/resources/document-zoning

of the pre-processing steps, the size of the vocabulary was reduced from the initial 46,698 words to 6,843 words. For downloading the journal articles, the Pubmed Central Open Access API[6] was accessed, following which the articles were annotated with the different zone types by executing a Python script written for this purpose[7].

The average number of sentences for the zones in the corpus is presented in Table 4.1.

Abstract	Introduction	Methods	Results	Discussion
11	27	64	88	53

Table 4.1: Average number of sentences for each IMRAD zone in the PLOS journal corpus.

4.4.2 Constructing a Corpus for the Technical Domain and a Novel Document Zone Annotation Schema

The aerospace domain is a new area of research for *document zoning*. Within the context of the SAMULET research project, a collection of four different corpora, referred to as Corpus A, Corpus B, Corpus C and Corpus D[8], have been compiled and manually analysed with the purpose of understanding and identifying the commonalities between them, which is necessary when building automatic document zone classifiers. These corpora contain several unstructured and semi-structured PDF documents, containing a mixture of natural language sentences, images and tables. They were written at different stages of an investigation process, which is typically initiated by a customer raising a request regarding an issue about a particular engine.

The corpus analysis revealed that the different corpora are related as they all discuss various service and maintenance operations conducted on an engine due to some issues or problems which occurred on that engine. *Six zone types* were identified, which are common in the reports as they follow a *problem-solving* perspective of an investigation. For example, these reports typically contain the *Metadata* zone, which introduces the main entities (e.g., engine, component) under investigation, then they continue with a *Problem description* zone, which describes the problems that occurred on these entities; next the *Instructions* zone gives typical instruction regarding what procedure should be taken, finally the *Decision* contains the decision taken after investigation. In addition, two other zones were found in these reports: the *Acknowledgement* zone, which is the formal part of the document, and the *Attachment* zone, which includes information about the further evidence taken during investigation (e.g., in the forms of images). These zone categories are also summarised in Table 4.2.

Comparing the proposed annotation schema to the existing zone classification schemas from the literature, it can be noted that some of the zone categories are unique to the proposed schema, and others share some similarities with the existing ones. For instance, the new zone categories are the *Instructions* and *Acknowledgement* zones, as they are not typical

[6] http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

[7] The Python script can also be found at https://sites.google.com/site/missandreavarga/resources/document-zoning.

[8] As the analysed documents belong to Rolls-Royce, due to confidentiality reasons these corpora have been anonymised.

Zone Category	Description
1. *Metadata* (Meta)	Contains general information about the report: the title of the report, the entities (e.g., engine, component) under investigation, and other entities (e.g., agents) participating in the investigation
2. *Problem description* (Prob)	Aims to describe the problem encountered on a specific entity (e.g., engine, component)
3. *Decision* (Inv)	Summarises the decision taken after investigation (e.g., the conclusion drawn)
4. *Instructions* (Ins)	Describes the general procedure to follow in a certain situation (e.g., a given problem)
5. *Acknowledgement* (Ack)	Denotes the formal part of the document, consisting of the details of the agents (e.g. name) and their signature
6. *Attachment* (Attach)	Contains further evidence attached to the investigation (mostly pictures, email, faxes)

Table 4.2: Proposed meta-knowledge annotation schema.

in scientific articles. The *Instructions* zone aims to describe the step-by-step procedure which should be followed to solve the problem or complete a procedure. These instructions are typically split into tasks and subtasks, and may make reference to some manuals (e.g., the engine manuals). The *Acknowledgement* zone also typical to these reports acknowledges the conclusions drawn from the report, and contains the signatures of the responsible agents.

The remaining categories share some commonalities with the zone categories already proposed in the literature, e.g., with the 4-way (OMRC) sentence classification schema [Lin et al., 2006]. The *Metadata* zone maps to the Metadata information of the research articles, (e.g., title of the article, authors of the article). The *Problem description* zone corresponds to the *Objective* zone category. The *Decision* zone serves the role of *Result* and *Conclusion* zones. The *Attachment* zone maps to the figures, and images describing the experiments in the articles. Furthermore, in this context it may also consists of other documents (such as emails, faxes) attached in support of the evidence collected during investigation.

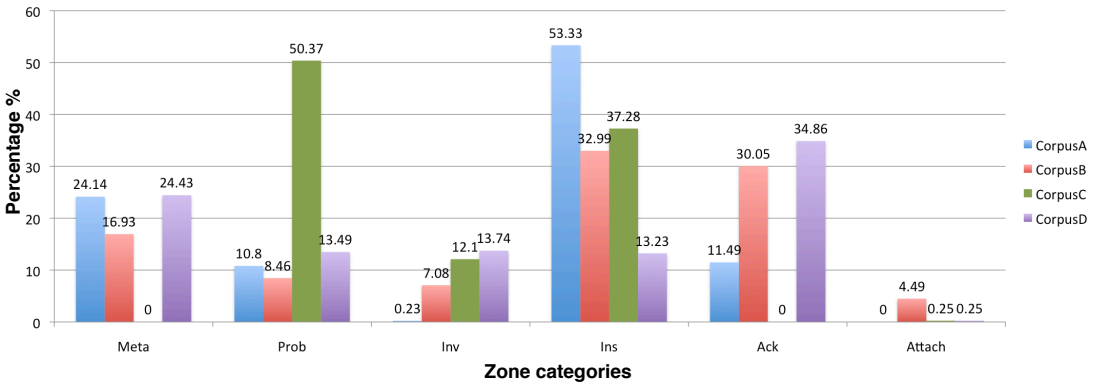

Figure 4.4: Distribution of zone categories across the different corpora.

Figure 4.4 shows the percentage of the zone types in each of the corpus. As can be observed, the majority of the zone categories (such as *Problem description*, *Decision*, *Instructions*) appear in each corpus. However, the *Metadata* and *Acknowledgement* zones are

only present in *Corpus A*, *Corpus C* and *Corpus D*.

Corpus name	*Meta*	*Prob*	*Inv*	*Ins*	*Ack*	*Attach*
Corpus A	36	3	22	76	17	35
Corpus B	30	49	13	107	13	25

Table 4.3: Average number of lines for each proposed zone category in the technical corpora.

Two of these corpora (Corpus A and Corpus B) were selected for evaluating the proposed graphical models. Two independent annotators annotated them, achieving an inter annotator Kappa agreement of 85%. The most common mistake made by the annotators was to omit several sentences from the beginning and end of the zones. This mistake was mostly due to the annotators annotating the plain text version of the documents, thus missing the layout and formatting information, which posed difficulties for them. The average length (number of lines) of the zones in the corpora is presented in Table 4.3.

During the the pre-processing step all the PDF documents were converted into plain text[9] and thus all the formatting information and figures were removed. Similar to the biomedical corpus, all numbers and stop words were also removed. Furthermore, due to the diverse format of the documents, consisting of tables and natural language text, the smallest unit of classification was considered to be the lines of the documents, as opposed to sentences. After the execution of the pre-processing steps, including the removal of words which occur in fewer than 5 documents, the size of the vocabulary of Corpus A was reduced from 4,153 to 1,023. For Corpus B, the initial vocabulary was reduced from 2,964 words to 1,340 words.

Corpus statistics for the two domain corpora are presented in Table 4.4.

Domain	Corpus name	Number of documents	Average number of sentences per document	Average number of distinct words
scientific domain	PLOS journal articles	1,106	241 ± 59	966 ± 174
technical domain	Corpus A	317	226 ± 295	329 ± 241
	Corpus B	372	394 ± 405	518 ± 273

Table 4.4: Corpus statistics of documents in the scientific and technical domains.

4.5 Evaluation

This section presents the empirical evaluation of the proposed graphical models on *document zoning* using two different domains, the biomedical domain and the aerospace domain, comparing the performance of these models to an alternative baseline model. In the next section, the experimental set-up is presented, followed by a discussion of the results. The main research question addressed in these experiments is the following: *Can probabilistic graphical models be used to recognise the structure of the documents?*

[9]The Apache PDFBox library available at `http://pdfbox.apache.org/` was used for converting the Pdf documents into plain text format.

4.5.1 Baseline Model

Considering that the *zoneLDA* and *zoneLDAb* models are the first two graphical models for zoning, the baseline model used to evaluate their performance, is also an LDA model. This baseline LDA model relies on the original LDA model described in Blei et al. [2003a], with the modifications that at inference time, for each sentence the topics of the words are examined, and the most likely topic among the words of the sentence is considered as the zone type for that sentence.

4.5.2 Evaluation Measures

In order to evaluate the efficacy of the proposed models, the *pairwise clustering F1* measure was employed.

Pairwise clustering measures the overlap between the generated clusters and the gold standard. To compute this metric, the implementation in Mallet[10] is used. Taking into account the gold standard for each pair of sentences, this computes the false positives and false negatives in order to decide whether the pair should be in the same cluster or not:

$$\text{Prec}_{\text{pair}} = \frac{|\text{clustered sentence pairs which should be clustered}|}{|\text{sentence pairs which are clustered}|}$$

$$\text{Rec}_{\text{pair}} = \frac{|\text{clustered sentence pairs which should be clustered}|}{|\text{sentence pairs which should be clustered}|}$$

$$\text{F1}_{\text{pair}} = 2 \times \frac{\text{Prec}_{\text{pair}} \times \text{Rec}_{\text{pair}}}{\text{Prec}_{\text{pair}} + \text{Rec}_{\text{pair}}}$$

4.5.3 Experimental Set-up

When evaluating the models in the scientific domain, the experiments were repeated 10 times using 60%-10%-30% split, considering 60% of the original corpus as training, 10% as development, and the remaining 30% as testing. For the case of technical domain, the original Corpus A and Corpus B were split into 45% training, 10% development and 45% testing. In each case the performance of the unsupervised graphical models was compared to the baseline LDA model on the same held-out test data. In particular, for all the graphical models, Gibbs sampling was run for 10,000 number of iterations and a burn-in of 500.

[10] http://mallet.cs.umass.edu/

4.5.3.1 Hyper-parameter Setting

Given that the values of the parameters can have different effects on the performance of the proposed models, both *zoneLDA* and *zoneLDAb* models were evaluated with different values for their parameters. First, the number of the word distributions were varied, considering $Z \in \{5, 50, 100\}$ for the biomedical domain, and $Z \in \{7, 50, 100\}$ for the aerospace domain. Furthermore, for varying the number of zone types both asymmetric and symmetric values were tried for α. This lead to a number of six different values for α.

In the first case, a symmetric Dirichlet prior is chosen with $\alpha = 0.1$, which discovers zone types that are sparse. In the second case, a symmetric Dirichlet prior is chosen with $\alpha = 1$. In the third case, a symmetric Dirichlet prior is chosen with $\alpha = 10$, which discovers zone types which are dense.

When considering asymmetric Dirichlet priors, the α values were initialised such that the $\alpha_i, i \in \{1, \ldots, Z\}$ values for the different zone categories were based on the development set. As such, in the fourth setting, the $\alpha_i, i \in \{1, \ldots, Z\}$ were set such that $\sum_{i \in \{1,\ldots,N_Z\}} \alpha_i = 0.1$. In the fifth case the $\alpha_i, i \in \{1, \ldots, Z\}$ were set such that $\sum_{i \in \{1,\ldots,N_Z\}} \alpha_i = 1$. And finally, in the sixth case $\alpha_i, i \in \{1, \ldots, Z\}$ were set such that $\sum_{i \in \{1,\ldots,N_Z\}} \alpha_i = 10$. For the β parameter the same value of 0.01 was assigned in each case.

4.5.4 Results and Discussion

This subsection evaluates the proposed graphical models according to the measure introduced in Subsection 4.5.2. The results obtained for the three corpora using F1 pairwise clustering measure are presented in the Figure 4.5, Figure 4.6 and Figure 4.7.

Starting with the results obtained for the *biomedical domain*, shown in Figure 4.5, it can be observed that the accuracy of the zoneLDA model slightly increases with the number of zone types learned. The zoneLDA model achieves an F1-measure over 30% for 50 and 100, having the best F1-measure of 35.22% for 50 zone types with an asymmetric Dirichlet prior (case "50/6" in Figure 4.5). In contrast, when looking at the results of the zoneLDAb model, the improvement obtained with different number of zone types is less significant. The best F1-measure of 32.08% is achieved with 5 zone types and an asymmetric Dirichlet prior (case "5/6" in Figure 4.5). Compared to the baseline LDA model it can be seen, that with a small number of number of zone types (e.g., 5), the baseline LDA model outperforms both zoneLDA and zoneLDAb models, but as the number of zone types increases the performance of the zoneLDA model becomes superior in most cases.

An explanation for the relatively low performance of these LDA based models can be that the selected PLOS journal articles cover a large range of sub-topics, exhibiting a large variations in vocabulary and topic, making the recognition of zones a very challenging task. Examining the errors made by the zoneLDA model, it can be noted, that the most difficult zone to identify was the Abstract zone, for which the zoneLDA model achieved an F1-measure of 1%. The second most difficult zone type was the Introduction zone, for which the performance was 8%. For the Discussion zone type the F1-measure was 21.4%, for the Methods zone 30.00% and for the Results zone 66.8%. Similar trends can be observed for the zoneLDAb model. The Abstract zone type still being the most difficult zone type to be discovered with an F1-measure of 2%. For the Introduction the zoneLDAb model achieved an F1-measure of 13.1%, and for the Discussion zone type it achieved an F1 of 25.1%. The

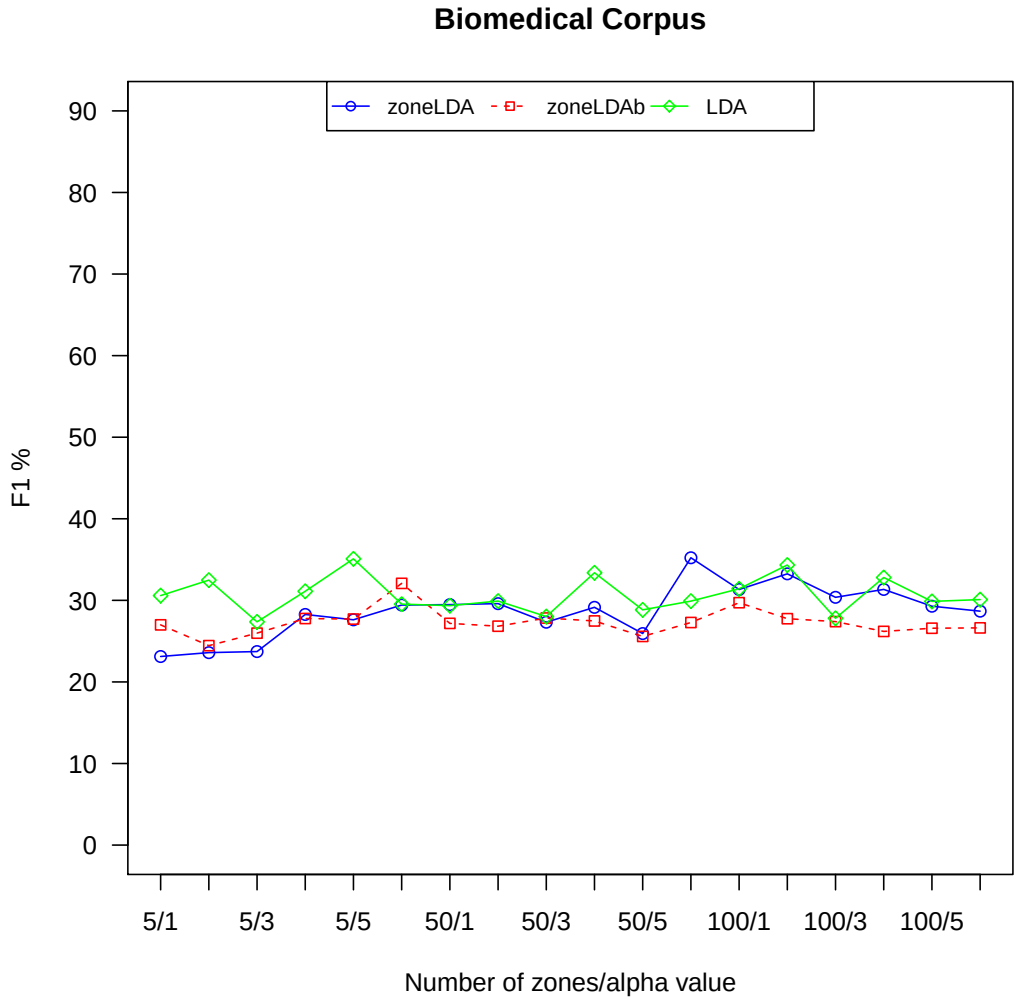

Figure 4.5: The performance of zoneLDA and zoneLDAb models over the biomedical corpus varying the number of zone types/alpha values. $Z \in \{5, 50, 100\}$, "/1" case denotes $\alpha = 0.1$, "/2" denotes α=1, "/3" denotes $\alpha = 10$, "/4" denotes $\sum_{i\in\{1,\ldots,N_Z\}} \alpha_i = 0.1$, "/5" denotes $\sum_{i\in\{1,\ldots,N_Z\}} \alpha_i = 1$, "/6" denotes $\sum_{i\in\{1,\ldots,N_Z\}} \alpha_i = 10$. The different values are connected by lines to help readability.

two best performances were achieved for the Methods zone type, with an F1-measure of 38.9%, and for the Results zone type an F1 of 41.5%.

When examining the results obtained by the baseline LDA model, however, the results look different. The worst results were obtained for the Discussion zone, with an F1-measure of 2%, the Methods zone type, with an F1-measure of 3.4%, and the Abstract zone type, an F1-measure of 5.6%. For the Introduction, the baseline model an F1-measure of 34.8%, and for the Results zone type an F1-measure of 43.2% was achieved.

The results obtained for aerospace Corpus A are presented in Figure 4.6. As can be seen, the zoneLDA model is more sensitive to the number of zone types learned. The best F1-measure of 64.03% was achieved with 100 zone categories using a symmetric Dirichlet prior (case "100/3" in Figure 4.6). Similar trends can be seen for the zoneLDAb model,

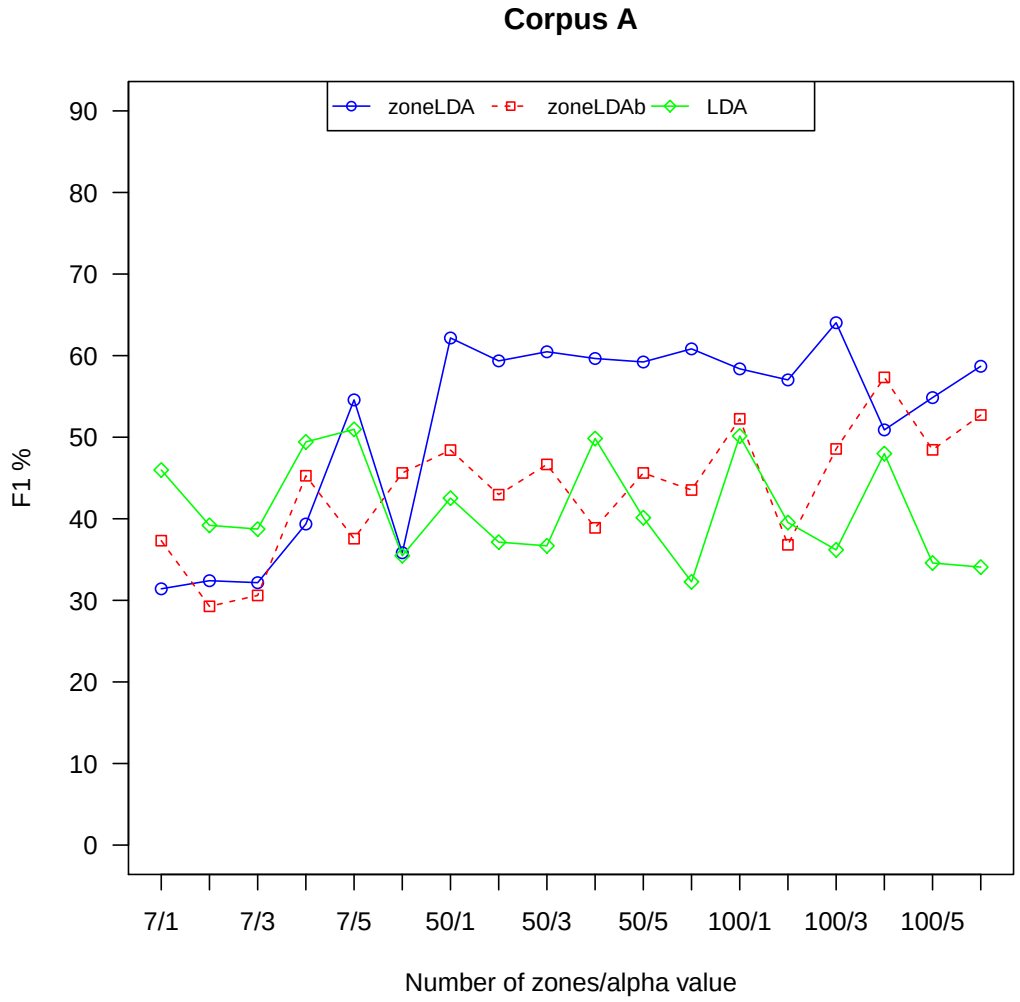

Figure 4.6: The performance of zoneLDA and zoneLDAb models over Corpus A varying the number of zone types/alpha values. $Z \in \{7, 50, 100\}$, "/1" case denotes $\alpha = 0.1$, "/2" denotes α=1, "/3" denotes $\alpha = 10$, "/4" denotes $\sum_{i \in \{1,\ldots,N_Z\}} \alpha_i = 0.1$, "/5" denotes $\sum_{i \in \{1,\ldots,N_Z\}} \alpha_i = 1$, "/6" denotes $\sum_{i \in \{1,\ldots,N_Z\}} \alpha_i = 10$. The different values are connected by lines to help readability.

whose performance improves with the number of zone types learned, achieving the best F1-measure of 57.33% using 100 zone categories with an asymmetric Dirichlet prior (case "100/4" in Figure 4.6). Compared to the baseline LDA model, it can also be noted that both zoneLDA and zoneLDAb models perform consistently better when having a large number of word distributions and clustering.

Finally, Figure 4.7 shows the results obtained in terms of F1-measure over the aerospace Corpus B. The performance of the zoneLDA model slightly increases with the number of zone types learned. The best F1-measure of 47.25% was achieved using 100 zone categories with an asymmetric Dirichlet prior (case "100/3" in Figure 4.7). The performance of the zoneLDAb model is also very similar, being over 40% for all the different cases when 50 or 100 zone categories are used, with highest values of 46.29% (case "100/4" in Figure 4.7) being

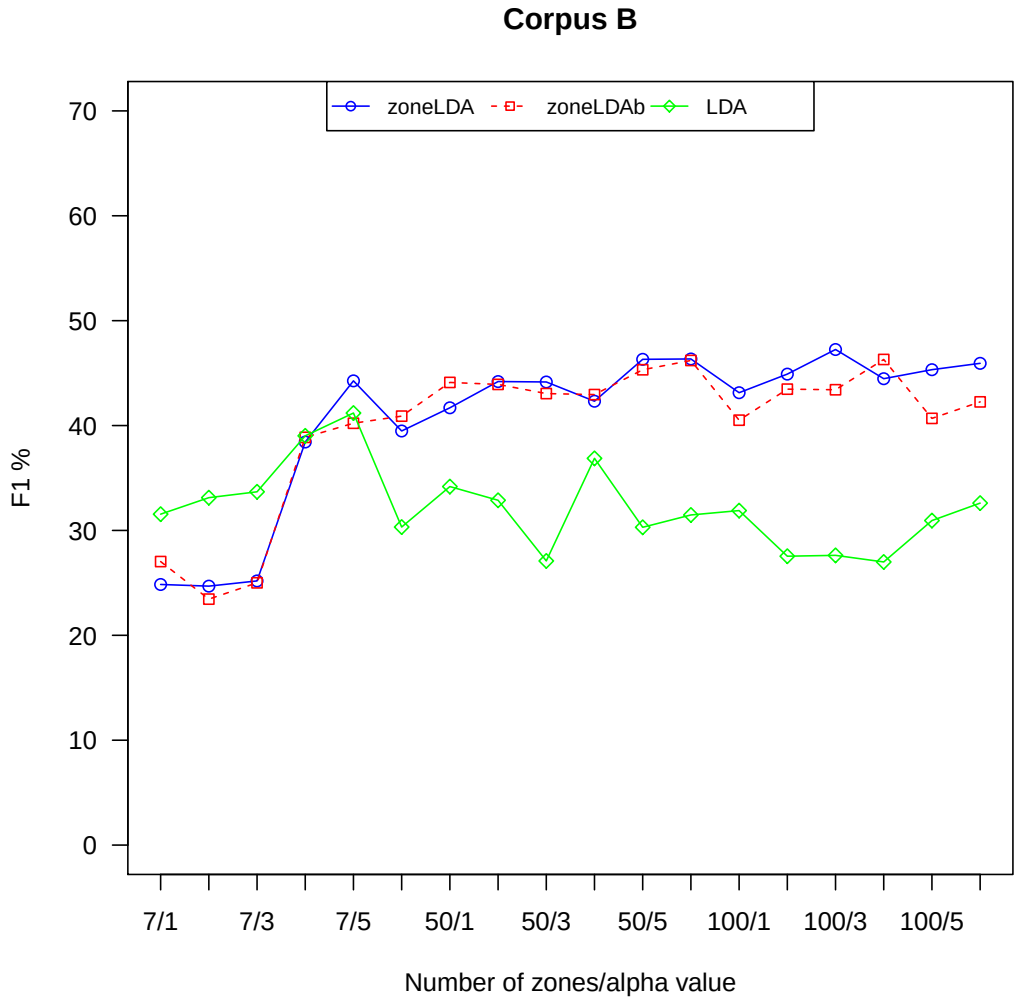

Figure 4.7: The performance of zoneLDA and zoneLDAb models over Corpus B varying the number of zone types/alpha values. $Z \in \{7, 50, 100\}$, "/1" case denotes $\alpha = 0.1$, "/2" denotes α=1, "/3" denotes $\alpha = 10$, "/4" denotes $\sum_{i\in\{1,\ldots,N_Z\}} \alpha_i = 0.1$, "/5" denotes $\sum_{i\in\{1,\ldots,N_Z\}} \alpha_i = 1$, "/6" denotes $\sum_{i\in\{1,\ldots,N_Z\}} \alpha_i = 10$. The different values are connected by lines to help readability.

obtained for 100 zone types with an asymmetric Dirichlet prior. Compared to the baseline LDA model, it can also be noticed that both zoneLDA and zoneLDAb models perform consistently better when having a large number of word distributions and clustering.

Overall, these results indicate that in general the performance of the zoneLDA and zoneLDAb models increases with the number of word distributions learned. For the majority of the cases, when the number of word distributions is more than 50, the zoneLDA model consistently outperforms the zoneLDAb model. This is because the background words discovered by the zoneLDAb model actually seem to contain zoning information.

When comparing the zoneLDA and zoneLDAb models with the baseline LDA model, it can furthermore be noted that when the number of word distributions relatively small (e.g. equal to the number of predefined zone classes: 5 for the scientific corpus, and 7 for the

technical corpora), the baseline LDA model outperforms both models. These results are not surprising because in such cases, both zoneLDA and zoneLDAb models discover coherent topics rather than zone types. On the other hand, when the number of word distributions learned increases these models exhibit a significant improvement over the LDA model. In this case, the discovered word distributions are less sensitive to topic information, allowing the zone information to be found.

These findings can be summarised as follows:

1. Probabilistic graphical models are useful for discovering the structure of the documents: both *zoneLDA* and *zoneLDAb* models perform better than a baseline LDA model in terms of F1 pairwise clustering performance.

2. The performance of the graphical models depends on the value of their hyper-parameters: increasing the number of zone clusters leads to higher performance.

4.6 Possible Future Directions

The proposed *zoneLDA* and *zoneLDAb* probabilistic graphical models have several advantages. First, they are *unsupervised*, requiring no annotated data for discovering the information structure of the documents. Second, they *do not rely on any domain or language specific tools* (e.g. part-of-speech tagging) to model the zones of the documents. Third, they allow *flexible modelling of the content of the documents ignoring the order of the sentences* in them.

Indeed, these models rely only on the lexical features (words) to discover the zones in the documents, making them more practical for many real word domains (such as aerospace), where there are no domain-specific resources (e.g., domain ontologies) available, or where the coverage of the ontologies is low. Given the simplicity of these models, several possible extensions could be explored:

- *Investigating more sophisticated graphical models*:
  The main assumption underlying the graphical models presented is that the order of the sentences in the domains is exchangeable, and the problem of document zoning is purely lexical (making use of only word features to model the zones).
  However, for certain documents in the domains, the order of the sentences may provide additional clues for the identification of zone types. For instance, in *scientific articles*, the first zone type is typically the Background, which is followed by the Introduction; in the technical *aerospace domain* also the first zone type to be presented is typically the Metadata zone. There have been different variants of LDA proposed in the literature which capture the transition between the different topics [Griffiths et al., 2004; Du et al., 2012]. One possible future direction could be to investigate the usefulness of such models for zoning. Another appealing future direction could be to investigate purely *non-parametric* graphical models [Teh et al., 2006], which allows the discovery of an arbitrary number of zone types.

- *Cluster labelling*:
  The main goal of the graphical models presented is to organise the sentences within the documents into different clusters, such that sentences within the same clusters are as

similar as possible, while sentences from different clusters are dissimilar. The created clusters are then considered as zones. In some applications, e.g. content filtering, however, the label (or zone type) of the cluster would also be directly needed in order to successfully complete the task. In such cases, cluster labelling approaches could be applied [Carmel et al., 2009], which rely on employing different statistical methods to map the important words from a cluster to a KS category or ontological concept (e.g. Wikipedia).

- *Incorporating graphical models with supervised machine learning approaches*:
  The graphical models presented provide an alternative (standalone) solution for discovering the information structure of the documents in a given domain. They also create a new latent semantic space for the documents, allowing to represent their content as a mixture of zone distributions. A possible future application of these models could be to apply them in a multi-domain scenario. For e.g., by learning the zone assignments between the sentences of the source and target domains (for e.g. as already used in NER adaptation [Guo et al., 2009; Nallapati et al., 2010] or text classification [Kadar and Iria, 2011]) and then using the newly learned zone assignment features to build a supervised transfer learning classifier by augmenting to augment the original feature spaces with these features.

Finally, another promising future direction could be to incorporate knowledge from KSs into graphical models.

4.7 Summary

This chapter addressed the problem of *within-document* TC in the absence of any annotated data. For this purpose, two unsupervised graphical models were examined for modelling the content of the domains. The *zoneLDA* and *zoneLDAb* models introduced, rely on clustering the sentences (or lines) in the documents based *only* on the lexical information about words in the domains. The main advantages of these models are that: i) they do not require any annotated examples or domain information; and ii) they ignore the order of the sentences in the documents, thus providing a more flexible way to model the content of the documents.

Experimental results on both scientific and technical domains showed promising results, outperforming a strong baseline graphical model, which first discovers the topics of the words within the sentences, and then infers the most likely zone for each sentence. The results further demonstrate that the performance of these models depends on the nature of the documents. The results obtained on the aerospace technical domains, consisting of regular documents with a restricted vocabulary, were reasonably higher (achieving an F1 measure up to 64%) than the results obtained on the biomedical journal articles spanning multiple sub-domains (yielding an F1 measure of 35.22%).

The next chapter continues the discussion of *within-document* TC, considering the situation when annotated data is available from a source domain. A range of adaptive TC models are presented, which can exploit the prior background information within the source domains. A set of domain similarity measures are also presented, which makes use of the graphical models presented.

Chapter 5

Supervised Transfer Learning for Document Zoning

5.1 Introduction

The previous chapter has shown that *probabilistic graphical models* can be used to automatically generate the *within-document* segments of *long documents* in large repositories. These *unsupervised* models exploit the lexical information within the domains, with the aim of inducing a new representation of the documents in which each document is represented by a *probability distribution over zones* and the zones as a *probability distribution over words*. These models are thus of great importance in situations where one can only make use of the lexical information within domains as domain information and domain ontologies are *scarce* or under-represented, as in the case of aviation maintenance domain or car crash management domains.

However, over the years, several advancements have been made in creating and maintaining linked knowledge sources, providing abundant amounts of information and knowledge about particular domains, such as Life Science, Geographic, Publications, etc. (as described Section 2.6). The most common examples are the biomedical ontologies, for example the overarching UMLS ontology, which encapsulates a broad spectrum of the biomedical subdomains, such as human anatomy, diseases, psychology, health care and microbiology. These knowledge sources can be used to enrich the representation of documents, exploiting semantic information about the concepts discussed in them.

This chapter presents different supervised transfer learning approaches, which can exploit the semantic information from KSs for *within-document* TC. For this purpose, a semantic meta-graph is generated from KSs. This meta-graph provides novel semantic features for transfer learning, aimed at reducing the distributional differences between domains. Further, this chapter also proposes novel unsupervised domain similarity measures for predicting the performance of an adaptive document zone classifier. The proposed measures are computed based on the lexical and semantic features of the zone segments in the source and target domains, with the zone segments being discovered by unsupervised graphical models.

The remainder of this chapter is organised as follows: Section 5.2 summarises the state-of-the-art approaches to *supervised* document zoning. Section 5.3 presents an adaptive TC

framework for *within-document* TC. Section 5.4 describes a novel set of domain similarity measures for *document zoning*. Section 5.5 presents the gold standard dataset used in the experiments. Section 5.6 then evaluates the proposed adaptive text classification models and domain similarity measures on a large number of domain pairs in the scientific domain. Possible future extensions are finally discussed in Section 5.7.

5.2 Related Work on Supervised Document Zoning

The classification of document zones into semantic classes has been proved to be valuable in many NLP tasks, including Information Extraction (IE) [Mizuta and Collier, 2004], Information Retrieval (IR) [Tbahriti et al., 2005], summarisation [Teufel and Moens, 2002], [Barzilay and Lee, 2004] and question answering [Caporaso et al., 2006]. For example, IE systems may target specific zones which contain evidence-rich results as opposed to other evidence-lean zones. Summarisation systems may produce summaries for each zone separately. Question answering systems may consider a specific zone from which to extract the correct answer. IR systems may filter out irrelevant documents based on whether they contain any zone relevant to the user query.

This section provides an overview of the traditional *supervised* and *semi-supervised* approaches to document zoning, which constitute the state-of-the-art solutions to document zoning. To remind the reader, the *unsupervised* approaches were reviewed in the previous chapter (Chapter 4).

5.2.1 Supervised Learning Strategies

The first class of approaches, *supervised* approaches make use of a large number of annotated data to build a document zone classifier for a particular domain, using widely known classifiers such as Naive Bayes [Teufel and Moens, 2002], Hidden Markov Model [Li et al., 2010], Maximum Entropy [Merity et al., 2009], SVM [Guo et al., 2011a; McKnight and Srinivasan, 2003] and CRF [Hirohata et al., 2008].

Most of these approaches have been applied to scientific articles in the context of *computational linguistics* [Teufel et al., 2009; Merity et al., 2009], *biology* [Mullen et al., 2005; Miyao et al., 2006; Lin et al., 2006; Caporaso et al., 2006; Settles and Craven, 2005; Liakata, 2010; Nawaz et al., 2010; Guo et al., 2011a; Agarwal and Yu, 2009; Hirohata et al., 2008] and *chemistry* [Liakata, 2010; Teufel et al., 2009], focusing on either the full text or abstract of the articles.

Teufel and Moens [2002] showed the usefulness of document zoning in producing executive summaries of scientific articles based on the Argumentative zoning (AZ) classification schema. This approach employed the supervised Naive Bayes classifier with various sentential features for the sentences, such as location features (capturing the position of the sentence), sentence length features, verb syntax features, citation features, meta discourse features (e.g. agent, action). Taking the rhetorical status of the sentences into account, this approach enables the creation of task-oriented, user-tailored summaries. Teufel et al. [2009] further extended the AZ classification schema and showed its applicability to both life sciences and computational linguistics.

Mullen et al. [2005] employed zoning to enable pinpointing and organising the factual

information extracted from a large number of full biomedical text articles. For this purpose a supervised SVM model was employed with various lexical (n-gram) and syntactical features (part of speech). They also showed that considering the context of the extracted information enables a much better interpretation of the extracted facts.

Miyao et al. [2006] used zoning to enable accurate retrieval of relational concepts, such as associations between protein-protein or gene-disease, from a large number of biomedical Medline abstracts.

Settles and Craven [2005] proposed a two-tier approach to information retrieval based on zoning using biomedical abstracts. In the first preprocessing step they used zoning to identify the different information zones of the articles, such as title, abstract, introduction, methods, results, and discussion. Then in the second stage, they performed zone level classification to make final document-level predictions.

Caporaso et al. [2006] used automatic document zoning as a pre-processing step in a question answering system. They used zoning to filter out parts of the documents which were likely to be irrelevant to user queries, thus enabling a more efficient retrieval.

Similarly, Lin et al. [2006] showed the important role of document zoning in building a clinical system, aiming to provide access to the information essential to particular patient treatment process. Their experiments revealed that using HMM with lexical and semantic features created by the LDA graphical model achieves good results for zoning, and these results are also competitive with those obtained by an SVM model.

Liakata [2010] highlighted the importance of automatically recognising negations and speculations from chemistry and biomedical articles. They used the Core Scientific Concepts (CoreSC) classification schema to capture this information, thus allowing to distinguish between the positive and negative outcome of the results influencing the conclusions drawn after an experiment.

Nawaz et al. [2010] considered annotating complex biomedical events from the Genia corpus using the meta-knowledge annotation schema, thus allowing to better distinguish between the various ambiguous interpretation of these events.

Guo et al. [2011a] applied document zoning to cancer risk assessment, aiming to estimate the probability of developing cancer from exposure, based on biomedical abstract articles available online in Medline. In their experiments they compared the Naive Bayes, SVM and CRF models using a range of different lexical and syntactic features. Their results revealed that the SVM model with lexical (uni-gram, bi-gram) and verb features consistently outperforms the CRF and Naive bayes models.

Li et al. [2010] showed the usefulness of zoning for classifying clinical patient notes into some predefined section names, such as history of present illness (HPI), family history (FHX), past medical history (PMH), past surgery history (PSH), allergies (ALL), medications (Meds), etc., thus enabling a better understanding of the notes. This means that the information relevant to the user queries can be answered more efficiently, for example when searching for information about a given patient, the family history zone should be ignored.

Subsequent work has also focused on comparing the performance of an IR system on abstracts and complex full texts [Lin, 2009]. Experimental results on Medline abstracts and full text articles from the TREC 2007 genomics evaluation show that IR based on full text articles does not allow more effective retrieval than that performed only on abstracts. Furthermore, IR based on document zones consistently outperforms approaches performed

on abstracts. A better solution, however, consists in combining the evidence from both document zones and full text.

5.2.2 Semi-Supervised Learning Strategies

In contrast to the *supervised* approaches discussed above, the second class of approaches, *semi-supervised* approaches require only a small amount of annotated examples, which are used to incrementally learn the document zone classifier. These approaches have only started to gain attention very recently [Guo et al., 2011b; 2013].

Guo et al. [2011b] proposed a semi-supervised approach for document zoning, which employs an active SVM classifier using active learning and self-training strategies. The active SVM classifier is built on a small amount of annotated data, and iteratively queries for unlabelled data for which the SVM has less confidence about its label. Further, the self training strategy ensures that both labelled and unlabelled data are efficiently exploited by training the classifier on both labelled and unlabelled examples annotated with the current classifier until a certain level of accuracy is reached. The proposed SVM classifier was evaluated for a variety of different features: such as location (location of a sentence within the text), words, bi-grams, part-of-speech, voice of the verb, etc. Experimental results on biomedical abstracts annotated with argumentative zones showed promising results. The best classifier, utilising all the features, outperformed the supervised SVM classifier.

Guo et al. [2013] presented a semi-supervised approach which employs the Maximum Entropy model with the Generalised Expectation (GE) criterion. This model makes use of several discourse and lexical constraints on the sentences for which labelled information is provided, thereby avoiding in this way the provision of labelled information about the zones itself. The discourse constraints contain features such as the location of the sentence, while the lexical constraints include features such as citations, references to tables, lists, tenses of verbs, etc. Experimental results on biomedical journal articles using the AZ zone schema showed that using both discourse and lexical constraints, the proposed classifier outperforms supervised approaches.

The main limitation of the above models is that they still rely on *in-domain machine learning approaches*, which only perform well when the distribution of the data remains the same across domains and text types. However, in many practical scenarios it is often necessary to extract zones from significantly different corpora/domains.

In order to address these limitations, this chapter proposes the use of transfer learning for document zoning. For this purpose different transfer learning techniques are presented, which make use of KSs to bridge the gap between the domains (described in Section 5.3). In addition, a set of novel *domain similarity measures* are proposed. These measure the similarity between multiple domains, serving as a proxy for the performance of a cross-domain document zone classifier (described in Section 5.4).

5.3 Ontology-driven Adaptive Document Zone Classification

This section describes a transfer learning framework for *adaptive document zone classification* of *long documents*, which is based on the framework introduced in Section 3.3. This

framework follows the scenario in which a *large amount of annotated source domain documents* and a *small amount of annotated target domain documents* are available.

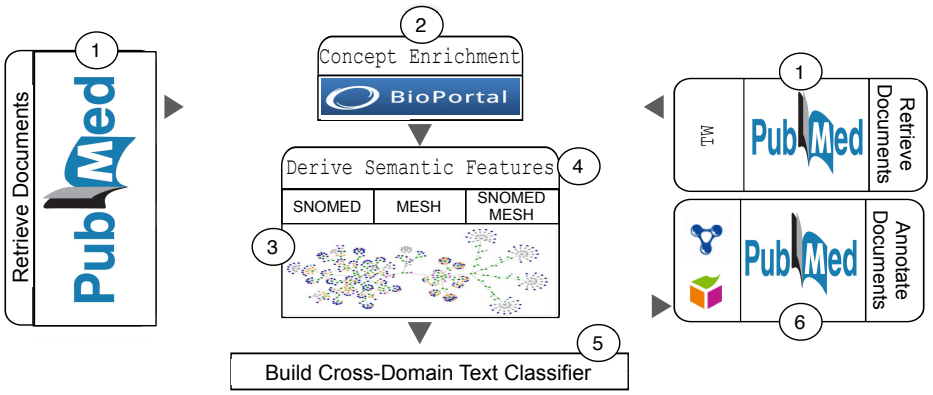

Figure 5.1: Architecture of the adaptive document zone classification framework using semantic features.

An overview of the individual components of the framework using biomedical domain ontologies (SNOMED-CT and MeSH) and datasets (from the Pubmed repository[1]) is presented in Figure 5.1, which can be summarised as follows: 1) *provision of annotated examples and content modelling*; 2) *concept enrichment using domain ontologies*; 3) *semantic concept graph generation*; 4) *pivot feature creation*; 5) *building adaptive document zone classifiers by employing different transfer learning techniques* 6) *evaluation of the adaptive document zone classifiers on held out documents.*

In the next subsection (Subsection 5.3.1), the motivation behind the biomedical domain ontologies employed is explained, together with an overview of their main characteristics. This is followed by a detailed description of each component of the framework.

5.3.1 Motivation

UMLS [Bodenreider, 2004] is the largest knowledge source for the biomedical domain, developed and maintained by the US government since 1986. This KS has been manually built by domain experts, covering a large amount of biomedical terms, which are organised in a metathesaurus, a repository of biomedical terminology and relationships. The main goal of this metathesaurus is to group names for the same concepts together from different biomedical knowledge sources. In order to achieve this, each concept is assigned a unique identifier (Concept Unique Identifier (CUI)), one or more concept names, and pointers to the other knowledge source vocabularies.

Over the years UMLS has been continually updated with new concepts and, thus, its size has significantly increased. According to Woods et al. [2006], the number of new concepts added to UMLS exceeded 300,000 between 1998 and 2002, and it further increased with 100,000 new concepts added between 2002 and 2003. The number of concepts to date are over two million[2].

[1] www.ncbi.nlm.nih.gov/pubmed

[2] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

The most widely used UMLS taxonomies in NLP are SNOMED-CT and MeSH. The main statistics about these KSs are summarised in Table 5.1.

SNOMED-CT [IHTSDO, 2010][3] serves as one of the largest parts of UMLS. It covers a wide range of clinical terms, from clinical findings, to symptoms, to diagnoses, to procedures, to body structures, to organisms. SNOMED-CT organises its concepts into its own ontology (*sct*), having each concept associated with several "descriptors" - names used to refer to the concept - such as a unique "Fully Specified Name (FSN)", a "preferred term", and one or multiple synonyms. The concepts are further connected to one another through the hyponym relation and other domain-specific relationships such as ("due to", "causative agent"). The number of unique classes in *sct* is 401,200, the maximum depth of the ontology is 28 and the maximum number of children is 2,712.

The MeSH taxonomy [Rogers, 1963][4] has been used to index biomedical journal articles. This indexing task is performed manually by a small group of highly qualified experts at the U.S. National Library of Medicine (NLM), who first read the full text of the journal article, and then assign the main concepts discussed within the articles as indexing terms to the articles. These main concepts are often refereed to as MeSH Headings or "descriptors", being associated with a definition and a list of synonyms for these descriptors. The MeSH controlled vocabulary (*msh*)[5] contains 245,885 classes which are organised into hierarchies representing subtopics (sub-hierarchies). Among these, 26,581 are main headings, which are grouped into 83 topical Subheadings (SHs). In addition, there are 203,658 Supplementary Concepts (formerly Supplementary Chemicals), which are structured in an almost flat structure, where most of the classes do not have any children assigned.

Semantic Features	SNOMED-CT (*sct*)	MeSH (*msh*)
Resources	0	0
Class *(Cls)*	401,200	245,885
Property *(P)*	0	0

Table 5.1: Statistics about *sct* and *msh* KS ontologies.

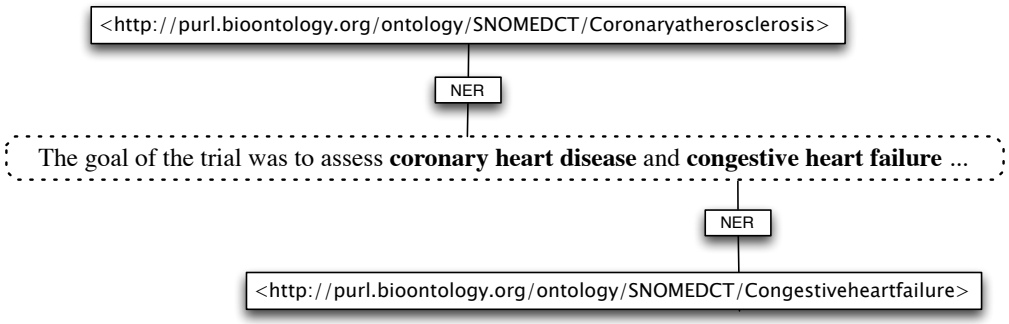

Figure 5.2: Example sentence mentioning different entities.

Overall, the main benefits of exploiting these domain ontologies (*sct*, *msh*) are that

[3] http://bioportal.bioontology.org/ontologies/SNOMEDCT
[4] http://www.nlm.nih.gov/mesh/
[5] http://bioportal.bioontology.org/ontologies/MESH

they provide a broad coverage of the biomedical domain and also allow the exploitation of semantic information about their concepts. Considering the example sentence depicted in Figure 5.2, one could exploit the semantic information about the entity resource <http://purl.bioontology.org/ontology/SNOMEDCT/Coronaryatherosclerosis> to provide additional information about the entity, e.g. being a type of *Heart disease* or *Disorder of coronary artery*. Given that each entity resource is associated to several ontological classes or concepts, thus additional contextual information can be provided for these resources, enabling the exploitation of various semantic structures related to these resources. The use of semantic structures within these domain ontologies could therefore help to provide a conceptual representation of the domain documents by incorporating additional contextual information about the concepts identified in them.

5.3.2 Provision of Annotated Data and Content Modelling

The initial step of the framework consists of the *provision of annotated data* for both source and target domains, and the *creation of an initial feature space* for modelling the content of these domain documents.

In order to compile a corpus of annotated documents for both source and target domains, the PubMed Central[6] repository was used, which provides a vast number of journal articles belonging to a range of different biomedical sub-domains. Following the process described in Mihăilă et al. [2012], for a particular biomedical sub-domain (e.g. health), the PubMed API[7] was employed for retrieving journal articles written in English, whose Broad Subject Term contains only the name of the sub-domain. The compiled biomedical corpora were further pre-processed and split into zones using the methodology described in Subsection 4.4.1.

Considering the *supervised transfer learning* scenario employed by this framework, thus the source domain consists of an abundant amount of labelled data from one biomedical sub-domain (e.g. health), while the target domain contains a small number of annotated documents and a large amount of unlabelled documents from another sub-domain (e.g. biology).

To construct an initial feature set for the domains, the simple BoW representation is used, where each word is weighted by TF-IDF (term frequency-inverse document frequency)[8]. This representation allows the zones to be represented based on what is discussed in them.

5.3.3 Concept Enrichment using Domain Ontologies

The next step of the framework relies on enhancing the representation of the documents in the source and target domains. In order to achieve this, the *entities and concepts are extracted* from the documents using the BioPortal REST API[9] according to different UMLS KSs. BioPortal [Salvadores et al., 2013] is a community-based ontology repository which provides access to all the UMLS KSs[10] via its public SPARQL endpoint[11]. The data in

[6] http://www.ncbi.nlm.nih.gov/pmc/

[7] http://www.ncbi.nlm.nih.gov/pmc/tools/oai/

[8] It also worth noting that the purpose of this thesis is to evaluate the added value of incorporating semantic features into cross-domain document zone classifiers. The incorporation of other state-of-the-art features such as part-of-speech into this framework thus serves the goal of future work.

[9] http://purl.bioontology.org/mapping/rest

[10] The list of all ontologies within BioPortal is enlisted at https://bioportal.bioontology.org/ontologies.

[11] http://sparql.bioontology.org/

BioPortal has been populated by researchers and practitioners, and stored in RDF format, consisting of ontologies and vocabularies, as well as data instances.

This framework exploits the SNOMED-CT and MeSH UMLS KSs for enrichment. These KSs are employed alone as well as in combinations, resulting in three different *concept enrichment scenarios*: i) *sct* - linking concepts to the SNOMED-CT ontology; ii) *msh* - linking concepts to the MeSH ontology; and iii) *sct+msh* - linking concepts to both SNOMED-CT and MeSH KSs.

5.3.4 Semantic Meta-graph Generation

In this step, the concepts within the documents are mapped into SNOMED-CT and MeSH URIs, allowing the incorporation of rich semantic information about concepts into document zone classifiers.

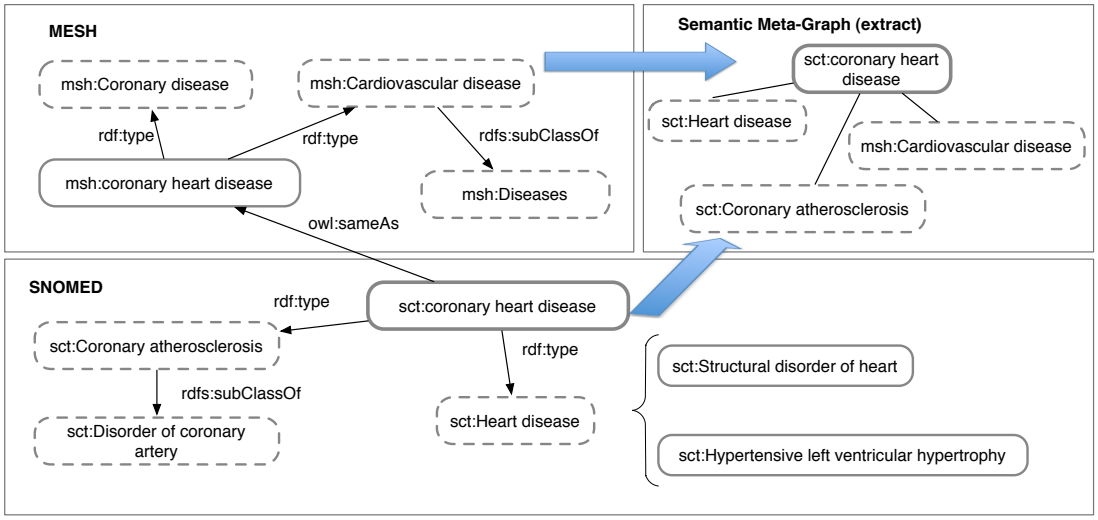

Figure 5.3: Deriving a semantic meta-graph from multiple biomedical KSs.

Figure 5.3 presents an extract of the semantic classes for the entity "coronary heart disease". For this entity resource, this framework retrieves all the <*subject, predicate, object*> instance triples associated with it, and exploits the semantic structure created from these triples at a meta-level using a semantic meta-graph, *Resource meta-graph*, constructed according to Definition 4. To remind the reader, let $\mathtt{G} := (R, P, C, Y)$ denote the *resource meta-graph* employed in this framework.

This *resource meta-graph* contains the complete information present in KS ontologies about a given resource. Given the entity e, the $G(e)$ represents a set of tuples $G(e) = (R, P, C, Y')$, consisting of the aggregation of all resources, properties and classes associated with this entity.

Taking into account the two different UMLS KSs exploited in this framework, three different *Resource meta-graph*s are considered: i) one from SNOMED-CT using *sct* ontology; ii) one from MeSH using *msh* ontology; and iii) another one from the *sct+msh* combined ontologies.

augmentation strategy		feature name	feature value
F	Baseline	BoW	coronary, heart, disease
F'_{A1}	Cls	Cls_1	$f_{TFIDF}(sct$:HeartDisease)
	Cls	Cls_2	$f_{TFIDF}(msh$:CardiovascularDiseases)
	Cls($sct + msh$)	$Cls_1 + Cls_2$	$f_{TFIDF}(sct$:HeartDisease), $f_{TFIDF}(msh$:CardiovascularDiseases)
	Cls($sct + msh_{CCA}$)	$Cls_1 + Cls_{2CCA}$	$f_{CCA}(sct$:HeartDisease, msh:CardiovascularDiseases)
F'_{A2}	parent(Cls)	parent(Cls_1)	$f_{TFIDF}(sct$:DisorderOfCardiovascularSystem)
	parent(Cls)	parent(Cls_2)	$f_{TFIDF}(msh$:Diseases)

Table 5.2: Example semantic augmentation strategies for the entity *coronary heart disease* using semantic features derived from the *resource meta-graph*. The first column stands for the augmentation strategies used to incorporate semantic features into a document zone classifier, the second column provides example features to which the augmentation strategies are applied, while the third column gives examples of possible values for each such feature. The example semantic features considered here are Cls_1, Cls_2, referring to semantic class features derived for *coronary heart disease* from the *sct* and *msh* ontologies respectively. For the sake of completeness, in the first row, the original feature space denoted by F, consisting of BoW features, is also presented. For this feature representation no augmentation strategy is applied.

For the semantic features further two different augmentation strategies are presented: F'_{A1} extending the F features with semantic features, and F'_{A2} augmenting the F features with semantic features derived from the class hierarchies of KSs (e.g. considering the parent classes of a class (parent(Cls))). In addition, for the F'_{A1} augmentation strategy two different ontology combination strategies are also presented: $Cls_1 + Cls_2$ consisting of a naive combination of the two semantic class features weighted with TF-IDF, and $Cls_1 + Cls_{2CCA}$ consisting of semantic features obtained after projecting these features into a latent space using CCA dimensionality reduction technique.

5.3.5 Pivot Feature Derivation and Combination

Having the semantic concepts graphs extracted, the following step consists of deriving *semantic pivot feature* from them, as follows:

Cls: Semantic class features: This feature set consists of a set of all the classes associated with an entity from a *Resource meta-graph*. For instance, for the entity "coronary heart disease", these features would be *sct*:*HeartDisease*, *msh*:*Cardiovascular Diseases*, and *msh*:*Coronary Disease*. The main intuition here is that entity classes can be characteristics of a zone type, serving as trigger words for that zone segment. For instance, the entity class "Symptom Finding" is more likely to be used in the "Results" zone, than in the "Introduction" zone.

Given that this framework exploits semantic concepts graphs from multiple ontologies (*sct*+*msh*), the combination of this semantic information is achieved by applying different *ontology combination strategies*:

sct+*msh* - *Naive Combination*: This strategy provides a simple approach for the combination of the semantic features of two ontologies by creating a joint feature set of the features obtained from the individual ontologies. The main idea behind this approach is to represent the content of the documents using the complete semantic information obtained from the two ontologies. Considering that the two ontologies only contain a partial overlap of the concepts discussed in them, this strategy provides complementary information and knowledge about concepts, resulting in a more complete and comprehensive view of the content/subject of the documents.

sct+msh_{CCA} - *Dimensionality Reduction*: While the previous strategy utilises all the available information from the employed ontologies, this information may often contain partial inconsistencies or noise (such as duplicate entities or concepts), due to the open nature of these resources. In order to avoid this, a more principled way for the combination of these ontologies is to apply canonical correlation analysis (CCA), which aims to remove the noise from the data by ignoring the irrelevant dimensions. This technique has been successfully applied for a variety of different problems dealing with multiple modal [Trivedi et al., 2011] or multi-domain data [Blitzer et al., 2007b].

CCA computes a low-dimensional shared embedding of both sets of features such that the correlation among the features of the two ontologies is maximised in the embedded space. Let $D_{sct} \in \mathbb{R}^{D_1 \times N}$ denote a dataset consisting of a set of pivot features from the *sct* ontology, and $D_{msh}^{D_2 \times N}$ denote a dataset comprising a set of pivot features from the *msh* ontology for the joint source and target domain dataset D. CCA seeks to find a linear projection $w_{f_{sct}} \in \mathbb{R}^{D_1}$ and $w_{f_{msh}} \in \mathbb{R}^{D_2}$, such that after projecting, the corresponding instances of the domains are maximally correlated (similar) in the embedded space. The correlation coefficient between these two datasets can be computed as follows:

$$\rho = \max_{w_{f_{sct}}, w_{f_{msh}}} \frac{w_{f_{sct}}^T \cdot D_{sct} \cdot D_{msh}^T \cdot w_{f_{msh}}}{\sqrt{(w_{f_{sct}}^T \cdot D_{sct} \cdot D_{sct}^T \cdot w_{f_{sct}}) \cdot (w_{f_{msh}}^T \cdot D_{msh} \cdot D_{msh}^T \cdot w_{f_{msh}})}}.$$

Considering that the correlation is not affected by the rescaling parameters $w_{f_{sct}}$ and $w_{f_{msh}}$, CCA is posed as an optimisation problem:

$$\max_{w_{f_{sct}}, w_{f_{msh}}} w_{f_{sct}}^T \cdot D_{sct} \cdot D_{msh}^T \cdot w_{f_{msh}}, \text{ subject to}$$

$$w_{f_{sct}}^T \cdot D_{sct} \cdot D_{sct}^T \cdot w_{f_{sct}} = 1, w_{f_{msh}}^T \cdot D_{msh} \cdot D_{msh}^T \cdot w_{f_{msh}} = 1.$$

Hardoon et al. [2003] showed that the above formulation is equivalent to solving following generalised eigen-value problem:

$$\begin{pmatrix} 0 & \Sigma_{D_{sct},D_{msh}} \\ \Sigma_{D_{msh},D_{sct}} & 0 \end{pmatrix} \times \begin{pmatrix} w_{f_{sct}} \\ w_{f_{msh}} \end{pmatrix} = \lambda \begin{pmatrix} \Sigma_{D_{sct},D_{sct}} & 0 \\ 0 & \Sigma_{D_{msh},D_{msh}} \end{pmatrix} \times \begin{pmatrix} w_{f_{sct}} \\ w_{f_{msh}} \end{pmatrix},$$

where $\Sigma_{D_{sct},D_{msh}}$ stands for the cross-covariance between D_{sct} and D_{msh}, while $\Sigma_{D_{sct},D_{sct}}$ and $\Sigma_{D_{msh},D_{msh}}$ stands for the covariances of the D_{sct} and D_{msh} datasets respectively.

5.3.6 Building Adaptive Document Zone Classifier

After the extraction of the pivot features, the next step in this framework consists of the incorporation of these features into the TC framework. For this purpose, the framework employs different *semantic augmentation strategies*, which provide alternative ways for the combination of the original lexical feature space and newly inferred semantic feature space.

5.3.6.1 Semantic Augmentation and Feature Duplication

The key idea behind the semantic augmentation strategies employed is to allow the explicit modelling of the *general (domain-independent)* and *domain-specific characteristics* of the source and target domains, which can both contribute to the performance of a document zone classifier. In order to achieve this, a modified version of the widely used Easy Adapt (EA) [Daumé, 2007] transfer learning approach is presented. EA has proven success on a range of problems, including sentiment classification, named entity recognition and part-of-speech tagging.

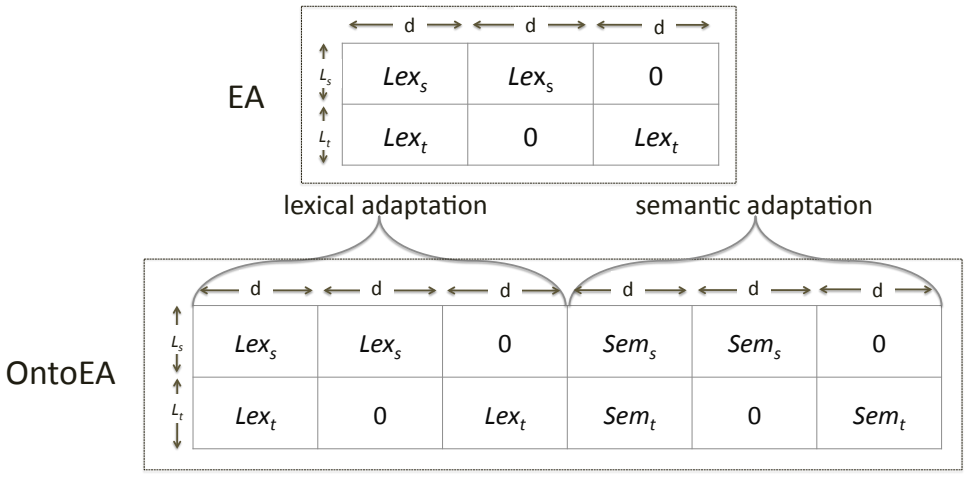

Figure 5.4: A diagram representation of the original EA and modified OntoEA transfer learning strategies.

An overview of the original feature duplication strategy underlying the EA approach, as well as its extension, named OntoEA, is presented in Figure 5.4. Algorithm 3 and Algorithm

Algorithm 3 Adaptive *within-document* TC using the original EasyAdapt (EA) approach.

1: ***Input***: L_S labelled source domain documents (e.g. articles related to health), L_T labelled target domain documents (e.g. articles related to biology), U_T unlabelled target domain documents, F_S feature set of the source domain documents, F_T feature set of the target domain documents.
2: Merge the lexical feature set of the first domain (F_S) with the lexical feature set of the second domain (F_T) into a common lexical feature set ($F_{Lex} = F_S \cup F_T$)
3: Augment the examples $x =< x_{lex} >$ from L_S with a source-specific copy of the original lexical features $< x_{lex}, x_{lex}, 0 >$ according to F
4: Augment the examples $x =< x_{lex} >$ from L_T, U_T with a target-specific copy of the original lexical features $< x_{lex}, 0, x_{lex} >$ according to F
5: Train a supervised classifier (e.g. SVM) on the annotated examples from both domains ($L_S \cup L_T$)
6: ***Output***: Annotated U_T target domain documents.

4 further summarise the main steps of EA and OntoEA, respectively. As can be observed, the extension in OntoEA is to perform feature duplication on both lexical (Lex) and semantic (Sem) features by applying the same EA feature augmentation strategy twice. The intuition here is that performing feature repetition on the semantic feature space could allow further reducing of the gap between domains, as domains which are distant in the lexical feature space, may become closer in the newly created semantic feature space.

Before the execution of feature duplication, first the lexical (F_{LEX}) and semantic (F_{SEM}) features are augmented together into a merged feature set $F = F_{LEX} \cup F_{SEM}$. For this purpose, two different semantic augmentation strategies are investigated. Examples for the various semantic features and semantic augmentation strategies employed for the entity *coronary heart disease* are provided in Table 5.2.

Semantic augmentation: This strategy augments the original lexical features (e.g BoW) with additional semantic information extracted for the entities appearing in a document. Given the **Cls** (i.e. semantic class features) feature set introduced in Subsection 5.3.4, the feature set F is extended into F'_{A1-Cls} by adding the class features extracted from the aggregation of the semantic meta-graphs of those entities appearing in the document x. Therefore, the expanded feature vocabulary size is $|F'_{A1-Cls}| = |F| + |F_{cls}|$ where $|F_{cls}|$ denotes the total number of these class features.

Semantic augmentation with concept generalisation: This augmentation strategy aims to further improve the generalization of a document zone classifier by exploiting the subsumption relation among classes within the SNOMED-CT or MeSH ontologies. Therefore in this strategy, instead of using the *typeOf* class cls of an entity, a more generic class of cls is considered, namely the set of parent classes of $cls(parent(cls))$. In this case the feature set F is enhanced with the set of parent classes of cls where $cls \in Cls$. Therefore the size of the augmented feature set F'_{A2-Cls} is computed as $|F'_{A2-Cls}| = |F| + |F_{parent(cls)}|$, where $|F_{parent(cls)}|$ denotes the total number of unique parent classes of cls.

Having the final augmented feature set F' created (step 7 in Algorithm 4), the main steps of OntoEA can be summarised as follows. For each instance $x =< x_{lex}, x_{sem} > \in F'$ an augmented feature vector is created $< x_{lex_{general}}, x_{lex_{source}}, x_{lex_{target}}, x_{sem_{general}}, x_{sem_{source}}, x_{sem_{target}} >$ consisting of a general, source-specific and target-specific version of both lexical

Algorithm 4 Adaptive *within-document* TC using the proposed OntoEasyAdapt(OntoEA) approach.

1: ***Input***: L_S labelled source domain documents (e.g. articles related to health), L_T labelled target domain documents (e.g. articles related to biology), U_T unlabelled target domain documents, F_S feature set of the source domain documents, F_T feature set of the target domain documents.
2: Merge the lexical feature set of the first domain (F_S) with the lexical feature set of the second domain (F_T) into a common lexical feature set ($F_{Lex} = F_S \cup F_T$)
3: Extract entities and concepts from both source and target domains
4: Exploit different semantic meta-graphs for the extracted concepts in both domains
5: Create semantic features from the semantic meta-graphs for both source F_{S_C} and target F_{T_C} domains
6: Merge the semantic feature set of the source domain (F_{S_C}) and target domain (F_{T_C}) into a common semantic feature set ($F_{SEM} = F_{S_C} \cup F_{T_C}$)
7: Merge the original common lexical feature space (F_{Lex}) with the semantic feature space F_{SEM}, ($F = F_{LEX} \cup F_{SEM}$)
8: Augment the examples $x =< x_{lex}, x_{sem} >$ from L_S with a source-specific copy of the original lexical and semantic features $< x_{lex}, x_{lex}, 0, x_{sem}, x_{sem}, 0 >$ according to F
9: Augment the examples $x =< x_{lex}, x_{sem} >$ from L_T, U_T with a target-specific copy of the original lexical and semantic features $< x_{lex}, 0, x_{lex}, x_{sem}, 0, x_{sem} >$ according to F
10: Train a supervised classifier (e.g. SVM) on the annotated examples from both domains ($L_S \cup L_T$)
11: ***Output***: Annotated U_T target domain documents.

and semantic features. Step 8 creates the augmented feature vectors for the source domain documents. Then, step 9 creates the augmented feature vectors for the target domain documents. In step 10 a classifier is trained on the augmented labelled instances, and tested on the remaining unlabelled target data in step 11.

5.4 Measuring the Similarity between Domains for Cross-domain Document Zoning

This section proposes various measures for quantifying the distributional differences between domains. Designing such a domain similarity measure can be extremely important, as it can serve as a proxy for the performance of a cross-domain document zone classifier. It is expected that the closer the two domains, the better the performance of a cross-domain document zone classifier [Pan and Yang, 2010]. In order to quantify this similarity, the documents are decomposed into *bag-of-zones*, and further into *bag-of-words* and *bag-of-entities*, as follows:

Definition 5 *Let* $\overrightarrow{d_s} = (z_{s_1}, z_{s_2}, \ldots, z_{s_{|Z|}})$ *define the* ***bag-of-zones*** *representation of the source domain* D_S*, where* z_{s_i} *contains all the paragraphs which were assigned to cluster* s_i*; and* $\overrightarrow{d_t} = (z_{t_1}, z_{t_2}, \ldots, z_{t_{|Z|}})$ *the* ***bag-of-zones*** *representation of the target domain* D_T*, where* z_{t_j} *contains all the paragraphs which were assigned to cluster* t_j*. Further, each zone is split into* ***bag-of-words*** *and* ***bag-of-entities****, where the entities were obtained following the entity enrichment process:* $\overrightarrow{z}_{s_{j\,BOW}} = (w_{s_1}, \ldots, w_{s_m})$, $\overrightarrow{z}_{s_{j\,BOE}} = (e_{s_1}, \ldots, e_{s_n})$ *corresponding to the representation for the source domain, and* $\overrightarrow{z}_{t_{j\,BOW}} = (w_{s_1}, \ldots, w_{s_m})$, $\overrightarrow{z}_{t_{j\,BOE}} = (e_{s_1}, \ldots, e_{s_n})$ *corresponding to the representation for the target domain. The weight for each features in* $\overrightarrow{z}_{BOW}$*, and* $\overrightarrow{z}_{BOE}$ *is TF-IDF.*

Given that gathering annotation for every domain is difficult, this section proposes *unsupervised domain similarity measures*, which rely only on the lexical and conceptual information present in the domains. That is, the proposed measures disregard any of the label information about the documents. In order to create the *bag-of-zones* representation of the domain documents, the unsupervised probabilistic clustering approach zoneLDA (see Subsection 4.3.2) is employed[12]. The main idea behind zoneLDA is to model the generative story in which the documents were created and assign each paragraph to one of the clusters.

Once the paragraphs are clustered, the similarity between two domains is measured by computing the lexical and ontological closeness of the paragraphs belonging to the individual zone types. The novel document similarity derived for cross-domain document zone classification can be computed as follows:

$$sim_{\text{intradoc}}(\overrightarrow{d_s}, \overrightarrow{d_t}) = sim((z_{s_1}, z_{s_2}, \ldots, z_{s_{|Z|}}), (z_{t_1}, z_{t_2}, \ldots, z_{t_{|Z|}}))$$

$$= \sum_{z_s, z_t \in z_1, \ldots, z_{|Z|}} \text{lexica\_sim}(z_s, z_t) + \text{onto\_sim}(z_s, z_t).$$

To compute the *lexical similarity* ($lexica\_sim(z_s, z_t)$) between the source and target zone pairs, classical *corpus similarity measures* are employed [Kilgarriff, 2001], including χ^2 statistics, a symmetric KL divergence metric, and cosine similarity:

- *Chi-squared (χ^2) test*: The χ^2 test measures the independence between the lexical feature sets ($F_{S_{LEX}}$ and $F_{T_{LEX}}$) of the zones in the training and test datasets. Given the $\overrightarrow{z}_{s_{BOW}}$ and $\overrightarrow{z}_{t_{BOW}}$ vectors, the χ^2 test can be computed as

$$\chi^2 = \sum \frac{(O - E)^2}{E},$$

  where O is the observed value for a feature, while E is the expected value calculated on the basis of the joint corpus.

- *Kullback-Leibler symmetric distance* (KL): Originally introduced in Bigi [2003], the symmetric KL divergence metric measures how different the $\overrightarrow{z}_{s_{BOW}}$ and $\overrightarrow{z}_{t_{BOW}}$ vectors are on the joint set of features $F_{S_{LEX}} \cup F_{T_{LEX}}$:

$$KL(\overrightarrow{z}_{s_{BOW}} || \overrightarrow{z}_{t_{BOW}}) = \sum_{f \in F_{S_{LEX}} \cup F_{T_{LEX}}} (\overrightarrow{z}_{s_{BOW}}(f) - \overrightarrow{z}_{t_{BOW}}(f)) \log \frac{\overrightarrow{z}_{s_{BOW}}(f)}{\overrightarrow{z}_{t_{BOW}}(f)}$$

- *Cosine similarity* measure: The cosine similarity represents the angle that separates the train and test vectors $\overrightarrow{z}_{s_{BOW}}$ and $\overrightarrow{z}_{t_{BOW}}$:

$$cosine(\overrightarrow{z}_{s_{BOW}}, \overrightarrow{z}_{t_{BOW}}) = \frac{\sum_{k=1}^{F_{S_{LEX}} \cup F_{T_{LEX}}} (\overrightarrow{z}_{s_{BOW}}(f_{S_k}) \times \overrightarrow{z}_{t_{BOW}}(f_{T_k}))}{\sum_{k=1}^{F_{S_{LEX}} \cup F_{S_{LEX}}} (\overrightarrow{z}_{s_{BOW}}(f_{S_k}))^2 \times \sum_{k=1}^{F_{S_{LEX}} \cup F_{T_{LEX}}} (\overrightarrow{z}_{t_{BOW}}(f_{T_k}))^2}$$

  It can also be noted that while the *cosine* measure captures the actual similarity between domains, the other two measures (KL, χ^2) measure the distance (inverse

[12]The motivation behind the selection of zoneLDA is that in the previous experiments it was shown to perform better than other probabilistic models for document zoning. The proposed similarity measures are, however, general and allow the incorporation of any other probabilistic graphical model suitable for this task.

similarity) between domains.

To capture the *ontological similarity* ($onto\_sim(z_s, z_t)$) between the domains, the *bag-of-entities* representation of the zones is used. Given two zones z_{s_j} and z_{t_j}, the ontological similarity is defined as follows [Mihalcea et al., 2006]:

$$sim(z_{s_j}, z_{t_j}) = \frac{1}{2}(\frac{\sum_{e \in z_s}(max(e, z_t) * idf(e))}{\sum_{e \in z_s}} + \frac{\sum_{e \in z_t}(max(e, z_s) * idf(e))}{\sum_{e \in z_t}}).$$

That is, for each entity e in the source zone ($e \in z_{s_j}$), an entity from target zone ($e \in z_{t_j}$) is found, such that the two entities have the highest similarity ($max(e, z_t)$) according to one *knowledge-based* similarity measure[13]. For this purpose, two different *knowledge-based* similarity measures are investigated: *path-based* similarity measures (sim_{lch}, sim_{wup}), and *information content-based* similarity measures (sim_{lin}, sim_{jnc}), both using *sct* and *msh* as reference ontologies:

- **Leacock & Chodorow**'s *path-based* similarity measure [Leacock and Chodorow, 1998]: This measure uses the normalised path between two concepts, which is computed as :

$$sim_{lch} = -log\frac{length(c1, c2)}{2 * MAX},$$

  where *length* is the number of edges on the shortest path in the ontology between two concepts and MAX is the depth of the taxonomy.

- **Wu & Palmer**'s *path-based* similarity measure [Wu and Palmer, 1994]: This measure takes into account the depth of the individual concepts c_1 and c_2 in the ontology, as well as depth of the least common subsumer (LCS):

$$sim_{wup} = \frac{2 * depth(LCS)}{depth(c_1) + depth(c_2)}$$

- **Jiang & Conrath**'s *information content-based* similarity measure [Jiang and Conrath, 1997]: This measure compares the sum of the information content of the individual concepts with that of their lowest common subsumer

$$sim_{jnc} = \frac{1}{IC(c_1) + IC(c_2) - 2 * IC(LCS)},$$

  where $IC(c)$ is the information content [Patwardhan et al., 2003] of a concept, and LCS denotes the lowest common subsumer, which represents the most specific concept that the two concepts have in common.

- **Lin**'s *information content-based* similarity measure [Lin, 1998]: This measure scales the information content of lowest common subsumer with the sum of information content of two concepts:

$$sim_{lin} = \frac{2 * IC(LCS)}{IC(c_1) + IC(c_2)}$$

[13] For the computation of various similarity measures, the UMLS-Similarity package was employed, available at http://search.cpan.org/dist/UMLS-Similarity/utils/query-umls-similarity-webinterface.pl

5.5 Compiling a Gold Standard Dataset for Cross-domain Document Zoning

In order to evaluate the effectiveness of the proposed transfer learning framework for document zoning, a list of biomedical sub-domain corpora has been compiled. Considering that to date there is no publicly available multi-domain dataset for document zoning, the dataset compiled by Mihăilă et al. [2012] is employed. This dataset was initially used for studying biomedical sub-domain variation on the full text of journal articles.

The original multi-domain dataset comprises of articles belonging to 20 biomedical sub-domains, which were selected from Pubmed by taking into account the subject heading associated with the documents, corresponding to the sub-domain of the documents[14]. For each such sub-domain a total of 400 documents were considered. These were further split into the IMRAD zones ($Z = \{Abstract, Introduction, Methods, Results, Discussion\}$) using the same methodology described in the previous chapter (Section 4.4). That is, the documents were first pre-processed by removing all the tables, figures, mathematical formulas, references, and metadata information (authors, affiliations, publication history), thereby disregarding additional clues which could help in the recognition of zones. Following this, additional pre-processing steps were performed, including the removal of stopwords and stemming of words (using Porter stemmer). When creating the lexical features for the classifiers, the feature spaces were further reduced to the top-1000 words weighted by TF-IDF for each domain.

Zone Name	Equivalent section labels
Introduction	Introduction, Background, Objectives, Review, Aims, Context
Method	Design and Method, Experimental Design, Materials and Methods, Experimental Design, Algorithm
Results	Result, Results, Main Results
Discussion	Discussion, Conclusion, Conclusions, Implications

Table 5.3: Zone name variations across multiple sub-domains.

The most challenging step in the creation of the final gold standard was the normalisation of the zone names for the documents. As illustrated in Table 5.3, there exists a large variation in the expressions used to refer to a particular zone type. In order to address these challenges, a list of manually tailored rules were fired. These rules were also manually validated to ensure the correctness of the resulting zone annotations.

For the purpose of the experiments conducted in this chapter, the number of sub-domain corpora was narrowed down to seven sub-domains: Biology (*Biol*), Cell Biology (*CellBiol*), Communicable Diseases (*Communi*), Health Services Research (*HealthS*), Medicine (*Medicin*), Public Health (*PublicH*), and Tropical Medicine (*Tropica*). *CellBiol* is a sub-area of biology, which studies cells (their structure and properties) and their interaction with the environment. *Communi* focuses on research findings related to clinically evident illness resulting from an infection, including, for instance, HIV disease, Diarrheal diseases, and Malaria. *HealthS* studies methods and concepts related to the financing, organisation,

[14]The complete list of biomedical sub-domains covered in the original dataset are: Allergy and Immunology, Biology, Cell Biology, Communicable Diseases, Critical Care, Environmental Health, Genetics, Health Services Research, Medical Informatics, Medicine, Microbiology, Neoplasms, Neurology, Pharmacology, Physiology, Public Health, Pulmonary Medicine, Rheumatology, Tropical Medicine, Virology.

Zone name	*Biol*	*CellBiol*	*Communi*	*PublicH*	*HealthS*	*Tropica*	*Medicin*
Abstract	1,043	461	1,418	1,613	1,639	1,478	1,754
Introduction	2,146	1,159	1,596	2,212	2,285	1,880	2,581
Method	4,287	4,881	3,210	4,455	5,400	3,819	3,971
Results	8,333	7,333	2,827	4,240	5,008	3,507	3,875
Discussion	3,349	3,777	3,220	4,060	3,681	3,041	3,302

Table 5.4: The total number of paragraphs for each IMRAD zone in the seven biomedical sub-domain corpora analysed.

delivery, evaluation, and outcomes of health services. The articles belonging to the *Medicin* sub-domain, cover all aspects of the medicine, including diagnosis, treatment, and prevention of illness. *PublicH* is concerned with research findings related to public health (including biostatistics and epidemiology) and health care (focusing on nursing and medicine). Finally, *Tropica* is a branch of medicine, which deals with health problems (e.g., tropical disease such as malaria) which occur in tropical and subtropical regions.

General statistics for the analysed biomedical sub-domains are provided in Table 5.4 and Table 5.5. As expected, the shortest zone type is the Abstract, followed by the Introduction. Concerning the other three zone types, it can be seen that in the majority of the cases the Discussion is shorter than the Results and Methods, while the length of the Results and Methods largely depends on the corpora. In the *Biol* and *CellBiol* domains the paragraphs within the Results zone are longer than in the Methods zone, while in the rest of the domains the opposite trend can be observed.

	Statistics	*Biol*	*CellBiol*	*Communi*	*PublicH*	*HealthS*	*Tropica*	*Medicin*
Lexical	AvgW/Para	131.31	130.47	102.00	98.11	89.77	111.95	110.78
	BoW	152,269	121,924	80,297	85,996	77,666	90,673	107,905
	BoE	2,155	1,644	2,058	2,084	1,629	1,755	2,423
Semantic	SMDcls	1,691	1,207	1,591	1,543	1,221	1,357	1,804
	MSHcls	2,465	1,747	2,153	2,105	1,703	1,796	2,517
	SMDcls/ent	2.40	2.33	1.91	1.79	1.80	1.83	2.01
	MSHcls/ent	2.15	2.11	1.55	1.35	1.80	1.52	1.64

Table 5.5: General statistics about the analysed biomedical sub-domain corpora, consisting of 400 documents/sub-domain. With regard to the lexical features: AvgW/Para stands for the average number of words per paragraph (zone), BoW denotes the unique number of words in the corpus, BoE refers to the unique number of entities in the corpus. Concerning the semantic features: SMDcls is the average of number of unique class features derived from *sct*, and MSHcls is the average number of unique class features created from *msh*.

Regarding the semantic features, the number of unique *msh* classes is higher than the number of *sct* classes in each sub-domain corpora. This indicates that a larger number of *msh* classes are used to enrich the domain documents. Further, concerning the average number of classes per entities, the values are very similar for the two KSs ontologies. On average, most entities have two *sct* and *msh* classes associated with them.

5.6 Evaluation

This section presents a series of experiments to evaluate the proposed *adaptive document zone classification framework* and *unsupervised domain similarity measures* using the different semantic pivot features and augmentation strategies described in Section 5.3.

Before discussing these experiments in detail, the baseline methods used in the experiments are presented in Subsubsection 5.6.1.1, and the evaluation measures are introduced in Subsection 5.6.2. Following this, the experimental setup is described in Subsection 5.6.3, and a discussion of the results is provided in Subsection 5.6.4.

5.6.1 Baseline Methods

5.6.1.1 Baseline Methods for Adaptive Document Zone Classification

The proposed adaptive text classification framework has been evaluated using different semantic features and augmentation strategies against several baseline models corresponding to state-of-the-art approaches for document zoning. These baseline models consist of the following features:

Bag-Of-Unigrams (BoW) Features: The uni-gram features captures the natural intuition to utilise what it is known about a particular zone segment, so that the features which are most indicative of a zone segment can be detected and the appropriate label(s) assigned. The BoW features consist of a collection of words weighted by TF-IDF (term frequency-inverse document frequency), capturing the relative importance of a word in a document with respect to its use on the whole corpus.

Bag-Of-Entities (BoE) Features: This feature set extends the lexical BoW features with entities extracted using available entity annotation services, e.g. BioPortal API, weighted by TF-IDF. In this case the value of the BoE features thus captures the occurrence of the entity and concept pairs $f_{BoE}('coronaryheartdisease' \wedge sct{:}\text{HeartDisease})$, where sct:HeartDisease corresponds to the most likely concept returned by BioPortal.

Considering the above baseline features, four strong baseline supervised machine learning models are employed:

- SRC_ONLY: a *source only* document zone classifier, in which an SVM model is built only on the source domain documents,
- TGT_ONLY: a *target only* document zone classifier, in which an SVM model is built only on the target domain documents,
- SRC_TGT: a *source-target* document zone classifier, in which an SVM model is built on both source and target domain documents,
- EA: the original *easy-adapt* document zone classifier, in which an SVM model is built on both source and target domain documents.

5.6.1.2 Baseline Domain Similarity Measures for Document Zoning

Given that to date there are no domain similarity measures proposed for document zoning, the baseline domain similarity measures used in the experiments are counterpart measures

of the proposed ones, which make use only of the lexical information in the domain documents[15]. That is, first the bag-of-words and bag-of-entities are merged into a singe vector $\overrightarrow{z}_{s_{BOW}} \cup \overrightarrow{z}_{s_{BOE}}$ for the source domain, and $\overrightarrow{z}_{t_{BOW}} \cup \overrightarrow{z}_{t_{BOE}}$ for the target domain. Next, the cosine, χ^2 and KL-divergence measures are computed over these vectors, as follows:

$$sim_{\text{baseline}}(\overrightarrow{d_s}, \overrightarrow{d_t}) = \text{cosine}|\text{KL}|\chi^2(\overrightarrow{z}_{s_{BOW}} \cup \overrightarrow{z}_{s_{BOE}}, \overrightarrow{z}_{t_{BOW}} \cup \overrightarrow{z}_{t_{BOE}})$$

5.6.2 Evaluation Measures

The measures used to evaluate the performance of the different document zone classifiers were the standard *precision*, *recall*, and *F1-measure*.

The *precision* (Prec_z) for a given zone z is computed as the ratio of the number of correctly annotated paragraphs to the total annotated:

$$\text{Prec}_z = \frac{|\text{correctly annotated paragraphs for zone z}|}{|\text{annotated paragraphs for zone z}|}$$

The *recall* (Rec_z) for a given zone z is the ratio of the number of correctly annotated paragraphs to the total number that should have been annotated:

$$\text{Rec}_z = \frac{|\text{correctly annotated paragraphs for zone z}|}{|\text{paragraphs which should have been annotated for zone z}|}$$

The evaluation is based on macro-averaged values in which the precision and recall values for the individual zone types are averaged. That is, the macro-average precision is computed as $\text{Prec} = (\sum_{z \in Z} Prec_z)/|Z|$, and the macro-average recall as $\text{Rec} = (\sum_{z \in Z} Rec_z)/|Z|$.

The macro-average *F1-measure* then provides a weighted combination of the two measures, defined as

$$\text{F1} = 2 \times \frac{\text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}$$

5.6.3 Experimental Set-up

Two different document zone scenarios were analysed and contrasted in the experiments: a single domain (or in-domain) case, in which case the baseline SVM document zone classifier trained on in-domain data only is employed (using the TGT_ONLY classifier), and a cross-domain scenario, in which case an SVM document zone classifier is trained on either out-of domain data alone (using the SRC_ONLY classifier) or combined with in-domain data (using the SRC_TGT, EA or OntoEA classifiers).

The experiments were performed using 80%-20% split for the target domain. For the TGT_ONLY document zone classifier 80% of the target domain documents is used. The same 80% is added to the 100% of the source domain documents for the cross-domain classifiers (SRC_TGT, EA and OntoEA). The SRC_ONLY classifier uses 100% of the

[15] In addition, it also worth noting that all the state-of-the-art domain similarity measures would require the domain documents to be annotated with zoning information, which is outside of the scope of the proposed measure. For this reason these measure are not directly comparable with the state-of-the-art domain similarity measures, e.g. A-distance [Rai et al., 2010].

source domain documents. Each of these document zone classifiers were evaluated on 20% of the target domain documents, using 5-fold cross-validation.

In light of the two document zone scenarios analysed, a series of experiments were conducted. The first set of experiments aims to assess the benefit of incorporating semantic pivot features into a document zone classifier. For this purpose the performance of the single-domain (SVM TGT_ONLY) and cross-domain (SVM SRC_TGT, EA and OntoEA) document zone classifiers are evaluated using two different resource meta-graphs, one derived from *sct* and one from *msh*. The main research question addressed are *"Do semantic meta-graphs built from KSs contain useful semantic features about entities for document zoning?" "Which KS ontology provides more useful information for document zoning?"*

In the second set of experiments, the effectiveness of the proposed adaptive OntoEA document zone classifier is compared against the baseline models in terms of annotation efficiency. The main research questions under investigation are *"To what extent does the OntoEA classifier exceed the performance of the baseline classifiers?" "How many annotated in-domain examples are required to build a reliable adaptive document zone classifier?"*

Finally, the third set of experiments examines different domain similarity measures for quantifying the accuracy of a document zone classifier. For this purpose, the correlation between the proposed domain similarity measures and document zone classifiers is computed. The goal of this analysis being to investigate the research questions *"Is it possible to predict the performance of a document zone classifier?" "Which domain similarity measure presents the highest correlations with the accuracy of a document zone classifier?"*.

5.6.4 Results and Discussion

5.6.4.1 The Usefulness of Semantic Meta-Graphs in Cross-domain Document Zoning

In-domain Scenario The goal of this first set of experiments is to investigate the impact of incorporating semantic KS-based features into an in-domain document zone classifier. For this purpose the SVM TGT_ONLY classifier built on in-domain data is assessed using semantic pivot features derived from *resource meta-graph*s and compared against two baseline features: BoW and BoE. The employed *resource meta-graph*s are generated from two biomedical KSs: *sct* and *msh* ontologies.

Considering that the proposed framework aims to investigate the usefulness of KS ontologies both independently and jointly, four different SVM TGT_ONLY classifiers have been created. The first two TGT_ONLY classifiers make use of individual KS ontologies: TGT_ONLY(*sct*) using semantic features from *sct*, and TGT_ONLY(*msh*) using semantic features from *msh*. The two other TGT_ONLY classifiers were using the combined KS ontologies: TGT_ONLY(*sct*+*msh*) and TGT_ONLY(*sct*+*msh* $_{CCA}$).

The results obtained for the semantic features using different feature augmentation and ontology combination strategies (as described in Subsubsection 5.3.6.1) in terms of F1 measure are summarised in Table 5.6[16]. For the sake of completeness it is mentioned here, that the results obtained for the upper class features are presented in the Section B.2, as they did not show a significant improvement compared to the semantic class features (the differences between the two features were less than 1% for F1.).

[16]The results obtained in terms of precision and recall are further presented in Section B.1.

TGT Domain	Semantic				Baseline	
	sct+*msh* $_{CCA}$	*sct*+*msh*	*sct*	*msh*	*BoW*	*BoE*
Biol	**0.718**	0.700	0.708	0.706	0.705	**0.709**
CellBiol	**0.815**	0.800	0.804	0.804	0.799	**0.805**
Communi	**0.663**	0.629	0.643	0.643	0.636	**0.641**
HealthS	**0.622**	0.592	0.603	0.603	**0.606**	0.604
Medicin	**0.711**	0.671	0.694	0.693	0.683	**0.693**
PublicH	**0.644**	0.619	0.632	0.627	0.619	**0.629**
Tropica	**0.653**	0.619	0.634	0.633	0.625	**0.633**
Average	**0.689**	0.661	0.674	0.673	0.668	**0.673**

Table 5.6: The performance of the SVM TGT_ONLY classifier in terms of F1 measure using semantic class (**Cls**) features extracted from two KS ontologies (*sct* and *msh*) and various baseline lexical features (*BoW*, *BoE*). The employed KS ontologies are evaluated both independently and jointly: *sct* and *msh* referring to the case when the individual KSs are used; *sct* +*msh* referring to the naive combination of semantic feature from the two KSs; and *sct* +msh_{CCA} corresponding to the scenario in which dimensionality reduction is applied over the semantic features of the two KSs. The best results obtained for the semantic and baseline features are shown in bold.

Inspecting the results obtained for the TGT_ONLY classifier using the baseline *lexical features* (*BoW*, *BoE*), it can be observed that the *BoE* features outperform the *BoW* features for most of the domains, although by a small margin. The best overall results further were obtained using *semantic features*. In particular, the TGT_ONLY classifier employing semantic class features from the *sct* and *msh* ontologies combined with dimensionality reduction (TGT_ONLY(*sct*+msh_{CCA})) achieved the best results. The highest improvement of 2.8% over the *BoW* features was achieved for the *Medicin* and the *Tropica* sub-domains (t-test, $p < 0.01$), while the smallest improvement of 1.3% was observed for the *Biol* ($p < 0.05$). Compared to the *BoE* features, the biggest gain of 2.2% was obtained for the *Communi* ($p < 0.01$), and the smallest gain of 1% for the *CellBiol* ($p < 0.05$).

Concerning the results obtained for the individual KS ontologies, the TGT_ONLY(*sct*) and TGT_ONLY(*msh*) classifiers achieved comparable results. The improvement of these classifiers on the baseline *BoW* features was also very small, less than 1% for F1.

Regarding the ontology combination strategies, it can be observed that the naive combination of domain ontologies did not perform very well. The performance of the TGT_ONLY (*sct* + *msh*) classifier achieved inferior results to the classifiers using individual KS ontologies (TGT_ONLY(*sct*), TGT_ONLY(*msh*)), and baseline features. An explanation for this may be that the two ontologies contain inconsistencies, which may well be addressed using CCA, confirming the superiority of the TGT_ONLY(*sct*+*msh* $_{CCA}$) classifier.

Cross-domain Scenario This section continues the discussion by describing the results obtained using semantic KS-based features in cross-domain document zone classification. In this case, the performance of the proposed SVM OntoEA cross-domain document zone classifier is compared against the SRC_TGT, TGT_ONLY, SRC_ONLY and EA baseline classifiers.

According to the different KSs and enrichment strategies employed, four different OntoEA classifiers were built: two using individual KS ontologies (OntoEA(*sct*) making use of

sct ontology and the OntoEA(*msh*) making use of *msh* ontology), and two using combined KS ontologies (OntoEA(*sct*+*msh*) and OntoEA(*sct*+*msh*)$_{CCA}$). In the same vein, four different versions of SRC_TGT and SRC_ONLY were created. These classifiers are evaluated against the SRC_TGT, EA and SRC_ONLY baseline models using lexical features in Table 5.7[17].

Analysing the results obtained for the baseline *lexical features*, it can be observed that the *BoE* features significantly outperform the *BoW* features for most domain pairs, considering all the SRC_TGT, EA and SRC_ONLY baseline classifiers ($p < 0.05$). These results are in agreement with the results obtained for the TGT_ONLY classifier. The biggest improvement on the *BoW* features can be observed for the SRC_ONLY classifier. In this case, a gain of over 10% is achieved in the majority of domain pairs (for 26 out of 42 domain pairs), and the highest gain of 55.9% is obtained for the *Medicin->Biol* domain pair. In the case of the SRC_TGT classifier, an improvement of over 5% can be seen for 20 domain pairs, and the highest improvement of 16.6% was observed for the *Biol->Medicin* pair. For the EA classifier, 23 domain pairs reached an improvement of 1.0-4.3% over the *BoW* features. The biggest improvement of 4.3% was obtained for the *Biol->HealthS* domain pair.

Looking at the *semantic class features*, similar to the TGT_ONLY classifier, the best results were achieved using semantic class features derived from multiple KS ontologies, combined with the CCA *ontology combination strategy*[18].

For all the three OntoEA, SRC_TGT, SRC_ONLY cross-domain classifiers, the (*sct*+*msh*)$_{CCA}$ ontology combination strategy significantly outperformed the *sct*+*msh* strategy, as well as the feature augmentation strategies employing a single KS ontology (*sct* and *msh*) ($p < 0.05$). The OntoEA (*sct*+*msh*)$_{CCA}$ classifier significantly outperformed the SRC_TGT(*sct*+*msh*)$_{CCA}$ classifier with an improvement of 2.4-7.5% (over 4% for 32 domain pairs), the EA(*BoE*) classifier with an improvement of 2.6-8.4% (over 4.0% for 38 domain pairs), the TGT_ONLY (*sct* + *msh*)$_{CCA}$ classifier with a gain of 1.1-6.3%, (over 3% for 28 domain pairs) and the SRC_ONLY(*sct*+ *msh*)$_{CCA}$ classifier with an improvement of 6.3%-32.8% (over 15% for 19 domain pairs). Further, the OntoEA (*sct*+*msh*)$_{CCA}$ classifier significantly outperforms the OntoEA classifiers using a single KS; an improvement of 2.8-7.2% can be observed against the OntoEA (*sct*) classifier, and of 3-7% against the OntoEA (*msh*) classifier.

Regarding the *sct* + *msh* ontology combination strategy, it can be seen that the performance of the cross-domain classifiers did not improve on the individual KS features (compare column 4 with 5 and 6; column 8 with 9 and 10, and column 16 with 17 and 18 in Table 5.7). These results are also in light of the results obtained for the TGT_ONLY classifier, indicating that the two KS ontologies may contain repetitions and redundant information about the concepts.

Comparing the individual KS ontology features, it can be noted that the performance of all three OntoEA, SRC_TGT and SRC_ONLY cross-domain classifiers is comparable using different *sct* and *msh* ontologies (compare column 5 with 6; column 9 with 10, and column 17 with 18 in Table 5.7). These results indicate that the two KS ontologies provide similar semantic information about biomedical concepts. However, when comparing the

[17]The results obtained in terms of precision and recall are further presented in Section B.1.

[18]For the sake of completeness, it is mentioned, that similar to the TGT_ONLY case, the results using upper-level concepts were only comparable to the results obtained using semantic class features, and for this reason those results are presented in Section B.2 only.

SRC Domain	TGT Domain	OntoEA				SRC_TGT						EA		SRC_ONLY					
		sct+msh CCA	*sct+msh*	*sct*	*msh*	*sct+msh CCA*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*	*BoW*	*BoE*	*sct+msh CCA*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*
Biol	*CellBiol*	**0.828**	0.789	0.795	0.796	**0.788**	0.774	0.779	0.771	0.695	**0.786**	0.787	**0.800**	**0.761**	0.733	0.737	0.740	0.348	**0.737**
	Communi	**0.689**	0.638	0.646	0.648	**0.642**	0.618	0.634	0.636	0.516	**0.636**	0.602	**0.637**	**0.524**	0.465	0.497	0.498	0.089	**0.498**
	HealthS	**0.634**	0.583	0.596	0.597	**0.594**	0.561	0.581	0.582	0.507	**0.588**	0.555	**0.598**	**0.425**	0.389	0.398	0.401	0.090	**0.397**
	Medicin	**0.760**	0.693	0.723	0.721	**0.707**	0.662	0.697	0.697	0.546	**0.702**	0.672	**0.699**	**0.640**	0.569	0.607	0.607	0.109	**0.607**
	PublicH	**0.662**	0.591	0.626	0.624	**0.623**	0.591	0.612	0.614	0.545	**0.614**	0.585	**0.616**	**0.467**	0.409	0.443	0.441	0.092	**0.443**
	Tropica	**0.684**	0.613	0.650	0.648	**0.639**	0.595	0.629	0.628	0.485	**0.635**	0.599	**0.628**	**0.543**	0.494	0.521	0.514	0.072	**0.524**
CellBiol	*Biol*	**0.770**	0.733	0.740	0.740	**0.704**	0.685	0.702	0.703	0.652	**0.700**	0.694	**0.702**	**0.644**	0.623	0.622	0.624	0.352	**0.623**
	Communi	**0.681**	0.647	0.640	0.641	**0.641**	0.614	0.631	0.633	0.599	**0.632**	0.610	**0.632**	**0.424**	0.392	0.402	0.401	0.321	**0.401**
	HealthS	**0.633**	0.594	0.601	0.600	**0.587**	0.570	0.580	0.579	0.577	**0.579**	0.578	**0.580**	**0.350**	0.333	0.327	0.329	0.274	**0.326**
	Medicin	**0.747**	0.694	0.709	0.710	**0.703**	0.658	0.694	0.697	0.665	**0.698**	0.679	**0.692**	**0.542**	0.494	0.512	0.513	0.372	**0.511**
	PublicH	**0.670**	0.608	0.631	0.631	**0.619**	0.602	0.611	0.614	0.603	**0.611**	0.596	**0.611**	**0.688**	0.633	0.357	0.358	0.289	**0.358**
	Tropica	**0.678**	0.623	0.647	0.646	**0.635**	0.599	0.626	0.626	0.602	**0.634**	0.611	**0.623**	**0.445**	0.414	0.420	0.420	0.329	**0.419**
Communi	*Biol*	**0.748**	0.693	0.719	0.718	**0.710**	0.688	0.705	0.707	0.632	**0.708**	0.695	**0.706**	**0.566**	0.542	0.535	0.542	0.055	**0.535**
	CellBiol	**0.828**	0.769	0.765	0.767	**0.787**	0.765	0.775	0.778	0.732	**0.753**	0.772	**0.777**	**0.607**	0.587	0.574	0.585	0.354	**0.575**
	HealthS	**0.676**	0.624	0.639	0.636	**0.605**	0.577	0.597	0.598	0.622	**0.600**	**0.599**	0.598	**0.531**	0.503	0.504	0.505	0.503	**0.538**
	Medicin	**0.764**	0.715	0.728	0.726	**0.703**	0.665	0.698	0.697	0.714	**0.701**	0.693	**0.698**	**0.645**	0.592	0.609	0.612	0.607	**0.609**
	PublicH	**0.696**	0.647	0.659	0.661	**0.632**	0.607	0.626	0.624	0.613	**0.621**	0.615	**0.627**	**0.606**	0.559	0.577	0.578	0.580	**0.578**
	Tropica	**0.710**	0.658	0.666	0.664	**0.646**	0.613	0.636	0.636	0.585	**0.640**	0.626	**0.635**	**0.635**	0.605	0.613	0.612	0.443	**0.612**
HealthS	*Biol*	**0.731**	0.675	0.683	0.682	**0.677**	0.666	0.660	0.661	0.622	**0.673**	0.664	**0.685**	**0.497**	0.488	0.477	0.475	0.072	**0.478**
	CellBiol	**0.828**	0.747	0.759	0.764	**0.776**	0.762	0.775	0.774	0.749	**0.750**	0.770	**0.775**	**0.500**	0.492	0.476	0.478	0.297	**0.477**
	Communi	**0.704**	0.645	0.662	0.665	**0.659**	0.619	0.642	0.640	0.655	**0.650**	**0.646**	0.643	**0.622**	0.561	0.589	0.593	0.560	**0.589**
	Medicin	**0.742**	0.683	0.698	0.700	**0.701**	0.665	0.693	0.691	0.690	**0.688**	0.689	**0.693**	**0.576**	0.518	0.536	0.539	0.527	**0.538**
	PublicH	**0.708**	0.648	0.669	0.673	**0.668**	0.607	0.630	0.626	0.655	**0.658**	0.616	**0.628**	**0.643**	0.598	0.616	0.618	0.610	**0.618**
	Tropica	**0.692**	0.645	0.653	0.652	**0.649**	0.610	0.639	0.638	0.588	**0.642**	0.627	**0.637**	**0.609**	0.567	0.582	0.578	0.421	**0.581**
Medicin	*Biol*	**0.763**	0.714	0.729	0.728	**0.710**	0.685	0.707	0.706	0.595	**0.719**	0.694	**0.707**	**0.661**	0.619	0.618	0.617	0.059	**0.618**
	CellBiol	**0.827**	0.794	0.792	0.792	**0.802**	0.763	0.798	0.802	0.755	**0.781**	0.793	**0.801**	**0.649**	0.623	0.610	0.615	0.378	**0.611**
	Communi	**0.721**	0.671	0.681	0.678	**0.659**	0.628	0.651	0.654	0.666	**0.671**	0.640	**0.651**	**0.652**	0.591	0.625	0.623	0.620	**0.624**
	HealthS	**0.673**	0.615	0.630	0.632	**0.606**	0.575	0.599	0.600	0.610	**0.619**	0.594	**0.599**	**0.503**	0.483	0.482	0.483	0.482	**0.505**
	PublicH	**0.683**	0.636	0.642	0.642	**0.629**	0.600	0.623	0.620	**0.640**	0.632	0.615	**0.624**	**0.588**	0.546	0.563	0.564	0.552	**0.559**
	Tropica	**0.697**	0.656	0.656	0.652	**0.641**	0.611	0.634	0.634	0.594	**0.648**	0.629	**0.633**	**0.633**	0.592	0.606	0.600	0.447	**0.601**
PublicH	*Biol*	**0.731**	0.688	0.686	0.686	**0.682**	0.669	0.682	0.680	0.610	**0.672**	0.682	**0.688**	**0.513**	0.496	0.488	0.492	0.075	**0.488**
	CellBiol	**0.828**	0.740	0.756	0.758	**0.775**	0.760	0.771	0.774	0.730	**0.746**	0.772	**0.773**	**0.537**	0.516	0.518	0.516	0.339	**0.515**
	Communi	**0.709**	0.654	0.663	0.666	**0.658**	0.620	0.650	0.648	0.664	**0.650**	**0.648**	0.646	**0.646**	0.590	0.617	0.614	0.627	**0.615**
	HealthS	**0.687**	0.638	0.652	0.648	**0.614**	0.586	0.604	0.604	0.637	**0.641**	0.603	**0.606**	**0.621**	0.576	0.597	0.594	**0.600**	0.596
	Medicin	**0.750**	0.692	0.709	0.709	**0.699**	0.661	0.692	0.690	0.696	**0.700**	0.691	**0.692**	**0.586**	0.547	0.555	0.551	0.550	**0.552**
	Tropica	**0.694**	0.648	0.654	0.653	**0.648**	0.609	0.634	0.637	0.590	**0.645**	0.628	**0.644**	**0.619**	0.572	0.592	0.591	0.437	**0.589**
Tropica	*Biol*	**0.750**	0.688	0.722	0.718	**0.712**	0.688	0.704	0.706	0.615	**0.703**	0.693	**0.704**	**0.609**	0.589	0.584	0.593	0.036	**0.586**
	CellBiol	**0.828**	0.770	0.765	0.762	**0.801**	0.761	0.799	0.800	0.761	**0.756**	0.771	**0.799**	**0.640**	0.614	0.607	0.613	0.358	**0.610**
	Communi	**0.723**	0.657	0.684	0.683	**0.661**	0.620	0.653	0.653	0.577	**0.645**	0.633	**0.653**	**0.659**	0.596	0.627	0.632	0.449	**0.631**
	HealthS	**0.681**	0.618	0.643	0.643	**0.606**	0.580	0.600	0.601	0.574	**0.600**	0.595	**0.599**	**0.552**	0.515	0.526	0.523	0.374	**0.527**
	Medicin	**0.764**	0.705	0.722	0.723	**0.710**	0.671	0.697	0.699	0.639	**0.703**	0.678	**0.697**	**0.637**	0.576	0.602	0.602	0.414	**0.604**
	PublicH	**0.694**	0.639	0.656	0.652	**0.631**	0.609	0.625	0.624	0.588	**0.627**	0.609	**0.624**	**0.590**	0.541	0.571	0.566	0.392	**0.571**
Average		**0.725**	0.671	0.684	0.684	**0.675**	0.645	0.666	0.666	0.628	**0.668**	0.656	**0.668**	**0.581**	0.541	0.546	0.546	0.356	**0.549**

Table 5.7: The performance of OntoEA, SRC_TGT, EA and SRC_ONLY classifiers in terms of F1 measure using semantic class (**Cls**) features extracted from two KS ontologies (*sct* and *msh*) and various baseline lexical features (*BoW*, *BoE*). The employed KS ontologies are evaluated both independently and jointly: *sct* and *msh* referring to the case when the individual KSs are used; *sct +msh* referring to the naive combination of semantic feature from the two KSs, and *sct* $+msh_{CCA}$ corresponding to the scenario in which dimensionality reduction is applied over the semantic features of the two KSs. The best results obtained for the individual classifiers are shown in bold.

performance of the OntoEA(sct) and OntoEA(msh) classifiers with the baseline EA(BoW) and EA(BoE) models, an improvement of 0.8-5.7% can be observed over EA(BoW), and 0.4-4.9% over EA(BoE) (except when porting to *CellBiol*). These results thus indicate that employing semantic class features from individual KS ontologies can also be beneficial in some scenarios, outperforming baseline models using lexical features.

Overall, considering the results obtained for both single-domain and cross-domain experiments, the following conclusions can be drawn:

- Resource meta-graphs built from KSs contain useful semantic features about entities for document zoning. In particular, incorporating semantic class features (**Cls**) from multiple KSs into both single-domain (TGT_ONLY) and cross-domain classifiers (OntoEA) gave a significant improvement over various state-of- the-art approaches.

- Combining the evidence about the semantic features from multiple biomedical KS taxonomies (SNOMED-CT and MeSH) via dimensionality reduction is beneficial for document zoning (OntoEA(sct+msh_{CCA})), showing a significant improvement over approaches considering a single KS (OntoEA(sct) and OntoEA(msh)).

5.6.4.2 Gain of the Proposed Model (OntoEA) over the Baseline Models

This section aims to investigate the gain obtained by the best adaptive classifier $OntoEA(sct+msh_{CCA})$ against the baseline classifiers ($TGT\_ONLY(sct+msh_{CCA})$, $SRC\_TGT(sct+msh_{CCA})$, $SRC\_ONLY(sct+msh_{CCA})$ and $EA(BoE)$), considering the performance of these models over the full performance curve[19].

Starting the comparison with the SRC_ONLY classifier, it can be seen that for all 42 adaptation scenarios, OntoEA consistently outperforms the SRC_ONLY classifier over the full performance curve. Using as little as 10% of the annotated in-domain data (corresponding to 32 annotated documents, under a 5-fold cross-validation setting), the biggest gain of 32.28% in F1 was obtained for the *HealthS*→*CellBiol* scenario ($p < 0.05$) (see Figure 5.5).

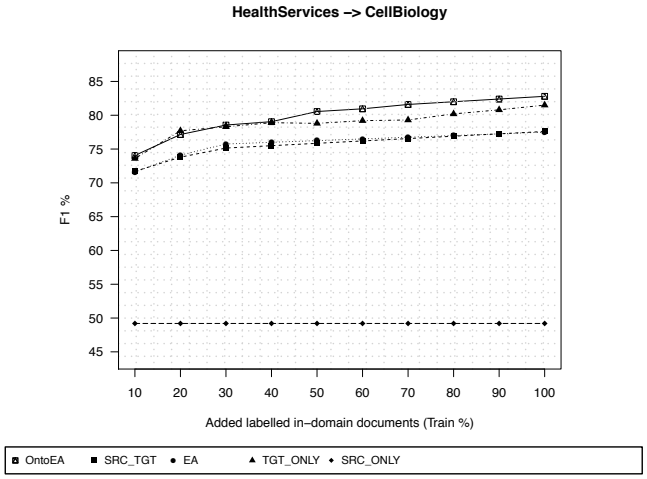

Figure 5.5: Performance of OntoEA on the *HealthS* → *CellBiol* domain pair. OntoEA consistently outperforms SRC_ONLY over the full performance curve.

[19]This section only reports some representative performance curves from the analysed 42 scenarios. The full list of performance curves are presented in Section B.3.

Concerning the other three baseline classifiers, three main trends were observed. In the first, OntoEA significantly outperforms all the three baseline models (SRC_TGT, TGT_ONLY and EA). This behaviour can be observed for most domain pairs: 31 out of 42 pairs[20]. An example for such a domain pair is shown in Figure 5.6.

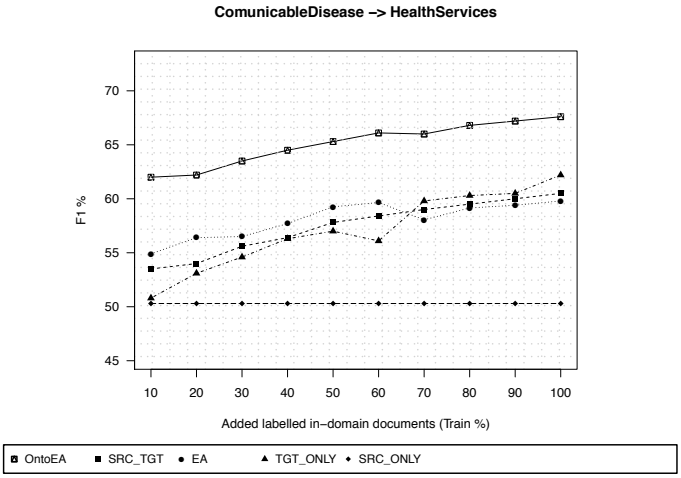

Figure 5.6: Performance of OntoEA on the *Communi→HealthS* domain pair. OntoEA significantly outperforms all the three baseline classifiers over the full performance curve.

The second most common trend corresponds to the situation when OntoEA performs as well as the SRC_TGT, TGT_ONLY and EA classifiers over the full performance curve. This case occurs for 8 domain pairs[21]. An example of such a domain pair is given in Figure 5.7.

Finally, the third trend represents a continuous gain of OntoEA classifier over the SRC_TGT, TGT_ONLY and EA models after a cutting point. This situation occurs for 3 domain pairs[22]. In such cases, the performance of OntoEA recovered using less than 50% of the annotated training data, showing a consistent improvement on all baseline models. For instance, for the *Medicin → CellBiol* pair only 30% of the annotated data was needed (see Figure 5.8), and for the *HealthS→CellBiol* pair, 40% was sufficient for OntoEA to outperform the baseline models.

In conclusion, considering the results obtained, the following conclusions can be drawn:

- The OntoEA classifier consistently and significantly outperforms the SRC_ONLY model over the full performance curve for all the adaptation scenarios.

- The OntoEA classifier is more effective than the TGT_ONLY, SRC_TGT and EA

[20]The complete list of domain pairs are the following: *Biol→CellBiol*, *Biol→Communi*, *Biol→Medicin*, *Biol→Tropica*, *CellBiol→Biol*, *CellBiol→Medicin*, *CellBiol→Tropica*, *Communi→Biol*, *Communi→HealthS*, *Communi→Medicin*, *Communi→PublicH*, *Communi→Tropica*, *HealthS→Communi*, *HealthS→Medicin*, *HealthS→PublicH*, *HealthS→Tropica*, *Medicin→Biol*, *Medicin→Communi*, *Medicin→HealthS*, *Medicin→PublicH*, *Medicin→Tropica*, *PublicH→ Biol*, *PublicH→Communi*, *PublicH→HealthS*, *PublicH→Medicin*, *PublicH→Tropica*, *Tropica→Biol*, *Tropica→Communi*, *Tropica→HealthS*, *Tropica→Medicin*, *Tropica→PublicH*. The corresponding performance curves are presented in Appendix B.

[21]The complete list of domain pairs are: *Biol→HealthS*, *Biol→PublicH*, *CellBiol→Communi*, *CellBiol→HealthS*, *Communi→CellBiol*, *HealthS→Biol*, *PublicH→CellBiol*, *Tropica→CellBiol*. The corresponding performance curves are presented in Appendix B.

[22]The complete list of domain pairs are the following: *CellBiol→PublicH*, *Medicin→CellBiol*, *HealthS→CellBiol*. The performance curves for the domains are provided in Appendix B.

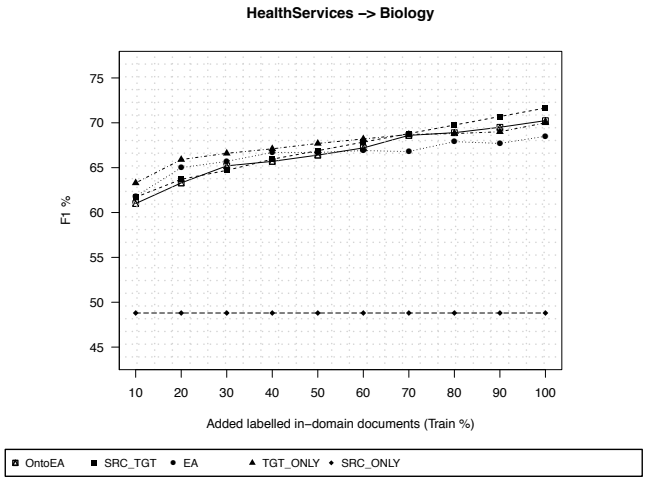

Figure 5.7: Performance of OntoEA on the *Tropica* → *CellBiol* domain pair. OntoEA achieves comparable results to SRC_TGT, TGT_ONLY and EA over the full performance curve.

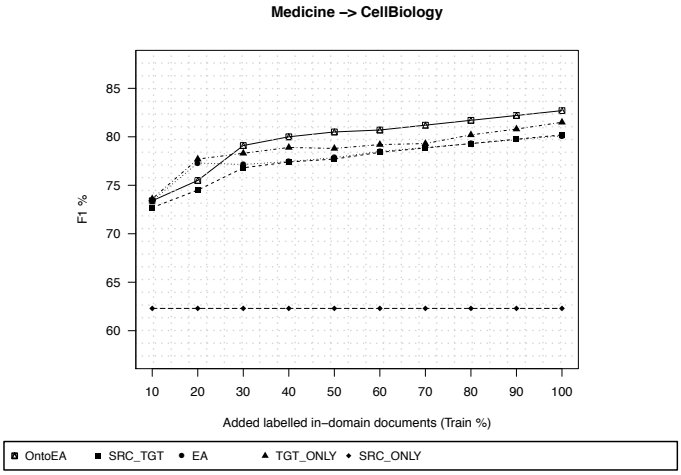

Figure 5.8: Performance of OntoEA against on the *CellBiol*→*PublicH* domain pair. OntoEA outperforms the baseline models after a cutting point.

classifiers, requiring a smaller number of annotations to achieve better results than these baseline models.

5.6.4.3 Evaluating Domain Similarity Measures for Document Zoning

As presented in the previous subsection, there is variation in performance levels between domains, suggesting that differences between the source and target domains affect the performance of a document zone classifier. In order to understand these variations, this subsection aims to investigate the similarity between the source and target domains at both lexical and conceptual levels, and to compute the correlation values between the similarity values and the performance in terms of F1 obtained by the OntoEA and TGT_ONLY classifiers. To assess the similarity measures, the *content-based lexical* measures and *hybrid measures* presented in Subsection 5.6.2 are employed. For the document zone classifiers, the best performing OntoEA($sct+\ msh_{CCA}$) and TGT_ONLY($sct+\ msh_{CCA}$) models are used.

Single-domain Scenario Figure 5.9 shows the correlations obtained for each domain between the similarity scores and the performance of the TGT_ONLY classifier at the end of the performance curve, utilising the full annotated training data. A positive correlation indicates that the performance of the classifier increases as the divergence decreases (the distributions are closer), while a negative correlation means that the performance increases as the divergence increases (the distributions are less similar). These figures show that among the *content-based similarity measures*, the KL divergence metric yields the best correlation scores. These scores exceed 60% for four domains. The second best score is the χ^2, followed by the *cosine* measure.

Among the *hybrid* measures, the combination of content-based KL and knowledge-based *lch* measures was found to achieve the highest correlation scores. The correlation scores in this case are relatively high, over 70% for all domains, lying between 71% and 84%. The second best correlation was obtained for KL+*wup*. This was followed by KL+*lin* and KL+*jcn* measures. Another important observation about these results is that, on average, all the four hybrid measures consistently outperform the three content-based similarity measures (with an improvement of 7.82-55.18% in absolute values). These results demonstrate that knowledge-based similarity measures play an important role in estimating the performance of an in-domain document zone classifier using KS-based features.

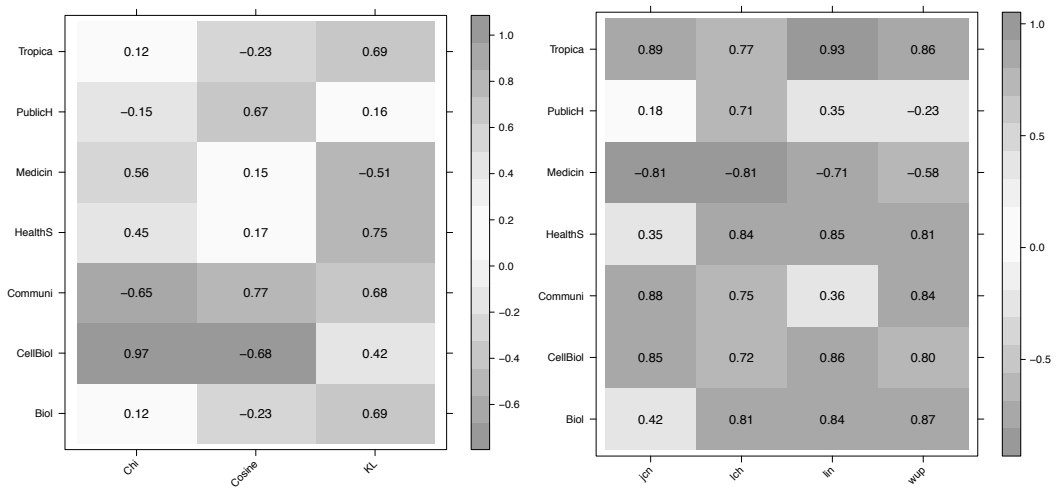

Figure 5.9: Pearson correlation values between the similarity measures and the performance of the TGT_ONLY (sct + msh_{CCA}) in-domain classifier using content-based (left), and hybrid measures (right).

Cross-domain Scenario The results obtained for the OntoEA model are presented in Figure 5.10. Of the *content-based* similarity measures analysed, the KL measure achieved the highest correlation values on average. This was followed by the χ^2 measure, and then the *cosine* measure. These results are thus in agreement with the results obtained for the TGT_ONLY classifier. The correlation values in this cross-domain scenario, however, are smaller: the values exceed 50% (in absolute terms) only in half of the adaptation scenarios (around 22 cases). One of the reasons for this behaviour could be that the different domains exhibit large lexical variations.

Focusing on the results obtained for the *hybrid measures*, however, the ranking of the

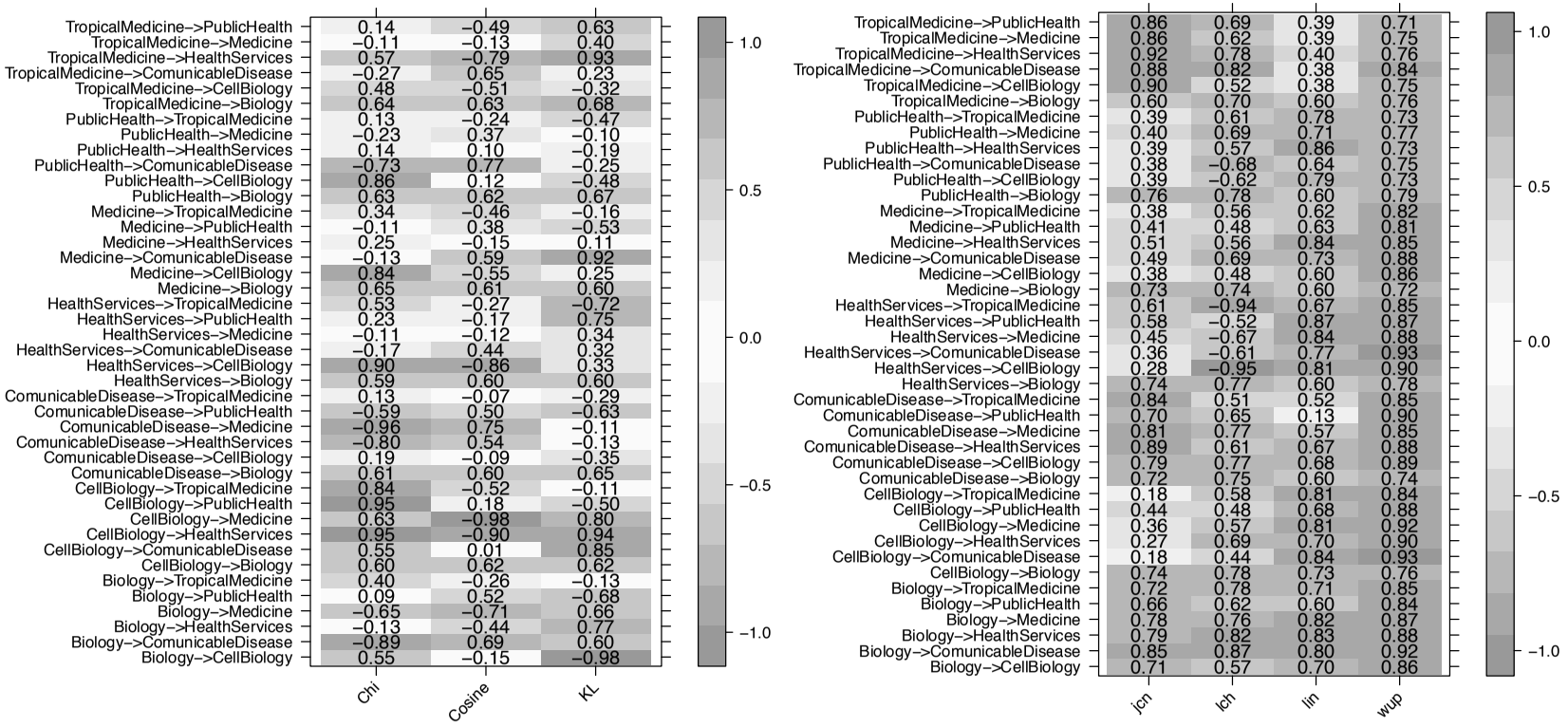

Figure 5.10: Pearson correlation values between the similarity measures and the performance of the OntoEA ($sct + msh_{CCA}$) cross-domain classifier using content-based (left), and hybrid (right) measures.

measures is slightly different to the results obtained for the TGT_ONLY classifier. In this case, the combined $KL+wup$ measure yields the best results. This is followed by the $KL+lch$, the $KL+lin$ and $KL+jcn$ measures. The correlation values for the $KL+wup$ values are reasonably high, over 70% (lying between 71% and 93%). Similar to the TGT_ONLY case, on average, all the *hybrid* measures outperformed the *content-based* similarity measures (with an improvement of 3.06-75.8% in absolute values). This further validates the importance of capturing the differences between domains over both the lexical and conceptual features. Another interesting observation is that, in the cross-domain scenario, it seems that the depth of the concepts plays a more important role than the path between the two concepts (resulted from the $wup > lch$) in capturing the closeness and similarity between domains. Further, the information content (IC) between concepts was also found to be less informative: both jcn and lin achieved inferior results to the wup and lch measures.

Given the above observations, the general findings about the hybrid measures are as follows:

- The performance of a document zone classifier can accurately be measured using hybrid measures, combining content-based and knowledge source-based measures defined over the concept graphs created from multiple KSs.

- These hybrid measures consistently and significantly outperform content-based lexical measures in both in-domain and cross-domain scenarios. The highest correlation values are obtained by the KL measure combined with path based-knowledge measures.

5.7 Possible Future Directions

The proposed adaptive document zoning framework has several advantages. Firstly, it *exploits the knowledge within domain-specific KSs to enrich the representation of the domains*, which is shown to reduce the lexical gap between domains. Secondly, *it makes use of the linked structure between KSs, allowing the incorporation of complementary information from KSs.* Thirdly, *it presents a principled way for the combination of multiple KSs.*
Despite of the success of this framework, several possible extensions could be explored:

- *Applying a different biomedical entity extractor:*
  The current transfer learning framework employs a particular entity extractor (the Bioportal API) for enriching the representation of the domain documents. The main advantages of this extractor are that it has a broad coverage of UMLS concepts and it provides semantic information about these concepts following linked data principles. The accuracy of this entity extractor largely determines the performance of the transfer document zone classifier, as it affects the number of correctly extracted entities, and the number of documents enriched with information about these entities.
  To date there have been different entity extractors developed for the biomedical domain, capturing specific biomedical entity types. The Open-Source Chemistry Analysis Routines (OSCAR) [Corbett and Copestake, 2008] specialised in four entity types including chemical molecule, chemical adjectives, enzymes and reaction. WhatIzIt[23]

[23] http://www.ebi.ac.uk/webservices/whatizit/info.jsf

captures six entity types: gene, protein, gene/protein, disease, drugs, metabolite. NeMine [Sasaki et al., 2008] covers more entity types such as phenomena, processes, organs and symptoms[24]. Despite of the popularity of these extractors, it is not very clear how do the performance of these different extractors compares, and whether BioPortal performs better than these extractors. A possible future direction in this regard could therefore be to conduct a systematic comparison of these different extractors to reveal the best performing one, and then to use that extractor within the proposed transfer learning framework.

- *Addressing the inconsistencies in biomedical knowledge sources:*
  Due to the manual process involved in the creation and maintenance of biomedical UMLS KSs, these KSs encounter some inconsistencies, as noted by the National Library of Medicine. For instance, Rector et al. [2011] pointed out that in the SNOMED-CT KS, diabetes is classified as a disease of the abdomen, while arteries of the foot is placed in the pelvis and myocardial infarction instead of being classified as ischemic heart disease. Similar topological inconsistencies concerning the MeSH KS were discussed by Antonio Jimeno Yepes [2013]. These inconsistencies affect the generalisation patterns learned between domains, and thus the performance of the cross-domain document zone classifier, as entities which should have been considered together are not grouped together. One possible solution to address this issue is to apply some preprocessing steps aimed at resolving these cases. For instance, the identification of MeSH concepts could be aided by considering the output of medical term indexers [Antonio Jimeno Yepes, 2013], while for correcting the inconsistencies in SNOMED-CT KS, different structural indicators could also be examined (such as the number of parents or the length of the concept's word).

- *Exploiting additional semantic structures and features for TC*:
  The proposed framework employs a particular semantic graph structure surrounding concepts in KSs, which makes use of the class information associated with concepts, and the is-a hierarchy among them. The analysed biomedical KSs, however, contain additional semantic information about concepts too, such as synonymy relationships or other domain-specific relationships (e.g. "due to", "causative agent"). Future work in this direction will thus consist of investigating whether these additional semantic relationships among concepts can further boost the performance of a cross-domain document zone classifier.
  Another possible future direction could also be to employ additional state-of-the-art features such as part-of-speech tags or cue phrases for enhancing the representation of domain documents for zoning. These features have been shown to improve the identification of zones for particular applications of document zoning within the same domain [Teufel and Moens, 2002]. However, to date no analysis has been conducted to investigate whether these features also help to recognise zones across multiple domains.

[24] A comparison of entity types covered by the different extractors is provided in Mihăilă et al. [2012].

5.8 Summary

This chapter presented a novel approach for the *within-document* TC of long documents, aimed at partitioning the documents into zone segments. The proposed approach introduced a modified version of the Easy Adapt (EA) algorithm, named OntoEA, which exploits the *semantic information* present in semantic concepts graphs derived from KSs, with the goal of reducing the distributional gap between domains. The feasibility of this approach was demonstrated by implementing document zone classification models that make use of semantic graph structures in multiple knowledge sources (SNOMED-CT and MeSH).

By exploring the research question *"Do semantic meta-graphs built from KSs do indeed contain useful semantic features about entities for document zoning?"*, it was found that resource meta-graphs created from both SNOMED-CT and MeSH KSs contain useful information for document zoning. The semantic class features extracted from these graphs serve as stable cross-domain features for adaptation. The best overall results were obtained by the combination of semantic features from these KSs using CCA dimensionality reduction. The proposed OntoEA($sct{+}msh_{CCA}$) model showed significant improvement upon the OntoEA model using a single KS, as well as several baseline models such as SRC_TGT, TGT_ONLY and SRC_ONLY models, and the EA model built on lexical features.

Through addressing the question *"How many annotated in-domain examples are required to build a reliable adaptive document zone classifier"*, it was demonstrated that the performance of OntoEA($sct{+}msh_{CCA}$) can reduce the human effort in annotating documents required for recognising zones in the target domain. For instance, in the case of the Comunicable Disease to HealthServices adaptation scenario, having 32 annotated documents (10% of the annotated data), OntoEA significantly outperformed the SRC_TGT, TGT_ONLY and EA baseline models over the full performance curve, achieving an F1 score of 67.6%.

These insights have provoked the final question *"Is it possible to predict the performance of a document zone classifier?"*. To address this question, several hybrid unsupervised domain similarity measures were introduced and evaluated over the concepts graphs. These results showed that the performance of OntoEA can be predicted with reasonably high accuracy (with a correlation above 70%) using the combination of KL divergence and *wup* path-based KS measure, significantly outperforming corpus-based similarity measures.

While classifying *long documents* from historical data can be useful for many applications, in order to obtain the big picture about an event, information often needs to be mined from other information sources as well. For instance, social media platforms have been found to provide up-to-date information about emerging events (e.g., emergency landings, natural disasters). The following chapter moves on to the presentation of *knowledge-driven* adaptive TC and domain similarity approaches for *short length* social media posts.

Chapter 6

Supervised Transfer Learning for Topic Classification of Social Media Posts

6.1 Introduction

The emergence of social media platforms (such as Twitter and Facebook) has allowed users to communicate news about emerging events (e.g., emergency landings, natural disasters and crimes) in a much faster way than traditional news sources. For instance, on April 29 2014, the first message about the emergency landing of Cobham Aviation flight appeared on Twitter[1]. Being informed about such events as they occur could be extremely important to authorities and emergency professionals as this would allow such parties to respond immediately.

While the previous chapters studied various adaptive text classification models for mining information from *long* documents archived in large repositories (e.g. corporate or scientific repositories), the main focus of this chapter is to mine information from *short documents* from social media platforms[2].

Dealing with *short* microposts to build supervised topic classification systems is a challenging task, due to the special characteristics of the messages: i) the *limited length* of microposts (limited to 140 characters), restricting the contextual information and cues that are available in normal long document corpora; ii) the *noisy lexical nature* of microposts, where terminology differs between users when referring to the same thing and abbreviations, misspellings and jargon are commonplace; iii) the *large topical coverage* of microposts, as the messages written by different users can cover a wide range of topics; iv) the *exponential increase in the rate of publication* of microposts, making the labelling of microposts difficult.

Linked Open Data cloud knowledge sources such as DBpedia and Freebase, however, provide an abundant source of structured data for a large number of topics which could potentially aid the topic classification of microposts. In particular, these sources exhibit

[1] http://www.theguardian.com/world/2014/apr/29/emergency-landing-at-perth-airport-after-suspected-engine-fire

[2] This framework has been designed in collaboration with Amparo E. Cano; all the experiments presented have been computed and analysed by the author of this thesis.

the following important characteristics: i) *they are constantly updated*; ii) *they cover a large number of topics*; and iii) *they provide a plentiful amount of annotated data for those topics*.

This chapter presents different transfer learning approaches which can exploit the information from KSs to build accurate topic classifiers of microposts. Firstly, these approaches make use of the *data* within these KSs as additional training data for building supervised topic classifiers, reducing the number of annotated tweets required. Further, due the short length of microposts, these approaches also exploit several graph structures surrounding concepts present in KSs to provide additional contextual information for microposts. In addition, this chapter also presents a study on the adaptability of a topic classifier, where a novel set of entropy-based measures are proposed for estimating the adaptability of a topic classifier making use of the enhanced document representation.

The remainder of the chapter is organised as follows: Section 6.2 reviews the state-of-the-art approaches in topic classification of microposts. Section 6.3 presents a novel framework for topic classification of *short* text messages using multiple KSs. Section 6.4 introduces a set of novel adaptability (or similarity) measures for topic classification. Section 6.5 describes the gold standard dataset used in the experiments. Section 6.6 evaluates the proposed adaptive topic classification models and domain similarity measures on a real-world dataset in the context of Emergency Response (ER) and Violence Detection (VD) domains. Finally, possible future extensions are described in Section 6.7.

6.2 Related Work on Topic Classification of Microposts

State-of-the-art approaches on topic classification of microposts can be divided into two main strands: approaches utilising a *single data source* (data from Twitter or blogs only) for topic classification (Subsection 6.2.1) and approaches utilising *knowledge sources* (such as DBpedia or Freebase) for topic classification (Subsection 6.2.2).

6.2.1 Single-domain Topic Classification of Microposts

The first class of approaches leverage information solely from the micropost content. They can be divided into the following sub-classes: *probabilistic graphical models* and *classification models*.

The first sub-class of approaches are based on *topic models*, which rely on the popular probabilistic Latent Dirichlet Allocation (LDA) model introduced in [Blei et al., 2003b]. Zhao et al. [2011] proposed an extended version of the LDA model, called TwitterLDA, which aims to detect the topics of short messages using only unlabelled data. Their approach relies on distinguishing between background words (words which occur in every topic), and content words (words specific to a topic). Experiments comparing TwitterLDA with traditional news media (e.g. New York Times) showed promising results, outperforming various other topic models.

Mehrotra et al. [2013] proposed various pooling schemas for improving the performance of the original LDA model for topic classification. These pooling strategies aim to aggregate microposts into longer documents (called "macro-documents"), which are more suitable for training LDA-based models. The pooling strategies evaluated were: author-wise pooling (pooling microposts according to an author), burst-score wise pooling (pooling microposts

according to a burst-score), temporal pooling (pooling microposts which are posted during major events by a large number of users), and hashtag-based pooling (pooling microposts according to a hashtag). Experimental results on three different datasets suggest that hashtag-based pooling leads to drastically improved topic modelling over unpooled schemes.

[Ramage et al., 2009; 2010], on the contrary, utilised annotated data for topic modelling. Ramage et al. [2009] introduced the LabelledLDA model, which extends the original LDA model by defining a one-to-one correspondence between LDA's latent topics and social media tags. Experimental results on a credit attribution problem, extracting tag-specific snippets from del.icio.us, were promising, outperforming supervised classifiers such as SVM. Ramage et al. [2010] further performed an extrinsic evaluation of the LabelledLDA model on a user recommendation task. In this case, the microposts were classified according to several dimensions including, e.g., style, substance, status, and other social characteristics of posts. Experiments showed promising results, achieving a performance comparable to those obtained using term frequency-inverse document frequency (TF-IDF) feature vectors built on tokenised microposts.

The second sub-class of approaches, *classification models*, are based on discriminative machine learning algorithms. Lin et al. [2011] proposed the combination of a language model with a supervised classifier for predicting the hashtags characterising a Twitter post. The features used for classification consisted of the perplexity of the unseen microposts. Tao et al. [2012] studied different topic-dependent and topic-independent features for topic classification. The topic-dependent features aimed to capture the relevance of the features to a topic (using keyword-based (lexical) and semantic-based relevance features). While the topic-independent features exploited various syntactic (e.g., hashtag) and semantic (number of entities, number of distinct entity types) micropost characteristics. Experimental results in the context of microblog search revealed that the topic-dependent features (the semantic relevance features) play an important role in this task, outperforming approaches which do not consider them.

6.2.2 Cross-domain Topic Classification of Microposts

The second branch of approaches exploit the information present in *individual KSs* (such as Wikipedia/DBpedia, Freebase and Probase) to detect the topic(s) of a tweet. The majority of these approaches employ *lexical* features (e.g.bag-of-words (BoW) or bag-of-entities (BoE)) extracted solely from the content of the documents (KS documents and micropost content).

Focusing on the approaches utilising the *Wikipedia or DBpedia KS* alone, Genc et al. [2011] proposed a model for mapping microposts to the Wikipedia articles most similar to them, employing a simple *BoW* representation for the text content. Their approach comprises two steps: mapping microposts to Wikipedia pages; and computing the semantic distance between microposts. For the computation of the semantic distance, a new measure is proposed, which approximates the distance between microposts by the link distance measure computed between the corresponding Wikipedia pages. Experimental results showed that this new distance measure outperforms the String Edit Distance [Levenshtein, 1966] and Latent Semantic Analysis [Dumais, 2004].

Shin et al. [2013] proposed a graph-based approach for detecting persistent topics (PT)

in microposts, which correspond to topics of long-term, steady interest to a user. For their graph based approach they introduced two novel scoring functions that measure the properties inherent to PT terms: regularity and topicality. They allow to distinguish between terms that represent persistent topics and terms which appear in static documents. Experiments showed that this approach outperformed other existing alternatives (including LDA and keyword extraction models).

Muñoz García et al. [2011] proposed an unsupervised approach for assigning topics to entities within microposts written in Spanish. Their approach first employs the Sem4Tags *POS* tagger [Garcia-Silva et al., 2010] to assign POS tags to a micropost. Following this process, a list of key phrases are identified, and the corresponding topics (DBpedia resource URIs) are assigned to them. This topic recognition phase exploits only local metadata, such as *BoW* features extracted from the keywords and contextual information in the form of neighbouring words to the keyword.

Vitale et al. [2012] proposed a clustering-based approach which enriches the *BoW* representation of the micropost using named entities extracted by the proposed Tagme system. The main idea behind Tagme is to assign the most likely topic to an entity, by taking into account the similarity between the topics returned by Tagme and Wikipedia categories for top-few categories. Experimental results showed that the approach incorporating these new *BoE* features into topic classification significantly outperformed approaches using *BoW* features only.

Pablo Mendes and Sheth [2010] proposed the Topical Social Sensor system, which allows users to subscribe to hashtags and DBpedia concepts to receive updates regarding these topics. Their approach relies on linking a micropost to DBpedia concepts derived from the entities contained with it. One of the main applications of the system is to detect the peak of a topic defined a priori.

Michelson and Macskassy [2010] proposed a model that discovers topics of interest of Twitter users based on their microposts. Their approach relies on first extracting and disambiguating the entities mentioned within a micropost. Following this process, a sub-tree of Wikipedia categories is retrieved for each entity and the most likely topic assigned.

Milne and Witten [2008] proposed an approach for assigning Wikipedia resources to key concepts within microposts. In their approach, a Wikipedia article is considered as a concept. Following this representation, a machine learning approach is presented, which employs different Wikipedia n-gram and Wikipedia link-based features.

Xu and Oard [2011] proposed a clustering-based approach which maps terms in microposts to Wikipedia articles. Their approach leverages the linking history of Wikipedia and the textual context information of terms to disambiguate the meaning of terms.

Meij et al. [2012] assign resources to microposts. In their approach, they make use of Wikipedia as a knowledge source, and consider a Wikipedia article as a *concept*. Their task then is to assign relevant Wikipedia article links to a tweet. They propose a machine learning approach which makes use of Wikipedia n-gram and Wikipedia link-based features.

Tao et al. [2012] studied various Twitter dataset-specific features (including whether a tweet contains a hashtag, or whether a tweet contains a URL) to identify whether a tweet is relevant to a topic, and showed that incorporating these features can help TC.

Looking at the approaches exploiting *Freebase KS*, Kasiviswanathan et al. [2011] proposed a clustering-based approach for topic detection, which makes use of entities and their

types gathered from Freebase. Subsequent work, classifying blog posts into topics [Husby and Barbosa, 2012], also demonstrated that selecting data from Freebase using distant supervision, in addition to incorporating features about named entities, is beneficial for topic classification.

There has been also work utilising *Probase KS* for topic classification. Song et al. [2011] proposed a probabilistic approach for mapping the terms within microposts to the most likely resources in Probase KS [Wu et al., 2012]. These resources were furthermore used as additional features in a clustering algorithm, achieving superior results to the simple *BoW* approach.

Although previous approaches have achieved a high level of success, they do still suffer from some limitations. Firstly, the vast majority of approaches still employ simple lexical features (BoW or BoE) derived solely from the content of the documents. Moreover, existing approaches [Muñoz García et al., 2011] consider the metadata of entities when detecting topics in microposts. The information is constrained by the NER service used (e.g., OpenCalais or Tagme), which often returns generic entity types [Rizzo and Troncy, 2011], ignoring more fine-grained semantic information described in external KSs. Secondly, these approaches still exploit a single KS when detecting topics in tweets, ignoring the possibility that multiple KSs may complement each other. Thirdly, these approaches often ignore the special characteristics of Twitter (such as URLs and hashtags). These Twitter-specific features could, however, provide additional information necessary for understanding the content of the messages.

Addressing the above limitations, the following section presents a cross-domain topic classification of microposts (Section 6.3), which exploits the information in multiple linked KS. Following this a novel set adaptability measures are also examined in Section 6.4.

6.3 Adaptive Topic Classification using Linked Knowledge Sources

This section describes a transfer learning framework for *adaptive text classification* of *microposts*, which is based on the framework introduced in Section 3.3. This framework follows a *supervised transfer learning setting*, in which a *large amount of annotated source domain documents*, and a *small amount of annotated target domain documents* are available. Compared to the case of *long documents* discussed in the previous chapter, the collection of annotated data for microposts constitute a very challenging task, due to the exponential rate of publication of these messages and the large coverage of topics discussed in them. In order to address these challenges and provide annotated data for training the classification models, in this framework, additional *data* from KSs (such as DBpedia and Freebase) is exploited.

The main stages of this framework can thus be summarised as follows: 1) *gathering of annotated data from KSs and domain content modelling* (Subsection 6.3.2); 2) *concept enrichment using KS ontologies* (Subsection 6.3.3); 3) *semantic meta-graph generation* (Subsection 6.3.4); 4) *pivot feature creation* (Subsection 6.3.5); 5) *building adaptive topic classifiers by employing different transfer learning techniques* (Subsection 6.3.6), 6) *evaluation of the adaptive topic classifiers on held out microposts*, depicted in Figure 6.1.

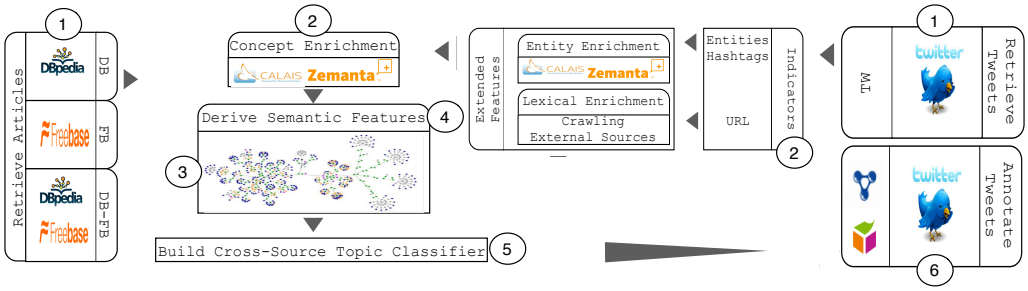

Figure 6.1: Architecture of cross-domain topic classifier using semantic features.

Before describing each of these individual steps in details, however, the motivation behind the selected KSs as well as an overview of their main characteristics are provided in Subsection 6.3.1.

6.3.1 Motivation

The Linked Open Data cloud consists of a large number of interlinked KSs, covering a range of different topics. Among these KSs, DBpedia and Freebase constitute some of the largest datasets built on a collaborative manner. These KSs contain factual information about a large number of entities of different domains, which is structured according to their own KS ontology. The relevance of these sources to Twitter is apparent, considering that both KSs and Twitter exhibit similar characteristics including for example that: i) *they are constantly edited by Web users*; ii) *their creation is done in a collaborative manner*; and iii) *they present a coverage on a large number of topics.*

Figure 6.2: Tweets exposing different contexts involving the same entity.

Some descriptive statistics about the DBpedia and Freebase KSs are depicted in Table 6.2. The first KS studied, DBpedia (*dbKS*)[3] is derived from Wikipedia[4]. In DBpedia [Bizer et al., 2009] each resource is harvested from a Wikipedia article which is semantically structured into a set of DBpedia[5] (*dbOwl*) and YAGO2[6] (*yago*) ontologies, with the pro-

[3]DBpedia, http://dbpedia.org
[4]Wikipedia, http://wikipedia.org
[5]http://wiki.dbpedia.org/Ontology
[6]http://www.mpi-inf.mpg.de/yago-naga/yago/

vision of links to external knowledge sources such as Freebase, OpenCyc[7], and UMBEL[8]. The Wikipedia articles are furthermore grouped into categories, which are represented using SKOS vocabulary[9]. The DBpedia dump version 3.8[10] classifies 2.35 million resources into DBpedia's ontology classes (*dbOwl*). These classes comprises 359 distinct classes, and 740,000 SKOS categories (*dbCat*), which form a subsumption hierarchy and are described by 1,820 different properties. Conversely, the *yago* ontology [Hoffart et al., 2012] is a much bigger and fine grained ontology. It contains over 447 million facts about 9.8 million entities which are classified into 365,372 classes, and 104 manually defined properties.

In contrast to DBpedia, Freebase[11] (*fbKS*) is a large online knowledge base which users can edit in a similar manner to Wikipedia. In Freebase [Bollacker et al., 2008], resources are harvested from multiple sources such as Wikipedia, ChefMoz, NNDB and MusicBrainz[12] along with data individually contributed by users. These resources are semantically structured into Freebase's own ontology (*fbOnt*), which consist of 1,450 classes and more than 7,000 unique properties.

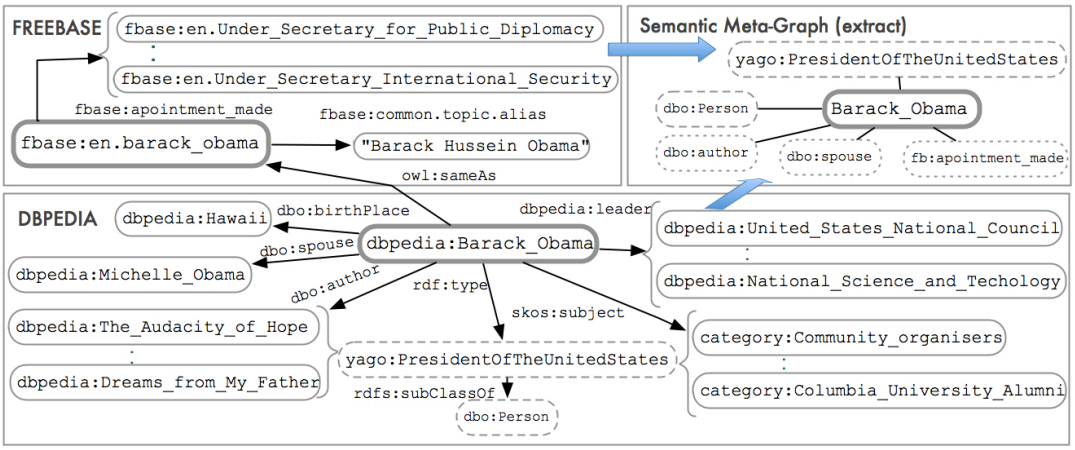

Figure 6.3: Deriving a semantic meta-graph from multiple KSs.

In summary, these ontologies (i.e. *dbOwl*, *yago*, *fbOnt*) enable a broad coverage of entities in the world, and allow entities to bear multiple overlapping types. One of the main advantages of exploiting these KSs is that each particular topic (e.g. `http://dbpedia.org/page/Category:Violence`) is associated to a large number of resources (e.g. `<http://dbpedia.org/page/Counter-terrorism>`), allowing to build a broad representation of a topic. In addition each resource is related to different ontological classes or concepts which provide additional contextual information for that resource, enabling in this way the exploitation of various semantic structures of these resources. The use of this structured knowledge enables the contextual enrichment of a micropost's entities by providing information that can help to disambiguate the role of a given entity in a particular context. Considering the tweets in Figure 6.2, although the entity *Obama* has different roles such as *president*, *Nobel laureate*, *husband*; the role of this entity will be defined by the contextual information

[7]OpenCyc, http://sw.opencyc.org/
[8]UMBEL, http://www.umbel.org/
[9]http://www.w3.org/2004/02/skos/
[10]This dataset was generated in 2012.
[11]Freebase, http://freebase.org
[12]Freebase Datasources, http://wiki.freebase.com/wiki/Data_sources

provided on the content of each tweet. Section 6.3.4 introduces the approach for leveraging this semantic contextual information by proposing the use of *semantic meta-graphs*.

6.3.2 Gathering Labelled Data from Knowledge Sources

The first step in the proposed framework consists of *collecting annotated data* for constructing the source domain data, and *creating an initial feature space for topic classification*.

For compiling a *source domain*, data is gathered from DBpedia and Freebase KSs. In particular, three different scenarios are investigated for building the *source domain*: i) **DB** - considering data from DBpedia only; ii) **FB** - considering data from Freebase only; and iii) **DB+FB** - considering data from both DBpedia and Freebase.

The data collection process consists in querying each KS for resources on specific topics, by accessing their publicly available APIs. In the case of DBpedia, for each analysed topic (e.g. Accident and Crime), SPARQL[13] queries are performed, which queries for all resources whose categories (*dcterms:subject*) and sub-categories (*skos:narrower*) correspond to the topic of interest. Table 6.1. shows some examples of the categories derived for the Accident and Crime topics.

Topic	DBpedia category
Accident (*DisAcc*)	*dcterms:subject* Category:Accidents *skos:narrower* Category:Aviation_accidents_and_incidents *skos:narrower* Category:Accidents_and_incidents_involving_airliners *skos:narrower* Category:People_involved_in_aviation_accidents_or_incidents ...
Crime (*Cri*)	*dcterms:subject* Category:Crime *skos:narrower* Category:Violent_crime *skos:narrower* Category:Crime_by_country *skos:narrower* Category:Crime_by_year ...

Table 6.1: Mappings between Topics of Microposts and DBpedia categories for some example topics.

Semantic Features	DBpedia (*dbKS*)			Freebase (*fbKS*)
	dbOwl	*dbCat*	*yago*	*fbOnt*
Resource	2.35×10^6		447×10^6	3.6×10^6
Property *(P)*	1,820		104	7,000
Class *(Cls)*	359	NA	365,372	1,450
Category *(Cat)*	NA	740,000	NA	NA

Table 6.2: Statistics about *dbOwl*, *dbCat*, *yago*, *fbOnt* KS ontologies.

In the case of Freebase KS, the Freebase Text Service API[14] is used to download the articles of the resources. The selection of resources being done, such that the domain name (the term used to describe topics in Freebase) of the resources matches the name of the topic of interest[15]. For some topics, such as Accident and Crime, for which there is no predefined

[13] http://www.w3.org/TR/rdf-sparql-query/
[14] http://wiki.freebase.com/wiki/Text_Service
[15] The collection of domains are enumerated at http://www.freebase.com/.

domain in Freebase, the selection of the articles is done based on the title of the article, ensuring that the selected articles' title contain these topics.

The Twitter dataset, serving as the *target domain* in this framework, furthermore consists of a set of tweets labelled with topics, which were collected between November 2010, and January 2011 [Abel et al., 2011].

Having both *source and target domain* data compiled, a simple BoW representation is employed for creating an initial feature space for topic classification. This representation allows both domains to be represented based on what it is discussed in the particular documents. In order to capture the importance of each word mentioned in these documents, furthermore, the TF-IDF weighting schema is applied.

6.3.3 Concept Enrichment

The second step of this framework aims to enrich the representation of both KS and Twitter documents using information about the entities and concepts mentioned in these documents.

In order to achieve this, two main steps are first performed: (i) *entity extraction* - employing the OpenCalais[16] and Zemanta[17] services for extracting the named entities in the documents; and (ii) *semantic mapping* - where the obtained named entities are mapped to their KS resource counterpart if it exists[18].

6.3.4 Semantic Meta-graph Generation

The mapping of entities to DBpedia and Freebase URIs allows the incorporation of rich semantic information into a topic classifier. In particular, the presented DB and FB KSs provide a rich source of information about concepts, and the exploitation of many useful structures about concepts.

Figure 6.3 presents an overview of the semantic features extracted for the entity "Barack Obama". Compared to the state-of-the-art approaches, rather than focusing on the <*subject, predicate, object*> instances associated with a resource, this framework focuses on each triple's semantic structure at a meta-level, and for that two meta-graphs are introduced: the *resource meta-graph* and the *category meta-graph*.

The first *resource meta-graph* exploits semantic information about an entity's KS resource. This semantic graph provides course-grained classification of entities by their types. The second graph is the *category meta-graph*, which exploits the semantic information extracted from the Wikipedia categories to which an entity belongs. This second graph can be effectively considered as a subset of the first one, as it groups similar entities belonging to the same topic under the same label. The *category meta-graph* thus categorises entities into more granular taxonomies.

Following the definition of *resource meta-graph* from Definition 4, let $\mathtt{G} := (R, P, C, Y)$ denote the *resource meta-graph* employed in this framework. The remainder of the reader, a resource meta-graph provides information regarding the set of ontologies, and properties used on the semantic definition of a given resource. The meta-graph of a given entity e can be represented as the sequence of tuples $G(e) = (R, P, C, Y')$, which is the aggregation of

[16]OpenCalais, http://www.opencalais.com
[17]Zemanta, http://zemanta.com
[18]Following this process, the percentage of entities without a deferenced URI is 35% in DBpedia, 40% in Freebase, and 36% in Twitter.

all resources, properties and classes related to this entity. In addition, two further notations can be introduced: $R(c) = \{e_1, \ldots, e_n\}$ for referring to the set of all entity resources whose *rdf:type* is class c; and $R'(c) = \{e_1, \ldots, e_m\}$ for denoting the set of entity resource whose type are specialisations of c's parent type (i.e. resources whose rdf:type are siblings of c).

Definition 6 (Category Meta-Graph) *A Category Meta-graph G_{cat} represents a qualified subset of the Resource meta-graph G in which all classes are of type skos:concept. This is defined as follows: $G_{cat} := (R, P, C', Y)$, where C' is a finite set whose elements are classes of type skos:Concept.*

Class	Category
dbOwl:Person	*dbCat*:Presidents_of_the_United_States
dbOwl:Author	*dbCat*:Obama_family
dbOwl:OfficeHolder	*dbCat*:Harvard_Law_School_alumni
yago:LivingPeople	*dbCat*:Democratic_Party_Presidents_of_the_United_States
yago:President	*dbCat*:United_States_presidential_candidates,_2012

Table 6.3: Top 5 features extracted from the DBpedia KS for the entity *Obama* of type Person.

For the sake of comparison, Table 6.3 and Table 6.4 present the top few class and category features derived from these graphs for two different entity types (*Obama* of type Person, and *Syria* of type Country). As it can be observed, the *dbCat* features group entities by topic, while the *dbOwl* features group entities by type[19].

Class	Category
dbOwl:Place	*dbCat*:Countries_of_the_Mediterranean_Sea
dbOwl:PopulatedPlace	*dbCat*:Arabic-speaking_countries_and_territories
yago:Country	*dbCat*:Eastern_Mediterranean_countries
yago:YagoGeoEntity	*dbCat*:Member_states_of_the_United_Nations
yago:MiddleEasternCountries	*dbCat*:Western_Asian_countries

Table 6.4: Top 5 semantic features extracted from the DBpedia KS for the entity *Syria* of type Country.

In light with the proposed three KS scenarios, three different *Resource meta-graphs* are constructed: (i) one from **DB** using the *dbOwl* and *yago* ontologies; (ii) one from **FB** using the *fbOnt* ontology; and (iii) another one from **DB+FB** using the joint ontologies. For the joint scenario the concepts from *dbOwl* ontology are used together with the the classes obtained after mapping the *yago* and *fbOnt* ontologies[20]. For the *Category meta-graph* a concept graph from DBpedia is derived, given that there is no category structure defined in Freebase. The three category concept graphs in this case correspond to (i) one from *DB* using the *dbCat* categories; (ii) one from *FB* using the *dbCat* categories obtained after mapping the FB URIs to DB URIs (iii) another one from *DB+FB* using *dbCat* categories.

[19]Further statistics about these semantic features are provided in Table 6.6.

[20]The mapping of Freebase entity classes to the most likely Yago classes was done by a combined element and instance based technique (`www.l3s.de/~demidova/students/master_oelze.pdf`) and is available at http://iqp.l3s.uni-hannover.de/yagof.html.

6.3.5 Pivot Feature Creation

Once a semantic meta graph has been constructed for a given entity, three main semantic features can be derived from it: *class*, *category* and *property* features. Among these features the *class* and *category* features are particular to a semantic meta-graph: *class* being extracted from the *resource meta-graph*, and *category* being derived from the *category meta-graph*; while the *property* features are common to both meta-graphs.

A description of each semantic feature can be given as follows:

Semantic class (Cls) features: Extracted from the *Resource meta-graph*, this feature set consists of all the classes an entity refers to. This set captures fine-grained information about this entity. E.g., for Barack Obama these features would be *yago:LivingPeople*, *yago:PresidentsOfTheUnitedStates*, *freebase:/book/author*, and *dbpedia:Person*. The main intuition here is that the relevance of an entity to a given topic could be inferred from an entity's class type. For example the class *yago:PresidentsOf-TheUnitedStates* could be consider more relevant to the topic "Politics" , than the class *yago:Singer*.

Semantic category (Cat) features: Extracted from the *Category meta-graph*, this feature set captures the Wikipedia categories an entity is related to. Similarly to the semantic classes, these categories provide additional fine-grained information about the entities, as entities about similar topics are grouped together in categories. For e.g. for Barack Obama these category features would be *cat:American_political_writers*, *category:People_from_Honolulu,_Hawaii*.

Semantic property (P) features: Common to both semantic meta-graphs, this feature set captures all the properties an entity is associated with. The intuition here is that given a context, certain properties of an entity can be more indicative of this entity's relevancy to a topic than others. For example, given the role of Tahrir Square in the Egyptian revolution, properties such as *dcterms:subject* could be more topically informative than *geo:geometry*. The relevance of a property to a given topic can be derived from the semantic structure of a KS graph by considering the approach proposed in Subsection 6.3.6.1.

6.3.5.1 Exploiting Twitter Specific Indicator Features

The above pivot features were derived by considering entities extracted from the sole content of the KS and Twitter documents. However, tweets can contain different Twitter specific *indicators* to external data sources (e.g. URL), which can provide additional *background information* necessary for understanding the content of the messages.

In order to exploit this additional background information, thus, two additional *indicators* are considered: i) *links indicators*, which consists of URLs posted within the tweets, ii) *hashtags indicators* (HSH), which are user generated markers (words or multi-word expressions) used to describe the topic of the tweets (e.g. #Egypt, #Obama). An example tweet highlighted with entities, links and hashtags is presented in Figure 6.4.

Corresponding to these *indicators*, two main *indicator features* can derived from the Twitter documents:

Bag-of-Links (BoL): The main rationale behind this feature set is to reduce the sparsity of microposts by incorporating lexical features derived from the external sources pointed

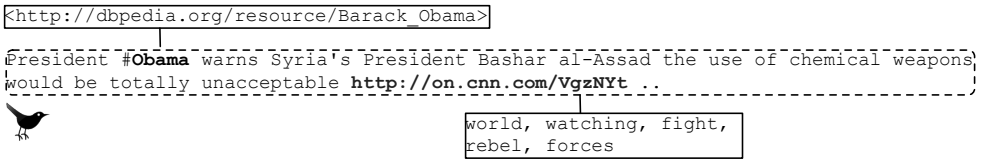

Figure 6.4: Enriching tweet content by using hashtags and links as indicators of external sources.

out by the *link indicators*. For this reason, a bag-of-links (BoL) is first constructed from the URIs mentioned within a tweet. Next each URI from the BoL is resolved and the content of the referenced web page parsed using the Jsoup API[21]. For each link then the following lexical features are created: i) *the title of the page (BoL(T))*: containing a concise description of the external URI content; ii) *the first paragraph of the page (BoL(1))*: introducing the main idea of the external URI content; and iii) *the last paragraph of the page (BoL(L))*: summarising the conclusions of the external URI content.

Bag-of-Hashtags (BoH): This feature set is viewed in a similar way as the entities mentioned within the documents. The main intuition here is that hashtags may refer to entities, which are described in the KSs, for e.g. #Egypt or #Obama. For this reason, first a bag-of-hashtags (BoH) is constructed from the hashtags mentioned within a tweet. Then for each hashtag, a semantic meaning is assigned. In order to achieve this, the output of the OpenCalais and Zemanta services is checked for potential DBpedia and Freebase URIs (e.g. #egypt will be associated with `dbpedia.org/resource/Egypt` and `freebase:Egypt`). If this process does not return any KS URI, then a series of SPARQL queries and regular expressions are fired, which are also manually revised and corrected if needed[22]. Once a KS URI for a hashtag is assigned, the new features for topic classification correspond to the *class* ($BoH(Cls)$), *category* ($BoH(Cat)$), and *property* ($BoH(P)$) features.

6.3.6 Building Adaptive Topic Classifier of Microposts

The final stage of the framework aims to build supervised topic classifiers corresponding to the different cross-domain scenarios, which make use of the generated KS pivot features. The Support Vector Machine (SVM) with polynomial kernel was selected as a base classifier, which will be detailed in Subsection 6.6.3.

For incorporating the presented pivot features into an adaptive topic classifier, this framework employs different *weighting strategies* for the pivot features and feature combinations (described in Subsubsection 6.3.6.1), as well as different *semantic augmentation* strategies for extending the initial feature spaces of both source and target domains (described in Subsubsection 6.3.6.2).

The goal of the feature *weighting strategies* is to capture the domain-dependent (specific) and domain-independent (generalisation) properties of the features, computed based on the structure of the KS ontologies. The *augmentation strategies* on the other hand provide alternative ways for the combination of the original lexical feature space and newly inferred

[21] http://jsoup.org

[22] Following this process, the percentage of hashtag entities without a dereferenced URI is 32.73% in DBpedia, and 32.79% in Freebase. The percentage of hashtag entities manually corrected is 5%.

semantic feature space.

Before describing each of these steps in details, for the sake of completeness, the Algorithm 5 provides an overview of the mains steps of the framework.

Algorithm 5 Adaptive *whole-document* topic classification exploiting multiple linked KSs

1: ***Input***: L_S labelled source domain (SRC) documents (consisting of DBpedia and Freebase articles), L_T labelled target (TGT) domain documents (e.g. microposts), U_T unlabelled target documents, F_S initial feature set of the source domain, F_T initial feature set of the target domain.
2: Merge the lexical feature set of the first domain (F_S) with the lexical feature set of the second domain (F_T) into a common lexical feature set ($F_{Lex} = F_S \cup F_T$)
3: Extract entities and concepts from both source and target domains
4: Exploit different semantic meta-graphs for the extracted concepts in both domains
5: Create semantic features from the semantic meta-graphs for both first F_{S_C} and target F_{T_C} domains
6: Merge the semantic feature set of the source domain (F_{S_C}) and target domain (F_{T_C}) into a common semantic feature set ($F_{SEM} = F_{S_C} \cup F_{T_C}$)
7: Exploit Twitter specific features indicators for the target domain
8: Create indicator features from these indicators $F_{T_{IND}}$ of the target domain
9: Merge the common lexical feature set (F_{Lex}) with the semantic feature spaces F_{SEM} and $F_{T_{IND}}$, ($F = F_{LEX} \cup F_{SEM} \cup F_{T_{IND}}$),
10: Augment the examples x from L_S with newly created semantic features according to F
11: Augment the examples x from L_T, U_T with newly created semantic features according to F
12: Train a supervised classifier (e.g. SVM) on the annotated examples from both sources ($L_S \cup L_T$)
13: ***Output***: Annotated U_T target data.

6.3.6.1 Semantic Pivot Feature Weighting

This framework proposes different weighting strategies for the semantic features derived from the entities and indicators: two of the weighting strategies capturing the importance of the features in the KS semantic meta-graphs; while the third one aiming to capture the relative importance of the features in the corpus. In addition to using these weights for quantifying the importance of pivot features for topic classification, these weights serve also another important role, that of filtering out irrelevant features for topic classification. This is done, by considering only the top few features with the highest weights for the incorporation into the classifier, as detailed in Section 6.5[23].

W-Freq: Semantic Feature Frequency: This weighting strategy provides a light-weight approach for weighting the ontological *class*, *category* and *property* features derived for both entities and Twitter specific hashtag indicators. It aims to enrich the feature space of a document (i.e KSs' article, or micropost) x by considering all the semantic meta-graphs extracted from the entity resources appearing in this document.

Formally, the frequency of a semantic feature f in a given document x with Laplace smooth-

[23]It is worth noting that this feature selection strategy also largely differs from state-of-the-art feature selection techniques [Forman et al., 2003] used in text classification, as they typically make use of the scores obtained for the features based on the text content only (e.g. occurrences of a feature in training positive- and negative-class training examples separately).

ing can be defined as follows:

$$W - Freq_x(f) = \frac{N_x(f) + 1}{\sum_{f' \in F} N_x(f') + |F|}, \tag{6.1}$$

where $N_x(f)$ is the number of times feature f appears in all the semantic meta-graphs associated to document x; and F is the semantic features' vocabulary. This weighting function captures the relative importance of a document's semantic features against the rest of the corpus; while the normalisation prevents bias towards longer documents.

While the **W-Freq** (semantic feature frequency) weighting function depends on the occurrences of features in a particular document, other generalised weighting information can be derived from a KS semantic structure to characterise a semantic meta-graph. To derive a weighted semantic meta-graph the following W-SG weighting strategy is proposed.

W-SG: (Class/Category)-Property Co-Occurrence Frequency: The rationale behind this weighting strategy is to model the relative importance of a property p (e.g. *dbOwl*: leader) to a given class *cls* (*yago*:President) or category *cat* (*dbCat*: United_States _presidential_candidates,_2012), together with the generality of the property in a KS's graph.

This weighting function computes how specific and how general a *property* is to a given *class* or *category* based on a set of semantically related resources derived from a KS's graph. In particular, given the semantic meta-graph of an entity e (i.e. $G(e)$), the relative importance of a property $p \in G(e)$ to a given class $cls \in G(e)$ in a KS graph $\mathcal{G}_{KS}$ can be computed by first defining the specificity of p to cls as follows:

$$specificity_{KS}(p, cls) = \frac{N_p(R(cls))}{N(R(cls))}, \tag{6.2}$$

where $N_p(R(cls))$ is the number of times property p appears in all resources of type cls in the KS graph $\mathcal{G}_{KS}$, and $N(R(cls))$ is the number of resources of type c in $\mathcal{G}_{KS}$. This measure captures the probability of the property p being assigned to an entity resource of type cls.

For example for the *Obama* entity, considering the *dbOwl*:leader property and *yago*:President class, the specificity value of *dbOwl*:leader in the DBpedia graph $\mathcal{G}_{DB}$ is computed as follows:

$$\begin{aligned} &\text{specificity\_DB(dbOwl:leader, yago:President)} \\ &= \{| <?headofstate, dbOwl:leader, ?leader>, \\ &<?headofstate, rdf:type, yago:President> \in \mathcal{G}\_DB|\}/ \\ &\{| <?headofstate, rdf:type, yago:President> \in \mathcal{G}\_DB|\} \end{aligned} \tag{6.3}$$

As indicated in Equation 6.2, the computation of the specificity value is independent of the entity e and differs according to the KS graph from which it is derived[24]. Higher specificity values indicate that the property p occurs frequently in resources of the given class cls.

Conversely, the generality measure captures the specialisation of a property p to a given class cls, by computing the property's frequency within other semantically related classes

[24]It might be worth mentioning that for each entity resource the specificity values for the properties are the same, capturing in this way the generalisation of the property for the same concept type.

$R'(cls)$. The generality measure of a property p to a class cls in a KS graph $\mathcal{G}_{KS}$ is defined, as follows:

$$generality_{KS}(p, cls) = \frac{N(R'(cls))}{N_p(R'(cls))}, \tag{6.4}$$

where $N(R'(cls))$ is the number of resources whose type is either cls or a specialisation of cls's parent classes. This measure captures the relative generalisation of a property p to a broader set of specialised sibling classes derived from cls, and its computation is independent of the entity e. In this case the generality of property *dbOwl*:leader given the class *yago*:President for the *DB* graph is computed as:

$$\begin{aligned}
&\text{generality_DB(dbOwl:leader, yago:President) =}\\
&\{| < yago:President,\ rdf{:}subClassOf,\ ?parent >,\\
&\quad <?group, rdf:subClassOf, ?parent >\\
&\quad <?agroup, rdf:type, ?group >\\
&\quad \in \mathcal{G}_DB|\}/\\
&\{| < yago:President,\ rdf{:}subClassOf,\ ?parent >,\\
&\quad <?group, rdf:subClassOf, ?parent >\\
&\quad <?agroup, rdf:type, ?group >\\
&<?agroup, dbOwl:leader, ?leader > \in \mathcal{G}_DB|\}
\end{aligned} \tag{6.5}$$

Higher generality values indicate that a property spans over multiple classes, and is less specific to a given class cls. These two measures (generality and specificity) of a property p to a given class cls are combined as follows:

$$W\text{-}SG(p, cls) = specificity(p, cls) \times generality(p, cls) \tag{6.6}$$

6.3.6.2 Semantic Augmentation

This section provides an overview of the semantic augmentation strategies supported by the framework. Examples for the various semantic features, feature combinations and semantic augmentation strategies employed for the entity *Obama* are provided in Table 6.5.

Semantic augmentation: This strategy (F'_{A1}) augments the initial lexical features (e.g *BoW* and *BoE* features) of the datasets with additional semantic information extracted for the entities appearing in them.

In the case of the *resource meta-graph*, for both **Cls** and **P** features, the original lexical feature set F has been extended with a set of unique **Cls** (including e.g., *dbOwl*:Author) and **P** (including for e.g. *dbOwl*:writer) features derived from this graph. In this case, the expanded feature space vocabulary size becomes $|F'_{A1_{Cls}}| = |F| + |F_{cls}|$ for the **Cls** features and $|F'_{A1_P}| = |F| + |F_p|$ for the **P** features, where $|F_{cls}|$ denotes the total number of unique class features added and $|F_p|$ denotes the total number of unique property features added. Furthermore, for the combined **Cls**+ **P** feature set this augmentation strategy creates the

augmentation strategy		feature name	feature value
F	Baseline	BoW	*Obama*
F'_{A1}	P(W-Freq)	P_1	$f_{W-Freq}(dbOwl{:}\text{leader})$
	P(W-Freq)	P_2	$f_{W-Freq}(dbOwl{:}\text{writer})$
	P(W-SG)	P_1	$f_{W-SG}(dbOwl{:}\text{leader}, yago{:}\text{President})$
	P(W-SG)	P_2	$f_{W-SG}(dbOwl{:}\text{writer}, dbOwl{:}\text{Author})$
	Cls(W-Freq)	Cls_1	$f_{W-Freq}(yago{:}\text{President})$
	Cls(W-Freq)	Cls_2	$f_{W-Freq}(dbOwl{:}\text{Author})$
	Cls+P(W-SG)	$Cls_1{+}P_2$	$f_{W-Freq}(yago{:}\text{President}), f_{W-SG}(dbOwl{:}\text{writer}, yago{:}\text{President})$
	Cls+P(W-SG)	$Cls_2{+}P_1$	$f_{W-Freq}(dbOwl{:}\text{Author}), f_{W-SG}(dbOwl{:}\text{leader}, dbOwl{:}\text{Author})$
F'_{A2}	parent(Cls)(W-Freq)	parent(Cls_1)	$f_{W-Freq}(yago{:}\text{HeadOfState})$
	parent(Cls)(W-Freq)	parent(Cls_2)	$f_{W-Freq}(dbOwl{:}\text{Thing})$
	parent(Cls)(W-Freq)+P(W-SG)	parent(Cls_1)+P_2	$f_{W-Freq}(yago{:}\text{HeadOfState}), f_{W-SG}(dbOwl{:}\text{writer}, yago{:}\text{President})$
	parent(Cls)(W-Freq)+P(W-SG)	parent(Cls_2)+P_1	$f_{W-Freq}(dbOwl{:}\text{Thing}), f_{W-SG}(dbOwl{:}\text{leader}, dbOwl{:}\text{Author})$

Table 6.5: Example semantic augmentation strategies for the entity *Obama* using semantic features derived from *resource meta-graph*. The first column stands for the augmentation strategies used to incorporate semantic features into a topic classifier, the second column provides example features to which the augmentation strategies are applied, while the third column gives examples of possible values for each such feature.

As possible semantic features two different features are considered: P_1, P_2 corresponding to top semantic property features, and Cls_1, Cls_2 referring to top semantic class features for *Obama*. These features are considered alone as well as in combination (for e.g. $Cls_1 + P_2$). For the sake of completeness, in the first row, the original feature space denoted by F, consisting of BoW features, is also presented. For this feature representation no augmentation strategy is applied.

For the semantic features further two different augmentation strategies are presented: F'_{A1} extending the F features with semantic features, and F'_{A2} augmenting the F features with semantic features derived from the class hierarchies of KSs (e.g. considering the parent classes of a class (parent(Cls))). For both augmentation strategies two different weighting strategies are presented: W-Freq corresponding to the semantic feature frequency weighting, and W-SG corresponding to the class–property co-occurrence weighting. When these strategies are applied for the feature combinations (e.g. $Cls_1 + P_2$), two additional features are added to the topic classifier (e.g. $f_{W-Freq}(yago{:}\text{President})$, $f_{W-SG}(dbOwl{:}\text{writer}, yago{:}\text{President})$).

novel feature set $F'_{A1_{Cls+P}}$, in which the feature set F is expanded with the properties' $< p, cls >$ tuple features derived from the semantic meta-graphs. In this case, the size of the expanded feature set is: $|F'_{A1_{Cls+P}}| = |F| + |F_p| \times |F_{cls}|$ (see the Cls_1+P_2 and Cls_2+P_1 examples in Table 6.5).

Similarly, for the *category meta-graph*, the expanded feature set becomes $|F'_{A1_{Cat}}| = |F| + |F_{cat}|$ for the **Cat** features, and $|F'_{A1_P}| = |F| + |F_p|$ for the **P** features. In this case, $|F_{cat}|$ refers to the total number of unique category features and $|F_p|$ denotes the total number of unique property features derived from this graph. Furthermore, for the combined **Cat**+**P** feature set this augmentation strategy creates the novel feature set $F'_{A1_{Cat+P}}$, in which the feature set F is expanded with the properties' $< p, cat >$ tuple features derived from this semantic meta-graph. In this case, the size of the expanded feature set is: $|F'_{A1_{Cat+P}}| = |F| + |F_p| \times |F_{cat}|$.

Semantic augmentation with generalisation: This augmentation strategy (F'_{A2}) aims to further improve the generalisation of a topic classifier by exploiting the subsumption relation among classes within the DBpedia or Freebase ontologies.

In the case of the *resource meta-graph*, the feature set F is enhanced with the set of parent classes of cls where $cls \in Cls$. Therefore the size of the enhanced feature set $F'_{A2_{Cls}}$ is computed as $|F'_{A2_{Cls}}| = |F| + |F_{parent(cls)}|$, where $|F_{parent(cls)}|$ denotes the total number of unique parent classes of cls. Similarly, the enhanced feature set $F'_{A2_{Cls+P}}$ which uses the **Cls**+**P** features is built by adding the $< p, parent(cls) >$ tuple features. The size of the $F'_{A2_{Cls+P}}$ is therefore: $|F'_{A2_{Cls+P}}| = |F| + |F_p| \times |F_{parent(cls)}|$, where $|F_{parent(cls)}|$ denotes the total number of unique $parent(cls)$ classes derived from this graph.

When applying this strategy over the *category meta-graph*, however, the subsumption relations among the SKOS categories are considered. In this case, the expanded feature set size for the **Cat** features is $|F'_{A2_{Cat}}| = |F| + |F_{parent(cat)}|$ and for the combined **Cat**+**P** features is $|F'_{A2_{Cat+P}}| = |F| + |F_p| \times |F_{parent(cat)}|$. In this case $|F_{parent(cat)}|$ stands for the number of unique parent SKOS classes of cat, and $|F_p|$ denotes the number of unique properties extracted from this *category meta-graph*.

6.4 Measuring the Topical Adaptability of Topic Classifiers

This section continues by investigating the benefit of employing the enhanced representation of the domains for measuring the adaptability of a topic classifier.

Understanding which semantic structures can improve the performance of an adaptive topic classifier could help in providing an estimation of the semantic adaptability of a KS graph to previously unseen lexical data. One such example could be, when wanting to apply the proposed framework on a different genre, longer posts e.g. blogposts or Facebook comments. Another situation could be, when wanting to build a topic classifier for a new topic (e.g. Politics), in which case one wants to have an a priori estimate about the similarity between KS data and Twitter data.

In light of the semantic features (f = {**Cls**, **P**, **Cat**}) and feature combinations (f = {**Cls**+**P**, **Cat**+**P**}) introduced in Subsection 6.3.5 a set of entropy-based measures are pro-

posed for topic similarity.

Entropy is an information theoretic measure which defines a probability distribution p[25] over a random variable X, capturing the dispersion of the variable f among the different classes in a given dataset T: $H_T(f) = -\sum_{f \in X} p(f) \log p(f)$. In the context of this framework, this measure was introduced as it allows to capture the semantic ambiguity and uninformativeness of a topic based on the entities mentioned in the documents and the KS structure[26]. That is, entities that are evenly distributed over multiple KS concepts/categories will have high entropy and thus topics mentioning these entities are less focused (more ambiguous) in the subject(s) they discuss.

A summary of the proposed measures can be given as follows:

1. **Topic-Class bag entropy** (Cls-Entropy): This was computed by taking the bag-of-classes for each topic derived from the *resource meta-graphs*, and measuring the entropy of that class bag, capturing the dispersion of classes used for a particular topic. In this context, *low entropy* indicates a focused topic, while *high entropy* indicates an unfocused topic, which is more random in the subjects that it discussed. This measure is defined as follows:
   $H_T(Cls) = -\sum_{j=1}^{|Cls_T|} p(cls_j) \log p(cls_j)$, where $p(cls_j)$ denotes the conditional probability of a concept cls_j, within the topic's concept bag Cls_T.

2. **Topic-Category bag entropy** (Cat-Entropy): This was computed by constructing the bag-of-categories for each topic derived from the *category meta-graphs*, and measuring the entropy of that category bag, capturing the dispersion of categories used for a particular topic. In this context, *low entropy* indicates a focused topic, while *high entropy* indicates an unfocused topic which is more random in the subjects that it discussed. This measure is defined as follows:
   $H_T(Cat) = -\sum_{j=1}^{|Cat_T|} p(cat_j) \log p(cat_j)$, where $p(cat_j)$ denotes the conditional probability of a category cat_j, within the topic's category bag Cat_T.

3. **Topic-Property bag entropy** (P-Entropy): This was computed by considering the bag-of-properties for each topic derived from the KS graphs, and measuring the entropy of that property bag, capturing the dispersion of properties used for a particular topic. In this context, *low entropy* indicates a focused topic, while *high entropy* indicates an unfocused topic which is more random in the subjects that it discussed. This measure is defined as follows:
   $H_T(P) = -\sum_{j=1}^{|P_T|} p(p_j) \log p(p_j)$, where $p(p_j)$ denotes the conditional probability of a property p_j, within the topic's property bag P_T.

4. **Topic-Entity bag entropy** (Entity-Entropy): This was computed by taking the bag-of-entities for each topic extracted by the named entity recogniser, and measuring the entropy of that entity bag, capturing the dispersion of entities used for a particular topic. In this context, *low entropy* indicates a focused topic, while *high entropy* indicates an unfocused topic which is more random in the subjects that it discussed. This

[25] In this thesis the shorthand notation p is used for $Pr(X = f)$. The capital P is reserved for the property features.

[26] Compared to previous content-based similarity measures (e.g. cosine), these measures can explicitly measure the informativeness of a topic by capturing the dispersion of the entities among different KS classes/categories according to the various semantic meta-graphs presented.

measure is defined as follows:
$H_T(Ent) = -\sum_{j=1}^{|Ent_T|} p(e_j) \log p(e_j)$, where $p(e_j)$ denotes the conditional probability of an entity e_j, within the topic's entity bag Ent_T.

5. **Entity-Class entropy** (EntityCls-Entropy): This measure is computed for each topic, by considering the bag-of-classes for each entity mentioned in a topic based on the extracted *resource meta-graphs*, capturing the dispersion of the entities in each classes. That is, *low entropy* indicates that the topic is less ambiguous, consisting of entities belonging to few classes, while *high entropy* refers to higher ambiguity at the level of entities.
$H_T(Cls|E) = -\sum_{j=1}^{|E_T|} p(e_j) H_T(Cls|E = e_j)$, where $p(e_j)$ denotes the conditional probability of an entity e_j within the topics' entity bag E_T, and $H_T(Cls|E = e_j)$ refers to topic class entropy given the entity e_j.

6. **Entity-Category entropy** (EntityCat-Entropy): In an analogy with the Entity-Class entropy, this measure is computed for each topic, by considering the bag-of-categories for each entity mentioned in a topic based on the extracted *category meta-graph*s. In this case, *low entropy* indicates that the topic is less ambiguous, consisting of entities belonging to few categories, while *high entropy* refers to higher ambiguity at the level of entities.
$H_T(Cat|E) = -\sum_{j=1}^{|E_T|} p(e_j) H_T(Cat|E = e_j)$, where $p(e_j)$ denotes the conditional probability of an entity e_j within the topics' entity bag E_T, and $H_T(Cat|E = e_j)$ refers to topic category entropy given the entity e_j.

7. **Entity-Property entropy** (EntityProperty-Entropy): Similarly, the bag-of-properties is taken for each entity mentioned in a topic based on the extracted KS graphs. In this context, *low entropy* indicates that the topic is less ambiguous, consisting of entities being associated to few properties, while *high entropy* refers to higher ambiguity at the level of entities.
$H_T(P|E) = -\sum_{j=1}^{|E_T|} p(e_j) H_T(P|E = e_j)$, where $p(e_j)$ denotes the conditional probability of an entity e_j within the topics' entity bag E_T, and $H_T(P|E = e_j)$ refers to topic property entropy given the entity e_j.

8. **Class-Property entropy** (ClsProperty-Entropy): This was measured by taking the bag-of-properties for each class appearing in each topic derived from the *resource meta-graphs*. In this context, *low entropy* indicates that a topic is less ambiguous, few properties spanning over multiple classes, while *high entropy* reveals high property diversity. The corresponding measure is defined as followed:
$H_T(P|Cls) = -\sum_{j=1}^{|Cls_T|} p(cls_j) H_T(P|Cls = cls_j)$, where $p(cls_j)$ denotes the conditional probability of a class cls_j within the topics' class bag Cls_T, and $H_T(P|Cls = cls_j)$ refers to topic property entropy for the class cls_j.

9. **Category-Property entropy** (CatProperty-Entropy): The category property entropy for each topic was computed in a similar way. In this context, *low entropy* indicates that a topic is less ambiguous, few properties spanning over multiple categories, while *high entropy* reveals high property diversity. The corresponding measure is defined as followed:

$H_T(P|Cat) = -\sum_{j=1}^{|Cat_T|} p(cat_j) H_T(P|Cat = cat_j)$, where $p(cat_j)$ denotes the conditional probability of a category cat_j within the topics' class bag Cat_T, and $H_T(P|Cat = cat_j)$ refers to topic property entropy for the category cat_j.

Considering that the goal is to estimate the performance of a topic classifier on a new unseen test dataset, the *entropy difference* (DE) measure is furthermore defined for capturing the differences between a training dataset -used to train a topic classifier-, and a test dataset -used to test a topic classifier. Let T_{train} and T_{test} be the probability distributions estimated from the training and test datasets. For instance, given the *Cri* topic, and the cross-domain topic classifier built on DBpedia KS data, the T_{train} training dataset corresponds to a dataset collected for the *Cri* topic from DBpedia, while the T_{test} dataset corresponds to the dataset collected from Twitter. According to the above entropy measures, for each semantic feature (e.g. $f = P$) and feature combination (e.g. $f = Cat + P$)[27], the entropy difference measure is defined as follows:

$$DE(f, T_{train}, T_{test}) = |H_{T_{train}}(f) - H_{T_{test}(f)}|. \tag{6.7}$$

Intuitively, having features (e.g. **Cls** or **Cat**) with low DE values means that the features have similar values with respect to the train and test datasets. It is also expected that the lower the DE values are, the better the performance of a topic classifier.

These measures will be examined in Section 6.6.4.4 by correlating them with the performance of different topic classifiers. The proposed framework was evaluated on the Emergency Response and Violence Detection domains. The following section introduces the datasets in which the proposed framework and topic adaptability metrics were tested.

6.5 Compiling a Gold Standard for Cross-Domain Topic Classification of Tweets

To analyse the impact of utilising semantic features in building adaptive topic classifier of microposts, the performance of the proposed strategies is evaluated using a large corpus of microposts and two large coverage linked KSs, namely DBpedia and Freebase. For evaluating the topic classification framework the Emergency response (ER) and Violence detection (VD) domains are considered, and thus relevant dataset to these domains are compiled.

The Twitter dataset (TW) was derived from Abel et al.'s dataset [Abel et al., 2011], comprising microposts collected from over a period of two months starting on November 2010. This dataset has been topically annotated with 17 OpenCalais topics[28], including e.g., the following topic labels: "War & Conflict" (*War*), "Law & Crime" (*Cri*) and "Disaster & Accident" (*DisAcc*). This collection has been manually re-annotated, ensuring to have 1,000 microposts for each of these topic labels. These microposts served as positive examples for each topic in this dataset. In order to mimic the imbalance issue posed on the detection

[27]For clarity it is mentioned here that for the feature combinations (e.g. $f = Cat + P$) is employed the conditional entropy measure (e.g. $H_{T_{train}}(f = P|Cat)$), as this provides a natural way for capturing the relationships among multiple semantic features.

[28]The full list of topics include: Business & Finance, Disaster & Accident, Education, Entertainment & Culture, Environment, Health & Medical & Pharma, Hospitality & Recreation, Human Interest, Labor, Law & Crime, Politics, Religion & Belief, Social Issues, Sports, Technology & Internet, Weather and War & Conflict.

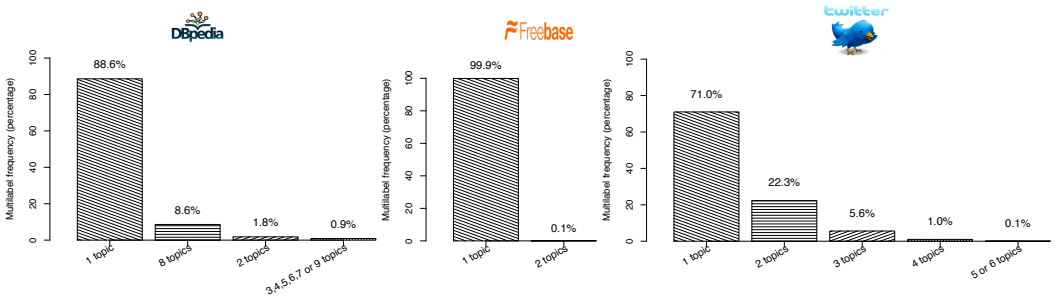

Figure 6.5: The multi-label distribution of the three gold standard datasets: DBpedia, Freebase and Twitter datasets. The numbers on the x axis represent the number of topics assigned to a document, ranging from 1 topic to 9 topics. The numbers on the y axis correspond to the percentage of documents labelled with different topics.

of microposts in this domain, in which a large proportion of microposts in a stream might be irrelevant to the topic of interest; a negative dataset was also built comprising a large collection of microposts which do not bear any relation to these three topics (i.e. *War*, *DisAcc* and *Cri*).

The Twitter dataset comprises 10,189 microposts annotated with up to six topic labels. The distribution of the examples belonging to multiple topics in each dataset is shown in Figure 6.5. In the Twitter dataset the majority of microposts are annotated with only one topic. In the case of the Freebase dataset, due to the nearly flat hierarchical structure of the domains, the majority of the articles belong to a single topic. In the case of the DBpedia dataset the majority of the articles belong to a single topic, and less than 1% of the articles are annotated with 3,4,5,6,7 or 9 topics.

Some notable events related to violence and ER discussed within these datasets include among others the "Mexican drug war", "Egyptian revolution", "Iranian Stoning Sentence", and "Indonesia Volcano Eruption". The DBpedia and Freebase topic datasets have been created as described in Subsection 6.3.2, by SPARQL querying these endpoints for all resources belonging to categories and subcategories of the skos:concepts of *War*, *DisAcc* and *Cri* respectively; keeping the resource's abstract or title as a document labelled with the given topic. Following this process, the final DBpedia dataset comprises of 9,465 articles, and the Freebase dataset consists of 16,915 articles.

Figure 6.6, 6.7 and 6.8 present the distribution of the top 15 entity types in the three datasets. As it can be observed the most frequent entity types are Country, Person, Organization, Natural Feature, Position and City.

In the *pre-processing* step, different steps were applied for both *lexical* and *semantic features*. When considering the *lexical features*, in order to obtain the BoW feaures for each document (i.e KS-derived article or micropost) the following steps were performed: removal of stopwords, lovercasing of words, stemming using Lovins stemmer [Lovins, 1968]. In addition, all Twitter-specific hashtags, mentions and URLs were removed, allowing to reduce the vocabulary differences between the KSs and TW datasets. The feature spaces were also reduced to the top-1000 words weighted by TF-IDF for each topic, which was found to perform better than using all the words.

When *obtaining semantic features*, using the BoE features derived from a document, SPARQL queried were fired for each entity's resource in DBpedia and Freebase. From these

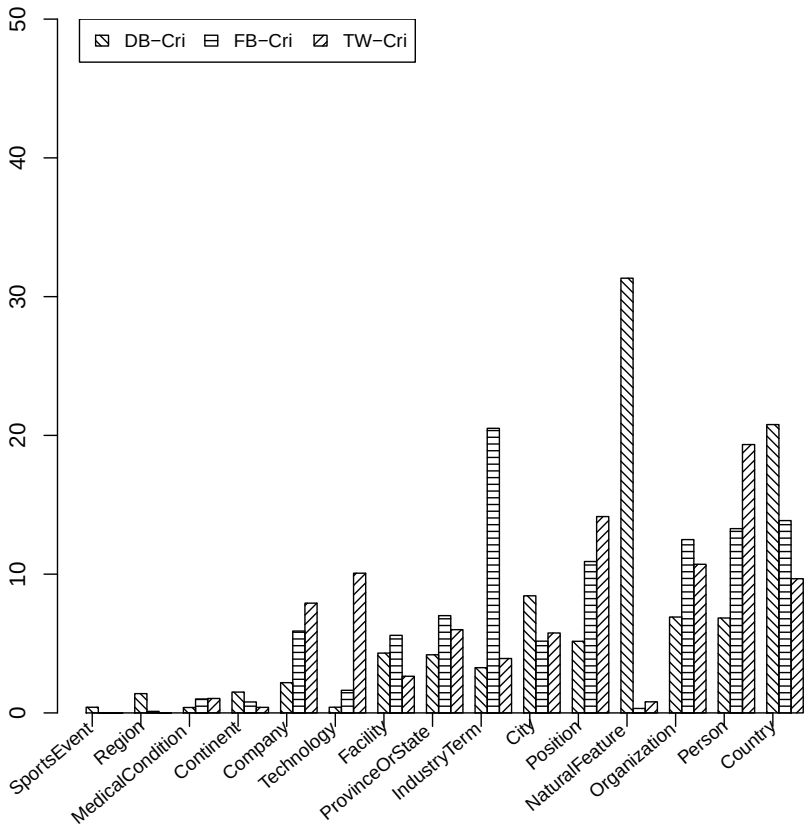

Figure 6.6: The distribution of top 15 entity types in the three gold standard datasets: DBpedia (DB), Freebase (FB) and Twitter (TW) datasets for the Crime (*Cri*) topic.

resources different semantic meta-graphs were built from each KS graph (i.e. $\mathcal{G}_{DB}$ and $\mathcal{G}_{FB}$) as indicated in Subsection 6.3.5. In addition, from both DBpedia and Freebase KS graphs some properties were disregarded containing general information about an entity (i.e common to each instance) e.g. rdfs:comment, abstract, wikiPageExternalLink from DbPedia and *type/object* from Freebase. These feature spaces were also reduced by considering for each entity type defined by OpenCalais (e.g. Person) the top 5 entity classes and top 5 properties derived from the different KS graphs. The same strategy is used for reducing the number of the category features, by selecting the top 5 for each OpenCalais entity type.

The statistics of the lexical and semantic features derived for these datasets are summarised in Table 6.6.

The *BOW* and *BOE* represents the size of vocabulary of the BOW and BOE features. *dbClass*, *yagoClass* and *fb* stand for the unique number of classes extracted from the DB, FB and DB+FB knowledge graphs; *dbCat* refers to the unique number of categories extracted from DB. *dbprop* counts the number of unique *dbOwl* properties, correspondingly *fbprop* counts the number of unique *fbOnt* properties. *cls/ent* refers to the average number of *dbOwl* and *yago* classes per entity; *cat/ent* stands for the average number of *dbOwl* categories, while *fbcls/ent* denotes the average number of *fbOnt* classes per entity. Similarly *prop/ent* denotes the average number of *dbOwl* and *yago* properties per entity, and

	Statistics	*DB*			*FB*			*TW*		
		DisAcc	*Cri*	*War*	*DisAcc*	*Cri*	*War*	*DisAcc*	*Cri*	*War*
Lex	BoW	8,837	8,837	8,504	2,078	4,596	2,574	3,218	3,197	2,781
	BoE	18,247	18,247	18,167	1,172	2,715	1,822	1,818	1,816	2,146
Semantic	dbCls	119	119	124	39	47	48	80	85	68
	yagoCls	3,865	3,865	3,864	351	834	922	1,480	1,795	1,275
	fbCls	1,289	1,289	1,215	394	713	641	881	915	772
	dbCat	9,275	9,275	8,796	783	1,844	1,807	3,252	3,878	3,087
	dbprop	4,105	4,105	4,215	1,229	1,849	1,871	2,544	2,457	2,422
	fbprop	1,090	1,090	1,065	420	586	554	834	869	696
	cls/ent	4.56	4.56	4.48	5.55	4.21	6.33	5.73	6.02	5.80
	cat/ent	5.45	5.49	5.34	7.76	5.80	8.89	7.49	8.20	8.72
	prop/ent	26.56	26.56	26.29	39.65	33.97	41.78	36.99	32.62	36.17
	fbcls/ent	7.30	7.30	7.12	15.89	12.68	15.57	11.98	11.66	12.49
	fbprop/ent	10.08	10.08	9.76	23.44	17.06	23.05	16.93	16.65	17.97
Indicator	%HSH							1.85%	2.78%	2.10%
	#HSH							233	220	198
	%URL							2.59%	6.41%	2.65%
	#URL							154	411	139
	dbCls(HSH)							29	23	20
	yagoCls(HSH)							150	169	171
	fbCls(HSH)							316	312	215
	dbCat(HSH)							29	23	20

Table 6.6: General statistics for the DBpedia (DB), Freebase (FB) and Twitter (TW) datasets used in the context of ER and VD for the two semantic meta-graphs analysed (*resource meta-graph* and *category meta-graph*). The rows labelled as *BoW* and *BoE* represent the size of the vocabulary of the *BoW* and *BoE* (without *BoW*) features. Statistics about the *resource meta-graph*: dbCls, yagoCls and fbCls stand for the unique number of classes extracted from the DBpedia and Freebase knowledge graphs. dbprop counts the number of unique DBpedia properties, and correspondingly fbprop counts the number of unique Freebase properties. Considering the *category meta-graph*: dbCat refers to the unique number of categories extracted from DBpedia knowledge graph.
cls/ent refers to the average number of *dbOwl* and *yago* classes per entity; cat/ent quantifies the average number of *dbCat* categories per entity, while fbcls/ent denotes the average number of *fbOnt* classes per entity. Similarly prop/ent denotes the average number of *dbOwl* and *yago* properties per entity, and fbprop/ent refers to the average number of *fbOnt* properties per entity.

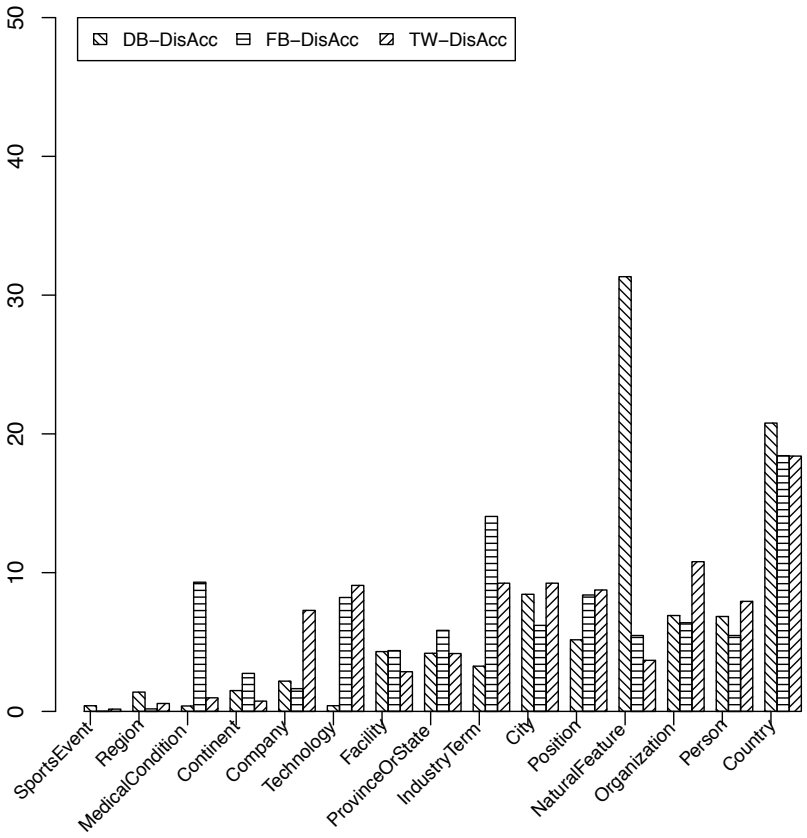

Figure 6.7: The distribution of top 15 entity types in the three gold standard datasets: DBpedia (DB), Freebase (FB) and Twitter (TW) datasets for the Disaster (*DisAcc*) topic.

$fbprop/ent$ refers to the average number of *fbOnt* properties per entity.

Comparing these statistics, it can be observed that the frequency of *dbOwl* categories (*dbCat*) are generally higher than those of *dbOwl* and *yago* classes. In addition, the average number of distinct categories (cat/ent) for an entity is mostly double the number of distinct classes per entities (cls/ent), indicating that the categories form much larger clusters than the classes.

In addition, it can be mentioned here that the DBpedia dataset contains the most number of entities for each topic, on average 22.24 entities per articles; while the number of documents without any entity is 69 (0.72%). In the case of Freebase, the average number of entities per article is 8.14, and the percentage of articles without any entity is 19.96% (3,377 articles). Lastly, the Twitter dataset consists of informative microposts mentioning at least one entity, the average number of entities per tweet is 1.73. In addition, it can observed that in all the three datasets the number of unique categories is higher than the number of unique classes, indicating that the datasets are more diverse in terms of categories than in terms of classes.

After concept generalisation, the number of unique *dbClass* classes reduces by 76%, the number of unique *yagoClass* classes reduces by 92%, and the number of unique *fbClass* classes by 88%. While in the case of category generalisation, the number of unique *dbCat* classes reduces by 42%.

Looking at the statistics about the Twitter specific indicators summarised in Table 6.6,

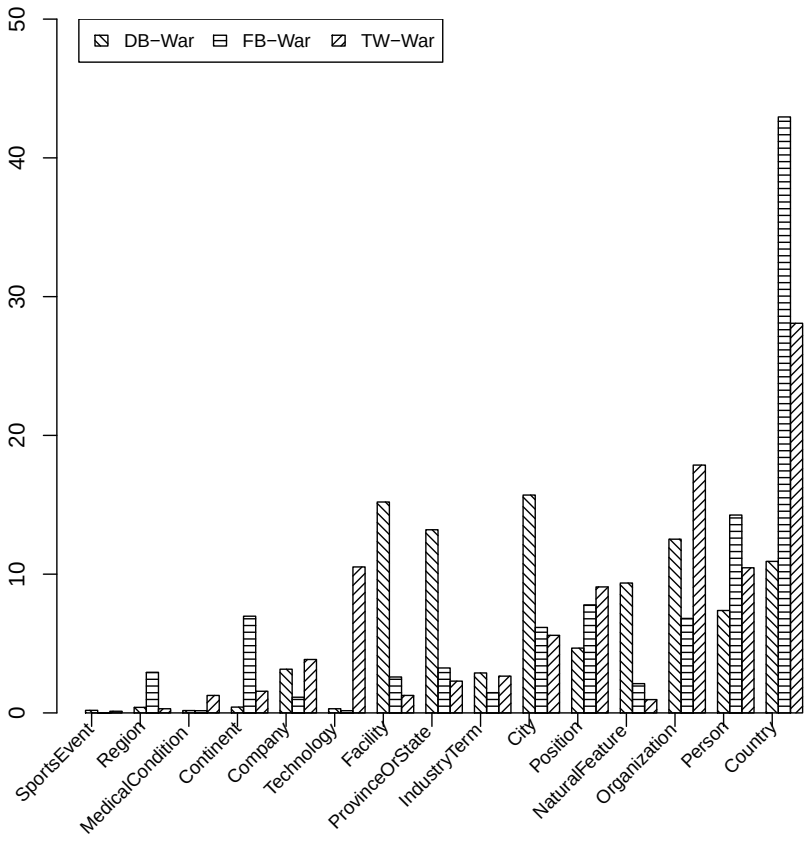

Figure 6.8: The distribution of top 15 entity types in the three gold standard datasets: DBpedia (DB), Freebase (FB) and Twitter (TW) datasets for the War (*War*) topic.

Topic	Example Hashtags
DisAcc	#haiti #pakistan #israel #oilspill #travel
Cri	#imigrants #illegal #jobless #topnews #cnn
War	#afganistan #iraq #alaska #nato #security

Table 6.7: Some example hashtags appearing in the analysed Twitter datasets.

it can observed that in the TW dataset the frequency of hashtags (HSH) and URLs is relatively low, indicating that only a small number of tweets contain external data source specific information. In total 2,386 (23,41%) microposts contain at least one hashtag; and

3,348 (32.85%) microposts contain at least one URLs. The number of unique hashtags is 1,784; while the number of unique URLs is 1,902. Some example hashtags mentioned in the Twitter datasets are illustrated in Table 6.7.

6.6 Evaluation

This section presents a series of experiments to evaluate the proposed *adaptive topic classification framework* and *topical adaptability measures* using the different semantic pivot features introduced in Subsection 6.3.5.

Before describing the experiment in details, however, first the baseline methods used in the experiments are introduced in Subsection 6.6.1, then the evaluation measures presented in Subsection 6.6.2. This is followed by the description of the experimental setup in Subsection 6.6.3, and a discussion on the results in Subsection 6.6.4.

6.6.1 Baseline Methods

6.6.1.1 Baseline Methods for Topic Classification

The proposed text classification framework has been evaluated using different semantic features, augmentation strategies against several baseline models corresponding to state-of-the-art approaches for TC. These baseline models consist of the following features:

Bag-Of-Unigrams (BoW) Features: The unigram features captures the natural intuition to utilise what it is known about a particular topic, so that the features, which are most indicative of a topic, can be detected and the appropriate label(s) assigned. The BoW features consist of a collection of words weighted by TF-IDF (term frequency-inverse document frequency) capturing the relative importance of a word in a document to its use on the whole corpus.

Bag-Of-Entities (BoE) Features: This feature set extends the lexical BoW features with entities and concepts extracted using available annotation services, e.g. OpenCalais API, weighted by TF-IDF. These web services annotate each entity with generic types. For example in the case of *Obama*, rather than recognise it as being of type *dbOwl*:President the majority of these services will annotate this entity with the label *Person* [Rizzo and Troncy, 2011]. In this case the value of the BoE features thus captures the co-occurrence of the entity and concept pairs $f_{BoE}(BarackObama \wedge Person)$.

Bag-Of-Concepts (BoC) Features: This feature set extends the lexical BoW features with concepts extracted with the OpenCalais API. The API provides one single (often generic) concept type for each entity. For example assuming that Barack Obama is annotated as *Person* by OpenCalais, this feature set captures the presence of the Person class type $f_{BoC}(Person)$[29]. This new baseline feature set provides an alternative comparison between the newly proposed semantic meta-graph derived features (*Cls*) and those obtained from the OpenCalais service.

[29]This comparison also allows us to investigate whether modelling each entity with more than one KS concept (in particular 5) is more suitable for TC than with a single one.

Part-of-Speech (POS) Features: Similar to the *BoE* feature set, this feature set aims to capture some generalisation patterns for the words. For this reason, the syntactical patterns within the documents are considered and used to extend the lexical *BoW* features. In this work the Ritter et al.'s Twitter NLP Tools [Ritter et al., 2011] is used, whose PoS tagger has been trained on short text messages.

Considering the above baseline features, two typical baseline supervised machine learning models are employed:

- *TW* (also called TGT_ONLY) *single-domain* topic classifier, in which an SVM topic classifier is built on microposts only (TW), and
- *KS* (also called SRC_ONLY) *cross-domain* topic classifier, in which an SVM topic classifier is built on sole KS (DBpedia and/or Freebase) data.

6.6.1.2 Baseline Content-based Measures of Topic Adaptability

The proposed topic adaptability measures (presented in Section 6.4) are also compared to various content-based domain similarity measures, which make use of the *lexical representation* of the domains.

Formally describing, let $\overrightarrow{d}$ represent a vector consisting of the BoW features weighted with TF-IDF occurring in each domain. Then for the corresponding two domains, $\overrightarrow{d_s}$ denotes the vector for the source domain and $\overrightarrow{d_t}$ denotes the vector for the target domain. Based on this lexical representation, the baseline content-based statistical measures employed are the χ^2 test, the symmetric Kullback-Leibler symmetric distance (KL) and the cosine similarity measures (*cosine*). The reminder of the reader these measures are described in details in Section 5.4.

6.6.2 Evaluation Measures

The evaluation metrics used to compare the performance of the different topic classifiers were the standard *precision*, *recall*, and *F1-measures*. The *precision* (Prec) is computed as the ratio of the number of correctly annotated microposts to the total annotated:

$$\text{Prec} = \frac{|\text{correctly annotated Tweets}|}{|\text{annotated Tweets}|}$$

The *recall* (Rec) is the ratio of the number of correctly annotated microposts to the total number that should have been annotated:

$$\text{Rec} = \frac{|\text{correctly annotated Tweets}|}{|\text{Tweets which should have been annotated}|}$$

The *F1-measure* provides a weighted combination of the two measures, defined as

$$\text{F1} = 2 \times \frac{\text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}$$

6.6.3 Experimental Set-up

Two different topic classification scenarios were analysed and compared in the experiments: a *single-domain* topic classification case, in which case the baseline SVM topic classifier trained on microposts only (TW) is employed (TGT_ONLY), and a *cross-domain* topic classification scenario, in which case an SVM topic classifier is trained on either the KS data alone (which is referred to KS or SRC_ONLY) or combined with Twitter data (which is referred to $KS + TW$ or SRC_TGT).

For evaluating these classifiers, the commonly used one-vs-all approach was employed [Bishop, 2007], in which the multi-label problem was decomposed into multiple independent binary classification problems. Following this approach, each TC system was evaluated using 5-fold cross-validation. The training dataset for the TW topic classification system consisted of 80% of the original Twitter data. For the KS classifier the training set consisted of the full KS data. For the $KS+TW$ classifier the full KS data was combined with 80% of Twitter data.

Each of these topic classifiers is evaluated on 20% of Twitter data, using 5-fold cross-validation. Furthermore, considering that the distribution of Twitter specific indicators is very sparse (as shown in Table 6.6), two different cases have been considered for the creation of Twitter test dataset: the *Full* (default) setting, in which case the complete Twitter data has been used (10,189 microposts), and a *Filt* setting (used mostly in the experiments evaluating the impact of Twitter specific indicators), in which only those microposts were considered which have at least one HSH or URLs (in total 4,778 microposts).

Considering the two different topic classification scenarios, a series of experiments were conducted. In the first set of experiments, the usefulness of the KS data is first evaluated, by comparing the performance of KS topic classifiers built on the individual KS (DB, FB) and joint KS data ($DB+FB$) against the performance of the TW topic classifier trained on microposts only. The main research questions addressed are *"Do KSs contain useful labelled data for building adaptive topic classifiers of microposts?" "Which KS data provides more useful information for topic classification?"*

In the second set of experiments the usefulness of semantic pivot features is evaluated, using different semantic meta-graphs and weighting strategies for the features.

In this case, the main research questions addressed are *"Do semantic meta-graphs built from KSs contain useful semantic features about entities for the topic classification of microposts?" "Which semantic meta-graph provide more useful semantic features for topic classification?"*

The next set of experiments then investigates the impact of Twitter specific indicators, answering the question of *"Does information derived from external data sources' indicators play an important role in topic classification of microposts?"*

Finally, the fourth set of experiments looks at the roles of the semantic features in predicting the adaptability of a topic classifier. For this reason various entropy-based measure were computed and correlated with the performance of SVM topic classifier. In this case the questions under investigation are *"Is it possible to predict the adaptability of a topic classifier?" "Which semantic feature can better represent the adaptability of a topic classifier?"*

6.6.4 Results and Discussion

6.6.4.1 The Usefulness of Knowledge Source Data in Cross-Domain Topic Classification

The goal of this first set of experiments is to investigate the usefulness and the relevance of the KS data alone for topic classification of microposts.

For this purpose the performance of three KS topic classifiers built on KS data alone is compared against the TW topic classifier built on microposts only, according to the two baseline feature sets: BoW and BoE feature sets.

The results obtained over the full performance curve (considering up to 80% of microposts) are presented in Figure 6.9[30]. As it can be observed, the TW topic classifier requires a sufficient amount of annotated microposts in order to significantly outperform the three KS classifiers. In the case of *DisAcc*, at least 993 annotated microposts were required for the TW classifier to significantly outperform $DB + FB$ ($p < 0.01$). For the *Cri* and *War* this number is 640 (see Table 6.8) ($p < 0.05$).

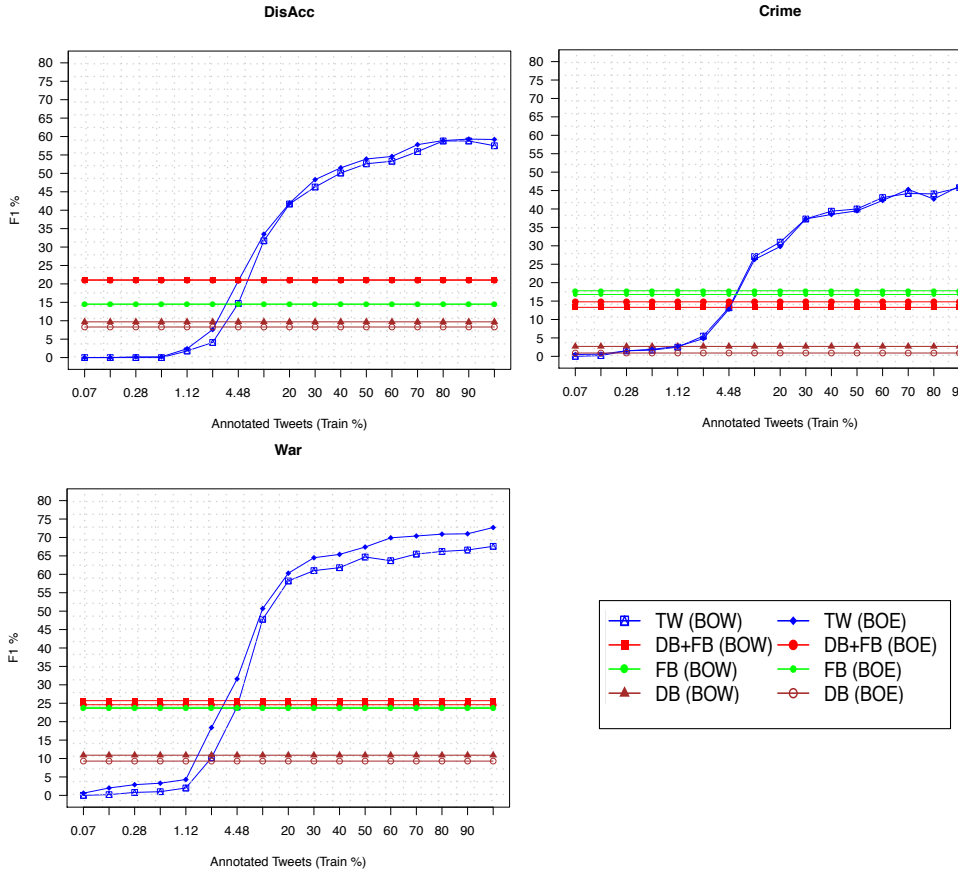

Figure 6.9: The performance in terms of F1-measure of the single-domain TW classifier and cross-domain DB, FB and $DB + FB$ classifiers over the full learning curve, using lexical features.

The comparison of the performance of the $DB + FB$ classifier against DB and FB has furthermore revealed that in the majority of the cases, the $DB + FB$ classifier outperforms

[30]The precision and recall curves are presented in Appendix C.

the individual DB and FB classifiers, indicating that the KSs complement each other.

DisAcc	*Cri*	*War*
993**	640*	640**

Table 6.8: Number of annotated tweets required for the Twitter classifier to beat the DB, FB and $DB + FB$ cross-domain classifiers. Significance levels: p-value $< 0.01^{**}$ 0.05^{*}.

Considering these results, the following general conclusions can be drawn:

1. In the absence of any or large amount of annotated microposts, the application of KS topic classifiers is beneficial.

2. The DBpedia and Freebase KSs contain complementary information about a specific topic, resulting in that the joint $DB + FB$ model significantly outperforms both DB and FB topic classifiers in the majority of the cases.

6.6.4.2 The Usefulness of Semantic Meta-Graphs in Cross-Domain Topic Classification

This section evaluates the usefulness of the semantic meta-graphs derived from KS first for the single-domain scenario, then for the cross-domain scenario.

Single-domain Scenario This section details the results obtained for the single-domain TC case. In these experiments three different single-domain TW classifiers were employed. These classifiers make use of single KS ontologies: TW(*dbKS*) and TW(*fbKS*); and the combined KS ontologies: TW(*dbKS*+*fbKS*). In particular, in the case of the *resource meta-graph*, *dbKS* denotes the *dbOwl* +*yago* ontologies, while in the case of the *category meta-graph*, *dbKS* stands for the *dbCat* ontology. These classifiers are evaluated against several baseline models, as presented in Table 6.10.

Looking at the performance of the baseline models, it can be observed that the best performance was achieved by the *BoE* features, which performed better than the *BoC* and *BoW* features. Further, the *POS* features did not improve on the baseline model using only *BoW* features. An explanation for this could be that the language in Tweets is quite complex, and exhibits less regularity than longer texts used from KSs (KS abstracts).

Comparing the results obtained for the best baseline feature -*BoE* feature- with those for the semantic features derived from the two semantic meta-graphs, it can be observed that the best results were obtained for the *resource meta-graph* for the combined *TW*(*dbOwl*+*yago* +*fbOnt*) scenario using the *P* features with the *W-SG* weighting strategy, which significantly outperforms the baseline lexical features (t-test with $p < 0.05$). In the case of the *War* category, the F1 measure increases by 2.8% with respect to the *BoW* features and 2.2% with respect to the *BoE* features; in the case of the *Cri* category the F1 measure increases by 2.3% with respect to the *BoW* feature and 0.6% with respect to the *BoE* features, while in the case of *DisAcc* an improvement of 1.5% over the *BoW* features can be observed. Further, for both semantic meta-graphs, the novel class–property co-occurrence weighting schema (*W-SG*) for the properties (*P(W-SG)*) shows a significant improvement over the feature frequency strategy (*P(W-Freq)*) (t-test with $p < 0.01$). These results demonstrate that capturing the importance of the property within a given semantic meta-graph (with

Dataset	Semantic graph	Features	TW(*dbKS+fbKS*)	TW(*dbKS*	TW(*fbKS*)
			*F*1	*F*1	*F*1
War	Baseline	BOW	0.800	0.800	0.800
		POS	0.798	0.798	0.798
		BOE	**0.806**	**0.806**	**0.806**
		BOC	0.784	0.784	0.784
	Resource	Cls(W-Freq)	0.790	0.796	0.803
		parent(Cls)(W-Freq)	0.792	0.791	0.803
		P(W-Freq/Cls)	0.803	0.800	0.800
		Cls+P(W-SG)	0.803	0.808	0.804
		parent(Cls)+P(W-SG)	0.803	0.802	0.809
		P(W-SG/Cls)	**0.828**	**0.817**	**0.816**
	Category	Cat(W-Freq)	0.818	0.820	0.817
		parent(Cat)(W-Freq)	**0.824**	**0.822**	**0.820**
		P(W-Freq/Cat)	0.811	0.811	0.809
		Cat+P(W-SG)	0.811	0.811	NA
		parent(Cat)+P(W-SG)	0.816	0.816	NA
		P(W-SG/Cat)	0.823	0.818	0.821
Cri	Baseline	BOW	0.602	0.602	0.602
		POS	0.597	0.597	0.597
		BOE	**0.619**	**0.619**	**0.619**
		BOC	0.590	0.590	0.590
	Resource	Cls(W-Freq)	0.597	0.599	0.605
		parent(Cls)(W-Freq)	0.604	0.603	0.607
		P(W-Freq/Cls)	0.604	0.604	0.606
		Cls+P(W-SG)	0.601	0.599	0.604
		parent(Cls)+P(W-SG)	0.604	0.601	0.607
		P(W-SG/Cls)	**0.625**	**0.612**	**0.616**
	Category	Cat(W-Freq)	0.613	0.613	0.609
		parent(Cat)(W-Freq)	0.613	0.613	0.605
		P(W-Freq/Cat)	0.610	0.610	0.608
		Cat+P(W-SG)	0.610	0.610	NA
		parent(Cat)+P(W-SG)	**0.616**	**0.616**	NA
		P(W-SG/Cat)	0.613	0.606	**0.614**
DisAcc	Baseline	BOW	0.709	0.709	0.709
		POS	0.696	0.696	0.696
		BOE	**0.728**	**0.728**	**0.728**
		BOC	0.680	0.680	0.713
	Resource	Cls(W-Freq)	0.705	0.707	0.703
		parent(Cls)(W-Freq)	0.705	0.706	0.706
		P(W-Freq/Cls)	0.690	0.706	0.703
		Cls+P(W-SG)	0.704	0.710	0.708
		parent(Cls)+P(W-SG)	0.708	0.709	0.704
		P(W-SG/Cls)	**0.724**	**0.718**	**0.715**
	Category	Cat(W-Freq)	0.712	0.714	0.710
		parent(Cat)(W-Freq)	0.716	0.716	**0.716**
		P(W-Freq/Cat)	0.715	0.715	0.711
		Cat+P(W-SG)	0.715	0.715	NA
		parent(Cat)+P(W-SG)	**0.719**	**0.719**	NA
		P(W-SG/Cat)	0.715	0.718	0.709

Table 6.10: The performance of the single-domain *TW* SVM topic classifiers using different KSs ontologies (DBpedia *dbKS*'s ontologies, and Freebase *fbKS*'s ontology) and two semantic meta-graphs derived from these KSs (*resource meta-graph* (Resource) and *category meta-graph* (Category)). The best results for the baseline, *resource meta-graph* and *category meta-graph* features for each topic and classifier are shown in bold.

respect to concepts in the *resource meta-graph* or to categories in the *category meta-graph*), improves the generality of the properties and the performance of the topic classifier for each topic.

While employing the P features has been shown to provide a positive gain over the baseline features for most of the topics, the usefulness of the semantic features and augmentation strategies merely depend on a number of factors. For instance, one of the factors which influences the performance of a topic classifier is the number of entities identified in a micropost. For instance, in the case of the *War* topic, a higher number of entities have been extracted than for the other two topics. This can explain the higher gain achieved for this topic,

resulted from a larger number of microposts being enriched. Further, the lower performance achieved by the *Cls* features, could be due to the level of ambiguity (measured as cls/ent) of the *Cls* features and their discriminative power for a given topic. Looking at the Table 6.6, it can be observed that there are a larger number of property features defined in KSs for an entity ($prop/ent$) than for a class (cls/ent, $fbcls/ent$). This allows the incorporation of very fine-grained information into TC, which indeed seems to improve the performance of the classifier upon the baseline features. In order to capture these factors and provide an insight into the usefulness of these features for topic classification, the reminder of the reader, a set of topic similarity measures are employed which will be evaluated in Subsection 6.6.4.4.

Inspecting the results obtained for the different taxonomies, similar trends were observed for the *resource meta-graph* and *category meta-graph*. That is, for both semantic graphs the *dbKS* ontologies (*dbOwl+yago* for *resource meta-graph*; and *dbCat* for *category meta-graph*) provide a significant improvement over the semantic features derived from *fbKS* ontology for the *War* and *DisAcc* topics, except for *Cri* (t-test with $p < 0.05$). This could be explained by the fact that in the *Cri* topic the entities extracted from the *dbKS* graph are more ambiguous than those found within the *War* and *DisAcc* topics (see cls/ent values in Table 6.6). Similarly, the entities extracted from the *fbKS* are less ambiguous in the *Cri* topic than in the other two topics (see $fbcls/ent$ values in Table 6.6). The best overall results were obtained by the combined *dbOwl+yago+fbOnt* and *dbCat+fbOnt* ontologies using the property features, indicating that the three ontologies contain complementary information (properties) about the entities.

Further, it was found that the augmentation strategies are beneficial for both semantic graphs. In the case of the *resource meta-graph*, different trends were found for the *fbOnt* and *dbOwl + yago* ontologies. When using *fbOnt* ontology, both (*parent(Cls)(W-Freq)* and *parent(Cls) + P(W-SG)*) showed a consistent improvement over the initial non-generalisation case (*Cls(W-Freq)* and *Cls+ P(W-SG)*) for each topic. However, when using the *dbOwl + yago* ontology encoding the very specific classes of the entities were found to be more beneficial for some topics (e.g. *War*). These results are understandable because after generalisation, the entities which have the same parent class in the KS graphs will be unified to the same semantic concept type, losing as a result the very specific meaning of the entity. In the case of *yago* ontology, the number of unique classes reduces by 92% after generalisation, while in *fbOnt*, the number of unique classes becomes 88% less. In the case of the *category meta-graph*, further, it was found that the *parent(Cat)(W-Freq)* and *parent(Cat) + P(W-SG)* features significantly improved over the *Cat(W-Freq)* and *Cat+ P(W-SG)* features for each topic (t-test with $p < 0.05$).

Cross-domain Scenario In this section the performance of the cross-domain topic classifiers is examined using the different semantic concept graphs and compared with the performance of the single-domain topic classifier.

Based on the three scenarios analysed, in these experiments six different cross-domain topic classifiers were employed. Among these cross-domain classifiers, four make use of individual KS ontologies: DB making use of *dbKS*'s ontologies, FB making use of *fbKS*'s ontology, DB+TW exploiting *dbKS*'s ontologies, FB+TW employing *fbKS*'s ontology. The remaining two cross-domain topic classifiers make use of the combined KS ontologies: DB+FB

and DB+FB+TW. In particular, in the case of the *resource meta-graph*, *dbKS* denotes the *dbOwl +yago* ontologies, while in the case of the *category meta-graph*, *dbKS* stands for the *dbCat* ontology. These classifiers are evaluated against several baseline models, as presented in Table 6.9.

Looking at the performance of the baseline models, a different trend can be observed compared to the TW only scenario. The syntactic classes provided by the POS taggers, in this cross-domain scenario, were found to be more beneficial, compared to the BoW cases. While for the BoE and BoC features, no improvement was observed upon the baseline BoW features. An explanation for this could be that the entities which appear in the TW dataset could be quite different from the entities appearing in the KS data for each topic, in which case exploiting the semantic information from KSs seems to be more beneficial.

Inspecting the best overall performance for the various features, feature weighting strategies and augmentation strategies, it was noticed that the *resource meta-graph* achieved the best results using the $DB(dbOwl + yago) + FB(fbOnt) + TW$ topic classifier.This classifier significantly outperformed the baseline single KS classifiers: by 11.9-30.7% (over $DB+TW$) and 13.4-31.4% (over $FB + TW$) (t-test with $p < 0.05$). Considering the *category meta-graph*, the improvements were slightly smaller, a significant improvement of 11.5-30.2% was observed over $DB + TW$ and 13-30.9% over $FB + TW$ (t-test with $p < 0.05$). Comparing the results against the TW baseline models, a significant improvement of 9.3%-28.2% can be observed over the $TW(dbOwl + yago + fbOnt)$ when using the *resource meta-graph*, and 8.9%-27.7% over the $TW(dbCat + fbOnt)$ classifiers when using the *category meta-graph*. Furthermore, the superiority of the TW topic classifier over the DB, FB and $DB+FB$ topic classifiers are in light with results obtained in Subsubsection 6.6.4.1, which demonstrated that outperforming the TW topic classifiers is extremely difficult using KS data alone.

Comparing the different enrichment strategies, similar trends can be observed for both *resource meta-graph* and *category meta-graph*. The best enrichment that consistently improved over the baseline for both concept graphs was the *W-SG* for *P*, indicating that encoding the specificity of a property for each semantic concept graph is beneficial for TC. For the *W-Freq* features, however, it was found that in the case of the *resource meta-graph*, the semantic augmentation by feature frequency (*Cls(W-Freq)*) and by generalisation (*parent(Cls)(W-Freq)*) worked consistently better than the baseline models. However, in the case of the *category meta-graph*, the performance of the *Cat(W-Freq)* and *parent(Cat)(W-Freq)* were only comparable to those of the baseline models.

Despite of the accuracy gain obtained with the *P* and *Cls* features for the $DB+FB+TW$ classifier, an interesting observation about these results is however, that the semantic features do not always improve upon the baseline models. For instance in the case of $DB+FB$ topic classifier, the results are comparable or slightly worst than those obtained by the BoW feature set ignoring semantic augmentation. An explanation for this could be that the distribution of entities in the DB and FB datasets may slightly be different to the one in Twitter. Further given that these classifiers do not make use of any microposts data, this mismatch provides challenges for the topic classifier. A possible reason for this could be the level of ambiguity of the entities in the different datasets. In order to capture the differences between the datasets and provide an estimation on the usefulness of the different semantic features, the reminder of the reader, a set of topic similarity measures were employed which will be examined in Subsection 6.6.4.4.

Contrasting the results for all three topics, it can be observed, that the biggest overall improvement was achieved for the *Cri* topic using the *resource meta-graph*. In particular, the $DB+FB+TW$ achieved an improvement of 31.4% over $FB+TW$. For the case of the *category meta-graph*, the $DB+FB+TW$ achieved an improvement of 30.9% over $FB+TW$.

Also for the *Cri* topic, it was observed, that the $FB+TW$ single KS classifier using BoW features performed better than the $DB+TW$ single KS classifier. However, when looking at the results obtained for the BoE features, the opposite trend was observed, the $DB+TW$ performed better than the $FB+TW$. An explanation for this could be that a relatively large number (3,377) of articles do not contain any entity, and thus are not semantically enriched.

Further, it was noticed that the coverage of entities is lower in the Freebase than in DBpedia. For example from the total number of entities extracted by OpenCalais a large proportion (40%) of the entities were not found in the Freebase KS, while in the case of DBpedia 35% of the entities were not assigned any URI. Regardless of this, an improvement in F1 measure was obtained for both semantic graphs when combining the two linked KSs. This thus indicates that the two linked KSs complement each other well. In one hand, Freebase brings its strength in content coverage for the topics, while DBpedia brings useful semantic evidence about the entities which are covered.

In conclusion, considering the results obtained for both single-domain and cross-domain scenarios for the various semantic features derived from the three KS graphs the main findings are as follows:

1. Semantic meta-graphs (both *resource meta-graph* and *category meta-graph*) built from KSs contain useful semantic features about entities for topic classification. In particular, incorporating semantic features about properties (**P**) using the novel class-property co-occurrence weighting schema (**W-SG**) proved a significant improvement over previous state-of-the-art approaches.

2. Combining the evidence about the semantic features from multiple linked KS taxonomies ($TW(db+yago+fb)$) is beneficial for TC, showing a significant improvement over approaches considering a single KS ($TW(db+yago)$, $TW(fb)$).

The Role of Semantic Concept Graphs in Single-domain and Cross-domain Topic Classification

In the previous section the overall performance of a topic classifier was compared using semantic features derived from two semantic meta-graphs (*resource meta-graph* and *category meta-graph*). In this section, the discussion focuses on the differences in roles of these semantic features in different TC scenarios.

Looking at the results obtained for the individual semantic features (**Cat**, **Cls**, **P**) different patters can be observed for the single-domain and cross-domain topic classification scenarios.

Inspecting the results obtained for the single-domain topic classifier, it can be noticed, that the performance of SVM topic classifier was consistently higher using the **Cat** features than using the **Cls** features for both **Freq** and **SG** weightings (see Table 6.10) (t-test with $p < 0.05$). These results indicate that the information about the category features seems to be more beneficial than the information about the classes in the single-domain TC scenario.

However, for the **P** features, it was found that the weights obtained from the *resource meta-graph* are better than those obtained from the *category meta-graph*. This behaviour could be understood by the fact that the *category meta-graph* consists of a larger number of **Cat** than the number of **Cls** in the *resource meta-graph*, and in addition the **Cat** are more ambiguous (less focused) than the **Cls** in terms of the number of properties associated to them.

In contrast to these observations, in the cross-domain topic classification scenarios different trends were found for the **Cat** and **C** features (t-test with $p < 0.05$). While for the TW only scenario, the **Cat** features worked better than the **Cls** features, in the cross-domain scenario the opposite trend was observed: the **Cls** features are more useful than **Cat** features. An explanation for this could be that, that the different datasets contain a larger number of **Cat** features than the **C** features (compare **dbCat** with **dbClass**, **yagoClass** and **fbClass** in Table 6.6), making it harder for the cross-domain classifiers to generalise over the **Cat** features, than for the **C** features.

In conclusion, considering the results obtained for both single-domain and cross-domain scenarios the main findings are as follows:

1. The semantic features derived from the two *resource meta-graph* and *category meta-graph* exhibit different roles (generalisation patters) in the different TC scenarios. The class features derived from the *resource meta-graph* exhibit better *generalisation patterns* in the cross-domain setting, while the category features derived from the *category meta-graph* are better suitable to encode the *specificity* of a topic in a single-domain setting

2. Despite of the differences in roles of the semantic features derived from the two semantic meta-graphs, incorporating semantic features from both semantic graphs is beneficial for TC, achieving performance superior to previous approaches utilising lexical features

6.6.4.3 The Usefulness of Twitter Specific Indicator Features in Topic Classification

This section continues by discussing the results obtained by incorporating the Twitter specific indicator features into a topic classifier. First the results obtained for the single-domain topic classification are presented. This is followed by the results obtained for the cross-domain topic classification case.

Single-domain Topic Classification Table 6.11 summarises the results obtained for the single-domain topic classification case using the indicator features alone, and combined with the previously presented semantic entity features. Considering the results obtained on the the *Full* TW corpus using the indicator features alone, it can be observed that the classifier built using *BoL* and *BoH* features improve upon the baseline classifier considering words only (BoW), except for the *DisAcc* topic using $BoL(T)$ feature. The best overall results were obtained by the BoH (Prop), achieving an improvement of 1.6% over the baseline for the *DisAcc*, 2.6% for the *Cri*, and 2.1% for the *War* topic. These results are in agreement with the results obtained in the single-domain topic classification case for the semantic entity features (in Subsubsection 6.6.4.1), which also showed that the property features provide useful information for topic classification, and also incorporating them into a topic classifier is more beneficial than utilising concept features.

Dataset	Semantic graph	Features	DB+FB	DB+FB+TW	DB	DB+TW	FB	FB+TW
			*F*1	*F*1	*F*1	*F*1	*F*1	*F*1
War	Baseline	BOW	**0.022**	0.905	**0.080**	0.793	0.226	**0.781**
		POS	0.013	**0.914**	0.061	0.797	**0.237**	0.776
		BOE	0.014	0.799	0.080	**0.802**	0.207	0.776
		BOC	0.020	0.700	0.091	0.761	0.199	0.752
	Resource	Cls(W-Freq)	0.017	0.916	0.075	0.792	0.226	0.765
		parent(Cls)(W-Freq)	**0.022**	0.917	0.077	0.795	0.226	0.776
		P(W-Freq/Cls)	0.017	0.911	0.086	0.789	**0.227**	0.773
		Cls+P(W-SG)	0.021	0.908	0.071	0.795	0.226	0.779
		parent(Cls)+P(W-SG)	0.021	0.907	0.069	0.787	**0.227**	0.774
		P(W-SG)	0.013	**0.921**	**0.103**	**0.796**	0.226	**0.787**
	Category	Cat(W-Freq)	0.021	0.914	0.080	**0.806**	**0.243**	0.793
		parent(Cat)(W-Freq)	0.024	0.915	**0.107**	0.799	0.242	0.789
		P(W-Freq/Cat)	0.023	0.910	0.088	0.793	0.242	0.781
		Cat+P(W-SG)	**0.026**	0.911	0.087	0.794	NA	NA
		parent(Cat)+P(W-SG)	0.021	0.913	0.090	0.797	NA	NA
		P(W-SG/Cat)	0.026	**0.917**	0.094	0.804	0.242	**0.794**
Cri	Baseline	BOW	0.025	0.898	**0.011**	0.573	0.240	0.583
		POS	0.025	**0.902**	0.009	0.592	**0.247**	0.582
		BOE	**0.052**	0.708	0.008	**0.600**	0.186	**0.593**
		BOC	0.021	0.531	0.022	0.553	0.250	0.559
	Resource	Cls(W-Freq)	0.021	**0.907**	0.016	0.564	0.233	0.581
		parent(Cls)(W-Freq)	0.019	**0.907**	0.014	**0.569**	0.237	0.585
		P(W-Freq/Cls)	0.021	0.904	**0.019**	0.566	**0.241**	0.580
		Cls+P(W-SG)	0.026	0.899	0.011	0.559	0.235	0.586
		parent(Cls)+P(W-SG)	0.024	0.899	0.011	0.563	0.240	0.582
		P(W-SG/Cls)	**0.028**	0.903	0.016	0.560	0.232	**0.593**
	Category	Cat(W-Freq)	0.021	**0.902**	0.013	0.575	0.258	0.586
		parent(Cat)(W-Freq)	0.021	0.901	**0.017**	0.588	0.244	0.588
		P(W-Freq/Cat)	0.022	0.899	0.013	0.579	**0.261**	0.582
		Cat+P(W-SG)	0.022	0.900	0.009	0.568	NA	NA
		parent(Cat)+P(W-SG)	0.022	0.899	0.016	0.589	NA	NA
		P(W-SG/Cat)	**0.024**	0.901	0.014	**0.592**	0.250	**0.594**
DisAcc	Baseline	BOW	0.004	**0.910**	**0.107**	0.684	**0.162**	**0.696**
		POS	0.017	0.903	0.052	0.688	0.159	0.679
		BOE	**0.076**	0.708	0.079	**0.722**	0.092	0.692
		BOC	0.023	0.602	0.155	0.656	0.176	0.627
	Resource	Cls(W-Freq/Cls)	0.004	0.915	**0.125**	0.679	0.162	0.691
		parent(Cls)(W-Freq)	0.004	**0.917**	0.109	**0.689**	**0.162**	0.692
		P(W-Freq/Cls)	0.004	0.910	0.123	0.677	**0.162**	0.681
		Cls+P(W-SG)	0.004	0.907	0.120	0.684	**0.162**	0.689
		parent(Cls)+P(W-SG)	**0.005**	0.908	0.112	**0.689**	**0.162**	0.693
		P(W-SG/Cls)	0.004	0.912	0.109	0.688	**0.162**	**0.695**
	Category	Cat(W-Freq)	0.003	0.911	**0.102**	**0.702**	0.166	**0.705**
		parent(Cat)(W-Freq)	0.004	0.911	0.088	0.694	0.162	0.688
		P(W-Freq/Cat)	0.004	0.907	0.096	0.681	0.162	0.694
		Cat+P(W-SG)	0.004	0.908	0.095	0.689	NA	NA
		parent(Cat)+P(W-SG)	0.005	0.909	0.098	0.687	NA	NA
		P(W-SG/Cat)	**0.005**	**0.913**	0.088	0.686	0.162	0.683

Table 6.9: The performance of the DB, FB and $DB+FB$ cross-domain SVM topic classifiers using different KSs ontologies (DB -using *dbKS*'s ontologies, FB -using *fbKS*'s ontology) and two semantic meta-graphs derived from these KSs (*resource meta-graph* (Resource) and *category meta-graph* (Category)). The best results for the baseline, *resource meta-graph* and *category meta-graph* features for each topic and classifier are shown in bold.

For the case, when the indicator features are combined with the semantic entity features, a slight improvement can be observed against the results obtained over both indicator features alone (Table 6.11), and semantic features alone (Subsubsection 6.6.4.1) (t-test $p < 0.05$). The improvement against the BOW features becomes 2.1% for *DisAcc*, 2.8% for the *Cri*, 3.0% for the *War*. For the property features, overall, it was again found that the *resource meta-graph* contain more useful information than the *category meta-graph* graph.

Inspecting the results on the *Filt* TW corpus, it was found that the $BoH(P)$ features perform the best. The improvement over the baseline classifier, however, was much bigger in two of the cases: 9.5% for the *Cri* topic, and 5.8% for the *War* topic. The improvement for *DisAcc* is 1.4%. An explanation for the small improvement for the *DisAcc* topic can be understood by the fact that the microposts belonging to the *DisAcc* topic contain the less number of HSHs and URLs, and therefore less number of microposts are semantically enriched. As in the case of the *Full* TW corpus, the best overall results were obtained using the *resource meta-graph*. Furthermore, the performance of the TW classifier slightly improvements when combining the indicator features with the semantic entity features.

Case	Semantic graph	Features	*DisAcc* $F1$	*Cri* $F1$	*War* $F1$
Full	Baseline	BOW	0.709	0.602	0.800
		BoL(1)	0.720	0.623	0.802
		BoL(L)	0.720	0.623	0.815
		BoL(T)	0.704	0.614	0.805
	Resource	BoH(Cls)	0.713	0.623	0.808
		BoH(P/Cls)	0.719	***0.628***	***0.821***
		Cls+BoH(Cls)	0.715	0.625	0.814
		P(SG/Cls)+BoH(P/Cls)	***0.730***	***0.630***	***0.830***
	Category	BoH(Cat)	0.712	0.619	0.808
		BoH(P/Cat)	***0.722***	0.627	0.820
		Cat+BoH(Cat)	0.718	0.621	0.818
		P(SG/Cat)+BoH(P/Cat)	0.728	0.628	***0.830***
Filt	Baseline	BOW-Filt	0.635	0.522	0.755
		BoL(1-Filt)	0.623	0.574	0.762
		BoL(L-Filt)	0.623	0.574	0.762
		BoL(T-Filt)	0.617	0.596	0.786
	Resource	BoH(Cls-Filt)	0.636	0.586	0.790
		BoH(P-Filt/Cls)	0.625	***0.617***	***0.813***
		Cls+BoH(Cls-Filt)	0.638	0.586	0.790
		P(SG/Cls)+BoH(P-Filt/Cls)	0.649	***0.618***	***0.817***
	Category	BoH(Cat-Filt)	0.621	0.525	0.756
		BoH(P-Filt/Cat)	***0.649***	0.592	0.758
		Cat+BoH(Cat-Filt)	0.643	0.577	0.791
		P(SG/Cat)+BoH(P-Filt/Cat)	***0.655***	0.610	0.803

Table 6.11: The performance of the single-domain SVM *TW* topic classifier using external *data source indicators*. The best results for the baseline, *resource meta-graph* and *category meta-graph* features for each topic are shown in bold.

Comparing the performance obtained for the different *BoL* features, it was observed, that in the *Filt* case, when most of the tweets have a URL inside them, the Title of the articles was found to be more informative of a topic. However, in the *Full* case, both the first and the last paragraphs of the webpages were found more beneficial than the title of the webpages.

Cross-domain Topic Classification The results obtained for the cross-domain scenario using the $DB + FB + TW$ classifier are presented in Table 6.12. For the *Full* TW case, the best overall results were obtained using the $BoH(P)$ features, similarly to the single-domain scenario. In this case, however, the $BoH(P/Cat)$ features achieved better results than the $BoH(P/Cls)$ features (t-test, $p < 0.05$). Furthermore, combining the indicator features with the semantic entity features, showed no significant improvements over the sole indicator feature case. The highest improvement over the baseline BoW features was 6.0% for the *DisAcc*, 6.3% for the *Cri*, 6.9% for the *War* topic. Compared to the best TW single-domain classifier, further an improvement of 24% was achieved for the *DisAcc*, 33.1% for the *Cri* and 14.4% for the *War* topic.

Comparing the two indicator feature, it was noticed that the results obtained for the $BoH(P)$ features outperformed the results obtained by the URL features. These results indicate that incorporating semantic information derived from KSs is very important in reducing the lexical gap between KSs and TW. In particular, the addition of new words derived from the external URL websites were found worst or achieved little improvement over the baseline BoW case (for *DisAcc*, *Cri*). With respect to the URL features, it was noticed that the performance of the classifier does not change drastically when utilising the first, last or the title of external URL websites. The difference in the performances is less than 1%.

Case	Semantic graph	Features	*DisAcc* $F1$	*Cri* $F1$	*War* $F1$
Full	Baseline	BOW	0.910	0.898	0.905
	Resource	BoL(1)	0.908	0.898	0.913
		BoL(L)	0.908	0.900	0.911
		BoL(T)	0.905	0.897	0.911
		BoH(Cls)	0.969	0.960	***0.974***
		BoH(P/Cls)	0.927	0.920	0.929
		Cls+BoH(Cls)	0.969	0.960	***0.974***
		P(SG/Cls)+BoH(P/Cls)	0.928	0.925	0.930
	Category	BoH(Cat)	0.967	0.960	0.973
		BoH(P/Cat)	***0.970***	***0.961***	***0.974***
		Cat+BoH(Cat)	0.969	0.960	0.973
		P(SG/Cat)+BoH(P/Cat)	**0.970**	***0.961***	***0.974***
Filt	Baseline	BOW-Filt	0.550	0.500	0.885
	Resource	BoL(1-Filt)	0.716	0.862	0.887
		BoL(L-Filt)	0.895	0.856	0.886
		BoL(T-Filt)	0.892	0.863	0.886
		BoH(Cls-Filt)	0.969	***0.945***	***0.973***
		BoH(P-Filt)	0.914	0.879	0.912
		BoH(Cat-Filt)	0.969	0.941	0.971
		BoH(P-Filt/Cat)	***0.970***	0.941	0.972
	Category	Cls+BoH(Cls-Filt)	***0.970***	***0.953***	***0.973***
		P(SG/Cls)+BoH(P-Filt/Cls)	0.883	0.858	0.881
		Cat+BoH(Cat-Filt)	0.969	0.938	0.969
		P(SG/Cat)+BoH(P-Filt/Cat)	0.962	0.941	0.962

Table 6.12: The performance of the DB+FB+TW cross-domain SVM topic classifier using various external *datasource indicators*. The best results for the baseline, *resource meta-graph* and *category meta-graph* features for each topic are shown in bold.

Examining the results obtained for the *Filt* case, a different trend can be observed, the $BoH(Cls)$ features achieves the best results in the majority of the cases (for the *Cri* and *War* topics). While for the *DisAcc* the $BoH(Cat)$ and $BoH(Cls)$ achieved comparable results. Furthermore, similarly to the *Full* case, there was no big improvement when combining the indicator features with the semantic entity features, except for the *Cri* topic. Considering the URL features, however, it can be noticed that the title of the websites seems to be more beneficial for TC, than the first or the last paragraphs. An explanation for this could be, that in this *Filt* scenario more microposts are affected by feature augmentation than in the *Full* scenario. In light with the results for the single-domain scenario, bigger improvement can also be observed over the baseline (up to 42% for *DisAcc*, 45.3% for *Cri*) in the Filt case than in the Full case.

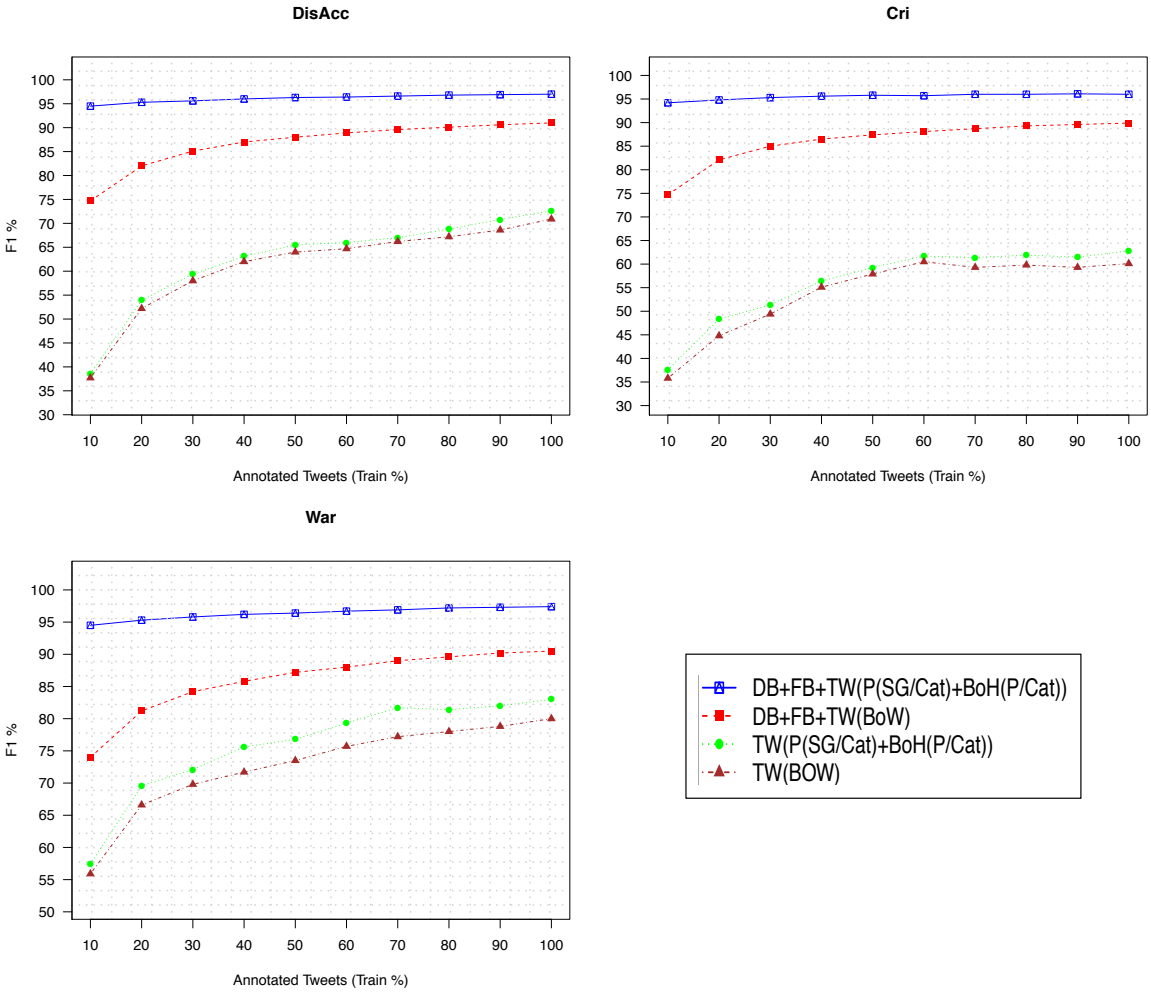

Figure 6.13: Performance curves in terms of F1 measure for the single-domain TW classifier and cross-domain DB+FB+TW classifier using lexical (BoW) and semantic features (**P** features from the *category meta-graph*).

Having the best TW single-domain and $DB + FB + TW$ cross-domain topic classifiers identified, an additional analysis was also conducted examining the performance of these classifiers over the full learning curve, utilising all the target domain data (*Full* case). Figure 6.13 shows the performance of the $DB+FB+TW(P(SG/Cat)+BoH(P/Cat))$, $DB+FB+TW(BoW)$ cross-domain classifiers against the $TW(P(SG/Cat)+BoH(P/Cat))$,

$TW(BoW)$ single-domain classifiers. It can be observed, that using as little as 10% of annotated tweets (815.12), the $DB{+}FB{+}TW(P(SG/Cat){+}BoH(P/Cat))$ classifier consistently and significantly outperforms the other classifiers over the full learning curve, for all the three topics. The baseline $DB + FB + TW(BoW)$ classifier was also found to be very effective, the number of annotated tweets required to outperform the TW single-domain classifiers was 10% for the *DisAcc* and *Cri* topics, and 20% for the *War* topic. These results thus demonstrate that the proposed adaptive DB+FB+TW (P(SG/Cat)+BoH(P/Cat)) classifier can drastically reduce the human effort of annotating tweets, being more effective than the baseline models. .

In conclusion, considering the results obtained for both single-domain and cross-domain scenarios the following findings can be drawn:

1. The Twitter specific indicator features provide useful information for topic classification. Incorporating the $BoH(P)$ features from the *resource meta-graph* into the single-domain TW classifier, and the $BoH(P)$ features from the *category meta-graph* into the cross-domain KS+TW ($DB + FB + TW$) classifier showed significant improvement over baseline models.

2. The combination of the hashtag indicator and semantic entity features was found beneficial for both scenarios. In the single-domain scenario, the combination of these two features achieved superior results to the classifiers using sole hashtag and sole entity features. In the case of cross-domain scenario, a significant improvement can be observed over the classifier using semantic entity features only, and the results are comparable to the classifier using only hashtags features.

6.6.4.4 Evaluating Topic Adaptability

The previous sections analysed the benefit of using semantic features derived from KS graphs for the topic classification task in both single-domain and cross-domain scenarios. These results have shown that there is variation in the performance levels between topics, suggesting that differences between the KS and Twitter datasets affects the performance levels.

In order to understand these variations, this last set of experiments aims to analyse and compare different topic adaptability measures, which can be used to estimate the performance of a topic classifier. First a series of content-based adaptability measures are compared, correlating their value with the performance of the $DB + FB$ and TW topic classifiers. Following this, the newly proposed entropy-based adaptability measures are compared against the best content-based adaptability measure found.

Content-based Adaptability Measures Figure 6.14 shows the correlations obtained using KL, *cosine* and χ^2 values and the performance in F1 of the $DB + FB$ and TW classifiers. A positive correlation indicates that the performance of the topic classifier increases as the divergence decreases (the distribution are more similar); while a negative correlation indicates that the performance increases as the divergence increases (the distributions are less similar).

As it can be observed, for the single-domain case, the Chi-TGT achieved the highest correlation values (in absolute terms) for all the three topics: in particular Chi-TGT(BoE) for the *DisAcc*, Chi-TGT(BoW) for the *Cri*, and Chi-TGT(BoE) for the *War*. A similar

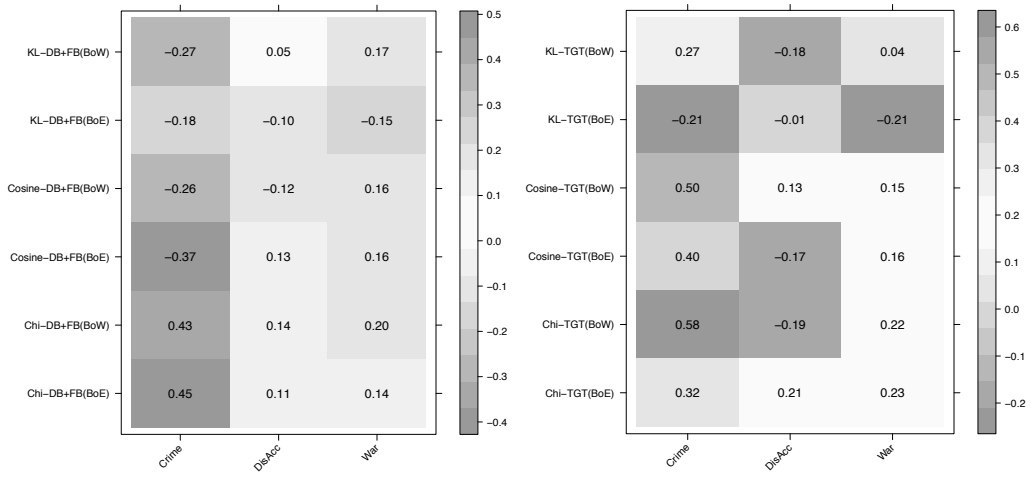

Figure 6.14: Pearson correlation values between the content-based adaptability measures and the performance of the $DB + FB$ cross-domain (left), and $TW(dbKS + fbKS)$ single-domain (right) topic classifiers.

trend can be observed for the cross-domain scenario, in which case, Chi-DB+DB(BoE) obtains the best correlation for the *DisAcc*, while Chi-DB+DB(BoE) for the *Cri* and Chi-DB+DB(BoE) *War* topics.

The second best values were achieved by the *cosine* measure for both single-domain and cross-domain TC scenarios, achieving higher correlation scores than the KL measure for *Cri* and *DisAcc*.

These results thus indicate that the χ^2 provides the best correlation scores for the adaptability of a topic classifier[31].

KS-based adaptability measures For examining the KS based adaptability measures, the *entropy difference* values were computed to capture the difference between the train and test datasets for each topic as introduced in Section 6.4.

In order to assess the relevance of a semantic feature type to the performance of a topic classifier, these metrics were analysed by considering the following cases:

1. Measuring entity dispersion (Entity-Entropy) - Since this metric captures only the entity dispersion in topics, it was correlated against topic classifiers build on BoE features;

2. Measuring class dispersion (Cls-Entropy, EntityCls-Entropy) - In this case the topic classifiers trained using **Cls** features was employed;

3. Measuring category dispersion (Cat-Entropy, EntityCat-Entropy) - In this case the topic classifiers built using the **Cat** features was employed; and

4. Measuring property dispersion (P-Entropy, EntityProp-Entropy, PropertyCls-Entropy, and PropertyCat-Entropy) - the topic classifiers using **P** features was considered.

[31]As χ^2 measure distance, the inverted value of $(\chi^2)^{-1}$ was used to measure the similarity between domains.

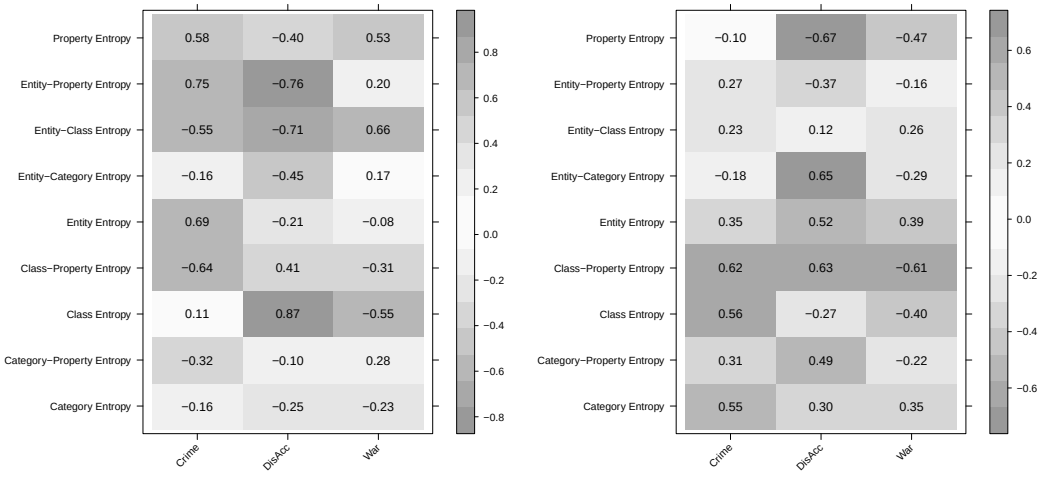

Figure 6.15: Pearson correlation values between entropy difference measures and the performance of the $DB + FB$ cross-domain (left), and $TW(dbKS + fbKS)$ single-domain (right) topic classifiers.

Figure 6.15 presents the Pearson correlation values obtained for each topic between the entity difference scores and the performance of the cross-domain ($DB + FB + TW$) and single-domain ($TW(DB + FB)$) classifiers in terms of F1 measure obtained using 80% of TW data for training (in addition to the KS data), and 20% TW data for test.

These figures show that in the cross-domain ($DB+FB+TW$) scenario, the *EntityProp-Entropy* yields the best correlation scores, over 70% in two out of three topics. When looking at the values obtained for *Cls-Entropy*, *Cat-Entropy*, *P-Entropy* and *Entity-Entropy* measures, it was observed that the *Cls-Entropy* showed the highest correlation values with the performance of the cross-domain topic classifiers. For the *DisAcc* and *War* these values were higher than 54%, however, for the *Cri* topic the correlation values were 11%. When examining the class dispersion measures, it can furthermore be noticed, that the *Entity-Cls* Entropy showed higher correlation than *Cls-Entropy*. In the case of the category dispersion values, for some topics (e.g. *DisAcc*) the *EntityCat-Entropy* was found better, while for others (e.g. *War*) the *Cat-Entropy* was more beneficial. Moreover, among the property dispersion values the *EntityProp-Entropy* values showed the highest correlation values.

Considering the results obtained for the single-domain TC ($TW(DB+FB)$) case, the *Cls-Prop Entropy* yields the best correlation value, over 60% for all the three topics. Among the *Cls-Entropy*, *Cat-Entropy*, *P-Entropy* and *Ent-Entropy* measures, however, the *P-Entropy* values were found the best. As opposed to the cross-domain case, among the class dispersion measures the *Cat-Entropy* values were higher than the *Entity-Cls Entropy* values. For the category dispersion measures, furthermore the *Cat-Entropy* values were higher than the *EntityCat-Entropy* values in two out of three topics.

These results indicate that in the single-domain case analysing a single-semantic feature (e.g. **P**, **Cls** or **Cat**) can provide a good estimate of the performance of the TC, while in the cross-domain scenario the representation of the topics seems to be more complex, requiring to model the entropy of two semantic features (in form of conditional entropy values). Nonetheless, among the property dispersion values, the best results were obtained

by the *ClsProp-Entropy* values.

A comparison was also made with the results obtained in the previous section using the χ^2 measure for both BOW and BOE features. According to these results, in the single-domain case the best correlation values obtained were: 21% (BOE) for *DisAcc*, 58% (BOW) for *Cri*, and 23% (BOE) for *War*; while in the cross-domain case, these values were 14% (BOW) for *DisAcc*, 45% (BOE) for *Cri*, and 20% (BOE) for *War*. As it can be observed, the novel entropy based-adaptability measures (*EntityProp-Entropy* for cross-domain topic classification and *ClsProp-Entropy* for single-domain topic classification) achieve better correlation with the performance of the topic classifier (an improvement above 30% in absolute values for the *Cri* and *DisAcc* topics), showing the usefulness of incorporating semantic features from KSs for enhancing the representation of a topic.

Given the above observations, the general findings about the entropy-based measures are as follows:

1. The adaptability of a topic classifier can accurately be measured by entropy-based measures defined over the concept graphs created for a topic from multiple linked KSs, outperforming previous content-based similarity measures derived from the sole text content

2. The usefulness of these entropy based measures varies among different topics and topic classification scenarios, however, the property based dispersion measures achieved best correlation values in both single-domain and cross-domain topic classification scenarios

6.7 Possible Future Directions

The proposed adaptive topic classification framework has several advantages: first: it *makes use of already existing KS data* for training a topic classifier of microposts; second it *exploits the structure and the knowledge from KSs* to improve the generalisation between domains, and thirdly it can *make use of the linked structure between the KSs through LOD* for providing a principled way for the combination of multiple KSs.
Despite of the success of this framework, several possible extensions could be explored:

- *Employing other NER extractors:*
  The presented semantic meta-graphs (both *resource meta-graph* and *category meta-graph*) have been shown to be both capable of providing contextual information about concepts in short text. The proposed topic classification framework makes use of various semantic features that are constructed from these semantic meta-graphs. By extracting the named entities the lexical feature space of a topic classifier is enhanced with additional contextual information about these concepts. In addition, this approach takes into account the information about concepts (e.g. resource type-hierarchies, resource properties) present in multiple semantic concept graphs of multiple linked KSs.

  One of the main factors which influences the performance of this approach is the performance of the named entity recogniser (NER) used to extract the named entities from short text messages. In this framework one of the most popular entity recogniser (e.g. OpenCalais and Zemanta) is employed for this purpose. Although there has

been several NER available [Rizzo and Troncy, 2011] for extracting entities for textual data, these approaches were built on newswire corpora, and therefore to date it is not well understood which one provides the best performance on microposts. Ritter et al. [2011] also showed that applying a generic NER tool yields a performance of 29% (in F1 measure) on tweets, while an NER tailored for tweets performs much better, achieving an F1 of 59%. Future work will thus concentrate in evaluating the framework using other NERs.

- *Improving the assignment of hashtags to DBpedia and Freebase URIs*:
  The assignment of hashtahs to DBpedia and Freebase URI uses a simple word matching approach. This however encountered various challenges, given that hashtags can often contain: 1. abbreviations (e.g. #nkorea `http://dbpedia.org/resource/North_Korea`); 2. contain compound words (e.g. #flightdelay `http://dbpedia.org/resource/Flight_delay`); and 3. some of the hastags may contain new abbreviations not present in KSs (e.g. #emfrmf) . For those cases no semantic meaning was assigned to them. In addition, one hashtag as any other word (#beirut) may have multiple meanings (e.g. the capital city of Lebanon; or a Lebanese governorate), and thus in order to assign the correct DB and FB URI one may apply a word sense disambiguation algorithm [Tatar, 2004] first, which takes into account not only the lexical form of a hashtag but also the context of the hashtag.

  Future work will thus aim at improving the automatic assignment of hashtags to DBpedia and Freebase URIs.

- *Accounting for the incompleteness and inconsistencies in KSs:*
  A second factor which have some drawback to the performance of this approach is the *incompleteness* and the *inconsistencies* within the KSs. For e.g. in Freebase the */crime/crime_ accuser* class is derived from a very generic */common/topic* class, while another related class type */crime/convicted_ criminal* extends the */people/person* class. One possible solution to overcome this problem could be to perform a cross-consistency validation, by investigating the overlapping properties between the entities assigned to the same entity classes, and consider the most likely entity classes [Dolby et al., 2009].

- *Investigating more sophisticated features: balancing the contribution of uni-grams and bi-grams by:*
  The proposed topic classification framework models the content of text using simple 1-gram (unigram) features. A possible extension of this approach could be to incorporate other ngram features into these models also, for e.g. 2-grams or a combination of 1-grams with 2-grams [Lampos, 2012].

- *Investigating the impact of different tweet normalisation approaches:*
  Given the vocabulary differences between KS data and microposts, one of the challenges faced by these models are the frequent usage of ungrammatical English words in microposts. Due to the restricted size of short messages, entities such as country names (*nkorea*) are often abbreviated, as in the following tweet: *"nkorea prepared nuclear weapons holy war south official tells state media usa"*. These irregularities results

in that current annotation services (including OpenCalais API) will ignore these entities, and therefore no semantic information will be exploited for these entities by the TC. A possible solution to address these challenges is to apply lexical normalisers especially developed for tweets [Han and Baldwin, 2011] to normalise these words to standard English terms[32].

6.8 Summary

This chapter explored the use of social knowledge sources (DBpedia and Freebase) for topic classification of *short* text messages, aimed at detecting the topic(s) of social media posts. The feasibility of this approach was demonstrated by implementing supervised classification models, which i) make use of the abundant amount of *data* within these sources as additional training data, and ii) exploit the *semantic information* present in KS concept graphs to enhance the representation of the documents.

Exploring the question *"Do KSs contain useful labelled data for building adaptive topic classifiers of microposts?"* it was found that both DBpedia and Freebase KSs provide valuable annotated data for training an adaptive topic classifier, and these KS data contain complementary information. The DB+FB classifier built on the joint KS data proved to be a competitive baseline, which can only be exceeded provided that sufficient amount of annotated microposts is available (more than 640 annotated posts for the Crime and War topics, and more than 940 labelled posts for the Disaster topic). Further the proposed adaptive DB+FB+TW classifier built on KS and Twitter data was shown to be very effective, achieving reasonably high accuracy against several baseline models, including TGT_ONLY (TW), SRC_TGT (DB+TW, FB+TW), and SRC_ONLY (DB+FB, DB, FB) models.

For enriching the documents, this chapter introduced and evaluated various semantic features derived from two distinct semantic concept graphs (*resource meta-graph* and *category meta-graph*). These enrichments were applied over the entities found in both KS documents and microposts. In addition to this, two other Twitter specific indicators (hashtags and URL) were also employed to further enrich the representation of tweets. For the hashtag features the concept graph enrichment strategies were applied, while for the URLs a list of keywords retrieved from the webpages were added to the lexical features of tweets. Experimental results revealed that the DB+FB+TW classifier using features from both semantic concept graphs improves upon the baseline models. The best overall results were obtained by employing the semantic property features extracted from the *category meta-graph* for both entities and hashtags, achieving significant improvement over various baseline models, including approaches considering sole lexical features (considering the SCR_TGT, TGT_ONLY and SRC_ONLY models), and approaches using semantic KS features about entities only (SRC_TGT). These experiments thus addressed the questions *"Do semantic meta-graphs built from KSs contain useful semantic features about entities for the topic classification (TC) of microposts?"* and *"Does information derived from external data sources'*

[32]Some initial experiments in this direction were already performed, employing a dictionary based approach for tweet normalisation. A lexicon from *http://www.noslang.com/* website was built, consisting of 5,407 abbreviation word pairs, and used to replace all abbreviations found in tweets with standard English terms. The initial results, however, showed no improvement upon the baseline model without normalisation. This indicate that further study needs to be conducted to investigate looking at other tweet normalisation approaches too.

indicators play an important role in TC of microposts?".

These observations have raised the final question *"Is it possible to predict the adaptability of a topic classifier?"*. To answer this question, this chapter introduced and evaluated various entropy-based measures defined over the semantic concept graphs. These experiments demonstrated that the performance of a topic classifier can be predicted with reasonably high accuracy using the property dispersion entropy measures. These results also showed a significant improvement over previous content-based lexical similarity measures.

Conclusions

Chapter 7

Conclusions and Outlook

The research presented in this thesis has examined how *text classification* can be better supported by harnessing knowledge from *domain knowledge sources* to enable the categorisation of documents in large heterogeneous repositories, spanning multiple domains and text types. A range of different repositories was explored, each capturing specific fragments of information related to events happening in the word (e.g. emergency landing). In order to provide a more complete and comprehensive picture about such events, this thesis presented techniques which extract information from these different repositories, performing text classification at multiple granularity levels. *Intra-document* text classification (a.k.a. *document zoning*) was used to recognise the structure of documents from *historical data* stored in organisational archives and large scientific (biomedical) repositories, and *whole-document* text classification (a.k.a. *topic classification*) was used to detect the topics of messages from social media repositories, providing up-to-date information about these events.

Chapter 2 introduced the task of text classification, and reviewed the state-of-the-art approaches on adaptive text classification using transfer learning techniques. Chapter 3 discussed the requirements of an adaptive text classification system, and presented an outline of a knowledge-driven approach for text classification. This approach relies on identifying a set of *pivot features* from the knowledge sources for adaptation, and then applying various *augmentation techniques* for incorporating these features into the adaptive text classifiers. First, adaptive text classification techniques for *document zoning* were studied. Chapter 4 discussed knowledge-poor approaches for document zoning which can be applied in the absence of any domain knowledge. Chapter 5 then presented methods for incorporating pivot features from domain knowledge sources into adaptive document zone classifiers in order to enhance the content of these documents. Finally, Chapter 6 described adaptive *topic classification* methods for Social Media posts making use of both data and knowledge from social knowledge sources. This chapter presented methodologies for gathering data from KSs for a particular topic/text class, serving as additional training data for building adaptive topic classifiers. In addition, a set of pivot features was derived by exploiting the underlying structure of KSs.

The current chapter summarises the main findings of this thesis and presents some future research directions. Section 7.1 returns to the research questions and claims presented in Chapter 1. Section 7.2 discusses the techniques and methodologies employed in this thesis with respect to the requirements set. Following this, Section 7.3 enumerates some possible

future directions on the overall research direction of transfer learning. Section 7.4 presents the closing statement to this thesis.

7.1 Research Conclusions

7.1.1 Analysis of Research Questions

The research conducted in this thesis centres around the main research question:

How can document classification be performed across multiple domains and text types?

In addressing this question, research was conducted on a range of repositories comprising documents belonging to different domains and text types. This lead to the development of novel approaches and methodologies for processing the vast amount of text within these repositories, producing as a result a semantic categorisation of their documents. The underlying text classification task involves the processing and analysis of document content, which may contain keywords, entities (instances of domain concepts) and special symbols, and mining this content in order to assign the correct semantic class(es) to the documents.

The analysis of different repositories revealed the main challenges in performing document classification across multiple domains. It was noted that the differences in vocabulary, style and language used in the different domains largely affect the performance of a text classifier. For instance the vocabulary used in documents about tropical medicine is very different to the vocabulary used in cell biology documents. Also, when dealing with social media repositories, the content of the documents can be very short, which restricts the contextual information of the documents, and can contain keywords, as well as special symbols such as hashtags, or links to external information.

In addressing these challenges it was highlighted the need for consistently representing the content of the documents across domains. For this purpose this thesis explored the use of semantics for document content enrichment. In order to achieve this, the use of knowledge sources was explored, and Semantic Web technologies applied as means for representing the knowledge within knowledge source ontologies. Further, besides exploiting the knowledge within knowledge source ontologies for enriching the content of the documents, this thesis also makes use of the textual data in knowledge source for providing additional training examples for transfer learning.

The main research question was further split into several research questions, as described in Chapter 1. These questions are the following:

1. *Is it possible to define automated techniques of text classification that are able to port across domains and text types?*

   Current approaches for building an adaptive text classification system rely on the selection of a set of *pivot features* for text classification, and application of different transfer learning strategies over these features to reduce the gap between domains. Traditional text classification approaches represent the content of the domain documents using the lexical information (words) found in them, and thus only consider lexical features as *pivot features* for adaptive text classification. This representation is, however, limited as it does not take into account any additional contextual information

about the documents. For instance, this representation ignores the entities mentioned in the documents, which could serve as potential *pivot features* for adaptation.

This thesis proposed the use of *semantic meta-graphs* for providing an alternative representation of the domain documents. This semantic meta-graph, introduced in Chapter 3, enables the exploitation of rich semantic information present in knowledge sources about entities and concepts found in the documents. This rich semantic information is then used to create novel *pivot features* for adaptation.

Throughout this thesis different semantic meta-graphs have been used for building *supervised adaptive text classifiers*. In particular, Chapter 5 and Chapter 6 presents two distinct *adaptive text classification* techniques, each creating semantic pivot features from different semantic meta-graphs.

OntoEA (Chapter 5) employs the resource meta-graph for document zoning, capturing the class information (type) associated with entities in knowledge sources. From this graph, OntoEA derives the semantic class features as pivot features for adaptive document zoning. Following this, OntoEA augments the original lexical feature spaces of domains with new semantic pivot class features using a feature duplication technique, which allows to automatically learn the domain-specific and domain-independent weight for the features. Experimental results comparing the performance of OntoEA with four baseline models, showed the superiority of this model: (it achieved an improvement in F1 measure of 2.4-7.5% against the classifier built on the joint source and target domain data (SRC_TGT), 2.6-8.4% against EasyAdapt, 6.3-32.8% against the classifier built on the source domain data (SRC_ONLY) and 1.1-6.3% against the classifier built on target domain data (TGT_ONLY)). These results demonstrate that this enhanced document representation provide a novel way for performing adaptive document zoning.

The adaptive topic classifier (Chapter 6) makes use of multiple semantic meta-graphs as well as data found in knowledge sources for topic classification. As an initial step annotated documents are retrieved from knowledge sources, corresponding to the source domain documents. Next two different semantic meta-graphs are employed for content enrichment: the resource meta-graph capturing the classification of entities based on their types, and the category meta-graph capturing the categorisation of concept based on their topics. From these graphs, three different semantic pivot features were considered: class, property and categories. This classifier then augments the original lexical feature spaces of domains with these new semantic pivot features, and also uses these graphs to assign appropriate weights for the features. Experimental results comparing the performance of topic classifier with two baseline models, showed the superiority of this model (achieving an improvement in F1 measure above 6% against the classifier built on the joint knowledge source data and microposts (SRC_TGT) and 14% against the classifier built on microposts (TGT_ONLY)). These results demonstrate that this enhanced document representation provide a novel way for performing adaptive topic classification.

2. *Can labelled data be gathered inexpensively to build adaptive text classifiers?*
One of the main bottlenecks of applying *supervised* transfer learning for text classification, is the need for a considerable amount of annotated data, which is often expensive

and time consuming to obtain. Instead, this thesis explores the use of data leveraged from knowledge sources to support such techniques. Chapter 5 proposed a methodology for topic classification which gathers annotated data for a particular topic from multiple social knowledge sources (DBpedia (DB) and Freebase (FB)). The collected data is then combined following linked data principles, and used to build several SVM topic classifiers: a baseline topic classifier using knowledge source data alone (called $DB+FB$ classifier), and another adaptive topic classifier using both knowledge source and target domain data (called $DB+FB+TW$ classifier). When comparing different classifiers, the results revealed that the $DB+FB$ classifier is difficult to beat, a considerable amount of annotated target domain data is needed to outperform it (more than 640 annotated tweets for the Crime and War topics, and above 940 labelled posts for the Disaster topic). Further, the combined adaptive topic classifier $DB+FB+TW$ also outperforms the topic classifier built on the target domain data. These results demonstrate that social knowledge sources can be used as background knowledge for topic classification.

3. *Is it possible to define a measure for quantifying the adaptability of a text classifier? (when to transfer)*
   This thesis presented two *domain similarity measures* for providing an estimate on the performance of an adaptive text classifier. In contrast to previous content-based domain similarity measures, these measures make use of the enhanced document representation exploiting contextual information about entities from semantic meta-graphs.

   Chapter 5 described an effective approach for measuring the *similarity between domains* for document zoning, requiring no labelled data. As an initial step, a graphical model introduced in Chapter 4 is employed for partitioning the paragraphs of the documents into zone clusters. Following this, the final similarity score is computed by the combination of statistical content-based and knowledge-based measures on the generated zone segments. The statistical measures are computed over the lexical representation of the zones, while the knowledge-based measures are computed over the concepts found in the zones. The experimental results correlating these similarity values with the performance of an adaptive document zone classifier showed superior results to the values obtained by content-based similarity measures (achieving an improvement of 3.06-75.8% in absolute value). These results demonstrate that the performance of an adaptive document zone classifier can be predicted using the enhanced document representation.

   Chapter 6 proposed entropy-based measures for quantifying the *similarity between domains* for topic classification. These measures aim to capture the informativeness of the source and target domains, by considering the entities discussed in them, and computing the entropy of each such entity over the employed semantic meta-graphs. The final similarity score is then computed by subtracting the entropy scores of the two domains. In doing so, lower entropy difference values indicate that the domains are close to one another, having features with similar values in the two domains. Experimental results showed that these entropy-based similarity measures achieved an improvement in correlation values (above 30% in absolute values for two out of three topics) compared to baseline measures. These results demonstrate that the perfor-

mance of an adaptive topic classifier can be predicted using the enhanced document representation.

4. *Is the effectiveness of adaptive methods comparable to in-domain machine learning methods?*
   This thesis presented two *adaptive text classification* techniques, each exploiting the semantic information from different semantic meta-graphs.

   In order to evaluate the effectiveness of these supervised adaptive text classifiers, extensive experiments have been conducted considering real world data spanning a range of different domains. When comparing the performance of the OntoEA to four strong baseline methods (EasyAdapt, the classifier built on the source and target data (SRC_TGT), the classifier built on the source domain data (SRC_ONLY), and the classifier built on the target domain data (TGT_ONLY)) in Chapter 5, experimental results demonstrated that OntoEA consistently outperforms existing models (as mentioned above). These results demonstrate that adaptive document zone classifiers outperform traditional machine learning models. When comparing the performance of the adaptive topic classifier to strong baseline methods (SRC_TGT, TGT_ONLY) in Chapter 6, experimental results demonstrated that this approach outperforms existing models. These results demonstrate that adaptive topic classifiers are superior to traditional machine learning models (as mentioned above).

7.1.2 Analysis of Claims

Chapter 5 and Chapter 6 presented the evaluation of adaptive text classification techniques using different semantic meta-graphs derived from multiple knowledge source ontologies. These results revealed that using certain pivot features from these semantic meta-graphs together with special feature combination strategies outperform baseline approaches. For instance, the *adaptive document zone classifier* presented in Chapter 5 explores semantic meta-graphs from complementary biomedical knowledge source ontologies (SNOMED and MESH). In this case, using semantic class pivot features from both ontologies combined with CCA dimensionality reduction strategy consistently outperform the classifier built on the joint source and target domain data (SRC_TGT), the classifier built on the target domain data (TGT_ONLY) and EA approaches in terms of precision, recall and F1 measure for all the analysed domain pairs. This ontology combination strategy was also found superior to the results obtained for the individual ontologies and the ad-hock (naive) combination of ontologies, which only achieved comparable results to baseline approaches.

The *adaptive topic classification technique* from Chapter 6 explored several pivot features created from two different variants of semantic meta-graphs constructed from complementary social knowledge sources (DBpedia and Freebase). Of the evaluated pivot features, the semantic property features exploited from the combined knowledge source ontologies were found to consistently outperform the SRC_TGT and TGT_ONLY classifiers, achieving high F1 measure, over 96% for all the analysed domain pairs. Comparing the results obtained using individual knowledge source ontologies, however, the property features only show improvement in terms of recall. Nonetheless, these results indicate that adaptive topic

classification techniques return accurate results, supporting the following claim:

- *Domain knowledge sources contain useful semantic structures from which pivot features can be obtained for adaptive text classification*

One can conclude that the explored knowledge sources provide valuable semantic information from which pivot features can be created for adaptive text classification. These sources support the contextual enrichment of domain documents, using these pivot features, which were found to decrease the gap between domains, yielding promising results for both document zoning and topic classification of microposts.

Besides exploiting the semantic information present in knowledge sources, another key contribution of this thesis is the use of data from knowledge sources to support adaptive text classification. In the introduction of this thesis it has been highlighted the importance of providing a large amount of annotated data for building high accuracy cross-domain text classifiers. Chapter 6 proposed an approach for collecting such labelled data from multiple social knowledge sources (DBpedia and Freebase) for building topic classifiers of microposts. This approach relies on exploiting the structure of these knowledge sources, and retrieving a set of documents whose topic correspond to the domain (or topic) of interest. This data is then used to build two topic classifiers of microposts: a baseline SRC_ONLY topic classifier on knowledge source data alone, and an adaptive text classifier which utilises this data as additional training examples to micropost data. The results from the evaluation quantified the performance of these classifiers against several baseline models, including the topic classifier built on microposts only. This evaluation revealed that the SRC_ONLY classifier yields very competitive results which is difficult to beat, and further the adaptive topic classifier outperforms the in-domain classifier with a large margin (as mentioned above). These results demonstrated that data obtained from social knowledge sources provides topic classification techniques with the necessary background information about a particular topic or domain. These results therefore support the following claim:

- *Data found in domain knowledge sources can be used to train an adaptive text classifier*

One of the key contribution of this thesis is the proposal of domain similarity measures for predicting the performance of an adaptive text classifier. By quantifying the similarity between the source and target domains an insight can be provided into the degree to which the classifier built on the source domain data performs well on the target domain. Designing a domain similarity measure which achieves high correlation with the performance of an adaptive text classifier, having a predefined threshold for the correlation, could thus predict the relevance of source domain data for the target domains. This could also be enable one to select a representative sample of source domain data for creating an adaptive text classifier for the target domain.

Chapter 5 and Chapter 6 have presented two domain similarity measures computed over the enhanced document representation: one for document zoning, and another one for topic classification. These measures both achieved correlation values above 70%. In particular, in the case of document zoning, the correlation figures were consistently higher than 70%, while in the case of topic classification only two out of the three domains obtained high correlation values. In order to understand the reasons behind the variations in correlation values for the topic classification task, future work will be conducted. Considering that

these correlation scores indicate a strong agreement with the performance of an adaptive text classifier - following the agreement scores found in the literature [Mihalcea et al., 2006], this suggest that the following claim is supported:

- *The accuracy of a text classifier can be measured as a function of conceptual representation of the domain documents*

For each of the proposed adaptive text classification technique this thesis presented an extensive evaluation using different domains, and knowledge sources. For instance, Chapter 5 evaluated the adaptive document zone classifier using seven biomedical sub-domains (corresponding to 42 adaptation scenarios), and two broad coverage and widely used biomedical knowledge sources (SNOMED and MESH). Applying the F1 measure, the proposed classifier consistently outperformed several baseline machine learning models with a gain between 1.1-6.3% (over TGT_ONLY) and 2.4-7.5% over SRC_TGT, which was shown to be statistically significant using the t-test ($p < 0.05$).

Similarly, Chapter 6 used three different emergency response domains to evaluate the proposed adaptive topic classification approach, considering two other broad coverage popular social knowledge sources (Freebase and DBpedia) for domain content enrichment. Experimental results over strong baseline models also showed a consistent improvement above 7% for each domain pairs. Despite of the limited number of domains, these results empirically demonstrated that adaptive topic classification techniques perform well when supported by semantic information from knowledge sources. These results thus validated the following claim:

- *Adaptive text classification techniques exploiting domain knowledge sources are able to achieve comparable results to in-domain machine learning approaches*

7.2 Analysis of Methodology Requirements

Chapter 3 has presented the state-of-the-art work for performing text classification across domains. Following the analysis of such techniques, a set of requirements were created, which adaptive text classification must fulfil. These requirement were presented in Chapter 3, together with the design of a knowledge-driven approach to adaptive text classification.

The proposed approach is divided into four main steps: annotated data gathering and content modelling, semantic meta-graph generation from knowledge sources, pivot feature creation and text classification. As discussed in Chapter 3, this approach follows a *supervised* transfer learning setting, which was found to be most suitable to the problem setting of this body of work. In light of this decision, several requirements which supervised techniques must fulfil were derived. These requirements are now analysed in relation to the presented work:

7.2.1 Requirements for Adaptive Text Classification

- *Perform adaptive text classification with minimal supervision:*
  Chapter 6 has presented an approach which alleviates the need for manually annotated data for topic classification of social media posts. This approach generates annotated

data from social knowledge sources (DBpedia and Freebase) for a particular domain (or topic), by exploiting the structure of these knowledge sources and selecting articles whose topic correspond to the domain of interest. This data is then used as additional training data for building an adaptive text classifier. Experimental results have shown that this adaptive text classifier outperforms the in-domain TGT_ONLY classifier with a large margin of up to 33.1% in terms of F1 measure.

Chapter 5 has presented an analysis for examining the performance of proposed adaptive document zone classifier. In these experiments several baseline models are compared against the adaptive classifier enhanced with KS pivot features. As discussed in the results section of Chapter 5, in the majority of the domain adaptation scenarios, the performance of this classifier was better than that of TGT_ONLY and SRC_TGT classifiers, requiring less number of annotated in-domain examples (less than 50% of annotated training data). These results thus revealed that this approach is feasible to reduce the number of annotated examples needed to build a document zone classifier.

- *Achieve classification accuracy comparable to supervised machine learning approaches:* The evaluation of the proposed adaptive text classifiers has provided an empirical evidence on the performance of the explored techniques against several baseline supervised machine learning approaches. These results presented in Chapter 5 and Chapter 6 indicate a comparable level of performance with supervised approaches.

  The adaptive *supervised* document zone classifier explored within Chapter 5 which combines multiple semantic meta-graphs via dimensionality reduction achieves comparable results in terms of F1 to supervised TGT_ONLY and SRC_ONLY classifiers for half of the analysed domain scenarios (around 20 domain pairs), but perform much better for the other half of the domain pairs (achieving an improvement of over 5% in F1 measure).

  For the *adaptive topic classifier* presented in Chapter 6, which combines multiple KS semantic meta-graphs in an ad-hoc manner and also makes use of twitter specific pivot features (about hashtags), the results obtained in terms of F1 measure were better than those of the baseline models, with a consistent improvement of 6% over SRC_TGT, and over 14% compared to TGT_ONLY. This improvement is also considerable higher to the results obtained without incorporating twitter specific pivot features, in which case the improvement over the SRC_TGT classifier were only comparable (around 2% in F1 measure). These results demonstrate that the twitter specific semantic features about hashtags play an important role, and the final classifier incorporating them is suitable for detecting the topics of microposts.

- *Enable the creation of pivot features from knowledge sources*
  The adaptive document zone classifier used in Chapter 5 enables the creation of pivot features from two biomedical knowledge sources (SNOMED and MESH). These features are used to augment the original feature spaces of both source and target domains. Employing the semantic class features as pivot features for transfer learning, the experimental results revealed that both knowledge source ontologies provide useful pivot features for text classification, and further these ontologies complement each other well, outperforming several baseline models. These results thus demonstrate that the

employed knowledge sources are representative of the biomedical domain and support adaptive document zoning.

Chapter 6 proposed strategies for the creation and weighting of pivot features from two social knowledge sources (DBpedia and Freebase). Of the evaluated pivot features, the semantic property features have been found the most effective, for which a novel feature weighting strategy was also presented capturing the importance of this feature for a given domain (topic). Experimental results revealed that DBpedia has a broader coverage of entities than Freebase for the analysed emergency domains, resulting in a larger number of pivot features being added to the adaptive topic classifier. Nonetheless both knowledge source ontologies have been found to provide useful pivot features which improve upon different baseline models. In addition, these knowledge sources were also found to complement each other well, the best overall results being achieved by the joint ontologies.

- *Be able to predict the performance of an adaptive text classification:*
  Chapter 5 presented a domain similarity measure for document zoning, which combines statistical corpus-based similarity measures with knowledge-based measures. The evaluation results correlating these values to the performance of the proposed adaptive document zone classifier showed relatively high correlation, above 70%, indicating the effectiveness of this measure in predicting the performance of an adaptive document zone classifier.

  Chapter 6 presented a domain similarity measure for topic classification, which computes the entropy difference between the source and target domain for a particular feature. Of the evaluated entropy measures, the entity-property entropy difference measure achieved the highest correlation values with the performance of an adaptive topic classifier, above 70% for two out of three domains, demonstrating the applicability of this measure to detect the performance of a text classifier.

- *Comply with the limitations and constraints posed by real-world application scenarios:*
  The presented text classification approaches rely on the application of shallow NLP techniques, employing simple lexical BoW features and semantic feature about entities, making them easily applicable to real-world scenarios imposing constraints. For instance, the lack of use of any formatting features by these techniques allows the application of tools used by largest majority of search engines, including corporate engines like FAST Enterprise Search Platform, which have a very large adoption in corporate environments.

  The approaches presented in Chapter 5 and Chapter 6 make only use of lexical BoW features as initial feature space, and semantic features about entities as semantic pivot features for adaptation. The extraction of entities being done by off-the-self entity recognition tools. Both models achieved superior results to several baseline models.

  The documents used in the zoning experiments reported in Chapter 4 were preprocessed using the FAST engine pipeline. As a result, only the textual information (words) were kept from the documents, ignoring any additional information such as tables and figures. The presented graphical models then employed a simple BoW representation over these documents for the identification of zones within the documents,

showing promising results over a baseline graphical model.

7.3 Future Directions

The research presented within this thesis covers a broad spectrum of work, however in certain cases the presented methods could further be extended. This section thus provides some possible future research directions on the general transfer learning problem, resulted from the detailed revision of state-of-the-art approaches and general observations obtained from this research:

- *Reducing the number of annotated examples needed for learning:*
  The majority of the work still rely on labelled data from the source [Jiang and Zhai, 2007a; Blitzer et al., 2006; Guo et al., 2009] and target domains [Dai et al., 2007b; Daumé III et al., 2010; Arnold et al., 2008]. However, it has been shown that creating and maintaining labelled data is both time consuming and expensive [Ciravegna et al., 2002; Zhang et al., 2010]. Although some methods leveraged unlabelled data in unsupervised [Dai et al., 2008; Huang and Yates, 2010] and semi-supervised manner [Dai et al., 2007a; Jiang and Zhai, 2007a], most of these methods were tested on specific IE tasks (such as part-of-speech-tagging, sentiment classification), and it is unclear whether these approaches are generalisable to other text classification tasks. A promising avenue of research in this direction could therefore be to investigate *semi-supervised* approaches which makes use of only a small amount of annotated data from the source domain and no annotated data from the target domain [Zhu, 2005].

- *Combining different transfer learning approaches into a unified framework:*
  Chapter 2 presented distinct transfer learning techniques which have been successfully applied to many Natural Language Processing tasks, including *instance-based* or *feature-representation-based* learning. While these individual approaches have been successfully applied on a variety of different problems, it is not very clear whether combining the advantages of the individual approaches could furthermore boost the performance of a transfer learning classifier [Jiang, 2008b; Japkowicz and Stephen, 2002].
  Previous efforts on combining these approaches in a straightforward manner [Jiang, 2008b] did not show consistent improvement, suggesting that a systematic and generalisable methodology would be required to investigate whether combining these approaches in beneficial for transfer learning.

- *Harnessing multiple knowledge sources:*
  The majority of transfer learning approaches exploit an individual knowledge source for learning [Pan and Yang, 2010]. These approaches can be beneficial when the selected knowledge source provide large enough coverage of the entities and data for a particular domain. When this situation does not hold, however, an appealing solution could be to exploit alternative knowledge sources for the domain at hand, for example from *linked data* [Bizer et al., 2009].

  Harnessing information from multiple knowledge sources, however, can pose further challenges in learning. One of the challenge arises when the same information about

an entity is represented in different ways, for example by using distinct names for the same property in the different knowledge sources: such as *inhabits* and *vote* in Freebase, while *resident* and *elect* in DBpedia. Without considering these repetitions, the same information could be represented twice, resulting in a much larger feature space. Another challenge comes from the *incompleteness* and the *inconsistencies* within the knowledge sources. For instance, in Freebase */crime/crime_accuser* class is derived from a very generic */common/topic* class, while another related class type */crime/convicted_criminal* extends the */people/person* class. In the case of the category structure of Wikipedia, it can also noticed that the category tree is not a strict taxonomy and does not always contain an *is-a* relationship [Gabrilovich and Markovitch, 2006].

Investigating alternative approaches for the combination of data and knowledge from multiple knowledge sources, (e.g. resolving redundancies and inconsistencies) are thus of increased importance when dealing with multiple complementary knowledge sources [Zhang et al., 2013].

- *Investigating multi-domain transfer learning scenarios:*
  The majority of transfer learning approaches have been designed for a single domain pair, using one source domain to learn from, and one target domain to evaluate the approach on [Pan and Yang, 2010]. However, for many real world application scenarios, e.g. corporate or large environments, there may be more than one annotated domain corpora available. A possible avenue of research could thus be to investigate *multi-domain* scenarios for transfer learning, when more than one source domain is used.

7.4 Closing Statement

This thesis explored the use of domain knowledge for text classification across multiple domains and text types. Three main research questions have been investigated concerning adaptive text classification: *what to transfer*, which identifies a set of pivot features from knowledge sources for reducing the distributional differences between domains; *how to transfer*, which studies various augmentation strategies for incorporating the pivot features into the text classifiers, and *when to transfer*, which quantifies the transfer ability of a text classifier as a function of pivot features.

A number of experiments have been conducted on two different text classification tasks (document zoning and topic classification) considering a large number of domains, contributing to a series of methods and findings which address these research questions. Empirical evidence from the evaluation of the explored adaptive text classification methods indicates that knowledge sources are valuable resources for reducing the gap between domains, achieving results superior to traditional supervised machine learning techniques.

Overall, the presented approaches demonstrated that knowledge sources: i) provide a rich set of semantic features which are stable across domains, ii) contain useful data which can provide additional training data for building adaptive text classifiers, and iii) provide an enhanced semantic representation of the documents, which can be used as a measure for predicting the adaptability of a text classifier, outperforming previous content-based lexical similarity measures.

Bibliography

Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. In *Proceedings of the 19th international conference on User modeling, adaption, and personalization*. Springer-Verlag, 2011.

Shashank Agarwal and Hong Yu. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics*, 25(23):3174–3180, December 2009. ISSN 1367-4811.

CharuC. Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 163–222. Springer US, 2012. ISBN 978-1-4614-3222-7.

Samir Al-Stouhi and Chandan K. Reddy. Multi-task clustering using constrained symmetric non-negative matrix factorization. In *SIAM INTERNATIONAL CONFERENCE ON DATA MINING (SDM) 2014*, 2014.

Samir Al-Stouhi and Chandan K. Reddy. Adaptive boosting for transfer learning using dynamic updates. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 60–75, 2011.

Rie Kubota Ando. Exploiting unannotated corpora for tagging and chunking. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 142–145, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Rie Kubota Ando and Tong Zhang. A high-performance semi-supervised learning method for text chunking. In *Proceedings of the Association for Computational Linguistics*, 2005.

Dina Demner-Fushman Alan Aronson Antonio Jimeno Yepes, James Mork. Comparison and combination of several mesh indexing approaches. *American Medical Informatics Association*, (4):432–440, 2013.

Andrew Arnold and William W. Cohen. Intra-document structural frequency features for semi-supervised domain adaptation. In *Conference on Information and Knowledge Management (CIKM)*, pages 1291–1300, 2008.

Andrew Arnold, Ramesh Nallapati, and William W. Cohen. Exploiting Feature Hierarchy for Transfer Learning in Named Entity Recognition. In *Proceedings of the Association for Computational Linguistics*, pages 245–253, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC*, pages 722–735, 2007.

Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *COLING (Posters)*, pages 36–44. Chinese Information Processing Society of China, 2010.

Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 113–120, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Neural Information Processing Systems (NIPS)*, pages 137–144, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2009.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *ICML*, pages 81–88, 2007.

Brigitte Bigi. Using kullback-leibler distance for text categorization. In *Advances in Information Retrieval, Lecture Notes in Computer Science Volume 2633*, 2003.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 2009.

Heather Blanchard, Andy Carvin, Melissa Elliott Whitaker, Merni Fitzgerald, Wendy Harman, and Brian Humphrey. The case for integrating crisis response with social media. 2012.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In *J. Mach. Learn. Res. 3*, pages 993–1022, 2003a.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. 2003b.

John Blitzer. *Domain Adaptation of Natural Language Processing Systems*. PhD thesis, 2008.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 2006.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortmanblind. Learning bounds for domain adaptation. In *Neural Information Processing Systems (NIPS)*, 2007a.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. ACL, 2007b.

John Blitzer, Dean Foster, and Sham Kakade. Domain adaptation with coupled subspaces. In *Conference on Artificial Intelligence and Statistics*, Fort Lauterdale, 2011.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, WI, 1998.

O Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32, 2004.

Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, 2008.

Jonathan Butters and Fabio Ciravegna. Authoring technical documents for effective retrieval. In *EKAW*, pages 287–300, 2010.

Jonathan Butters and Fabio Ciravegna. Using similarity metrics for terminology recognition. In *Proceeding of the International Conference on Language Resources and Evaluation (LREC)*, 2008.

Amparo Elizabeth Cano, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. Making sense of microposts (#msm2013) concept extraction challenge. 2013.

Amparo Elizabeth Cano Basave, Giuseppe Rizzo, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. Making sense of microposts (#microposts2014) named entity extraction & linking challenge. In *4th Workshop on Making Sense of Microposts (#Microposts2014)*, pages 54–60, 2014.

Gregory Caporaso, William Baumgartner, Hyunmin Kim, Zhiyong Lu, Helen L. Johnson, Olga Medvedeva, Anna Lindemann, Lynne M. Fox, Elizabeth K. White, K. Bretonnel Cohen, and Lawrence Hunter. Concept recognition, information retrieval, and machine learning in genomics question-answering. In Ellen M. Voorhees and Lori P. Buckland, editors, *Text REtrieval Conference (TREC)*, volume Special Publication 500-272. National Institute of Standards and Technology (NIST), 2006.

David Carmel, Haggai Roitman, and Naama Zwerdling. Enhancing cluster labeling using wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 139–146, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. Learning probabilistic linear-threshold classifiers via selective sampling. In *COLT*, pages 373–387, 2003.

Ciprian Chelba and Alex Acero. Adaptation of maximum entropy capitalizer: Little data can help a lot. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 285–292, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Massimiliano Ciaramita and Olivier Chapelle. Adaptive parameters for entity recognition with perceptron hmms. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP 2010)*, pages 1–7, Uppsala, Sweden, July 2010. Association for Computational Linguistics (ACL).

Massimiliano Ciaramita and Altur Yasemin. Named-entity recognition in novel domains with external lexical knowledge. In *Advances in Structured Learning for Text and Speech Processing Workshop*, 2005.

Fabio Ciravegna, Alexiei Dingli, Daniela Petrelli, and Yorick Wilks. Timely and non-intrusive active document annotation via adaptive information extraction. In *In Proc. Workshop Semantic Authoring Annotation and Knowledge Management (European Conf. Artificial Intelligence)*, pages 7–13, 2002.

Michael Collins. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 1–8, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

Peter Corbett and Ann A. Copestake. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics*, 9(S-11), 2008.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai S. Shwartz, and Yoram Singer. Online Passive-Aggressive Algorithms. *J. Mach. Learn. Res.*, 7:551–585, December 2006. ISSN 1532-4435.

Andrew M. Dai and Amos J. Storkey. The grouped author-topic model for unsupervised entity resolution. In *Proceeding of International Conference on Artificial Neural Networks (ICANN)*, 2011.

Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Transferring naive bayes classifiers for text classification. In *Proceedings of the 22nd National Conference on Artificial Intelligence*. AAAI Press, 2007a.

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *ICML*, pages 193–200, 2007b.

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Self-taught clustering. In *ICML*, pages 200–207, 2008.

Hal Daumé, III. Bayesian multitask learning with latent hierarchies. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 135–142, Arlington, Virginia, United States, 2009. AUAI Press. ISBN 978-0-9749039-5-8.

Hal Daumé, III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Francesco Dinuzzo and Bernhard Schölkopf. The representer theorem for hilbert spaces: a necessary and sufficient condition. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 189–196, 2012.

Julian Dolby, Achille Fokoue, Aditya Kalyanpur, Edith Schonberg, and Kavitha Srinivas. Extracting Enterprise Vocabularies Using Linked Open Data. In *Proceedings of the 8th International Semantic Web Conference (ISWC2009)*. Springer, 2009.

Gregory Druck, Gideon Mann, and Andrew McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 595–602, New York, NY, USA, 2008a. ACM. ISBN 978-1-60558-164-4.

Gregory Druck, Gideon Mann, and Andrew McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 595–602, New York, NY, USA, 2008b. ACM. ISBN 978-1-60558-164-4.

Lan Du, Wray Buntine, Huidong Jin, and Changyou Chen. Sequential latent dirichlet allocation. *Knowledge and Information Systems*, 31(3):475–503, 2012. ISSN 0219-1377.

Susan T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 2004.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Jenny Rose Finkel and Christopher D. Manning. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 602–610, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-41-1.

George Forman, Isabelle Guyon, and André Elisseeff. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 2003.

Evgeniy Gabrilovich and Shaul Markovitch. Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge. In *Proceedings of Twenty-First AAAI Conference on Artificial Intelligence*. AAAI Press, 2006.

Andres Garcia-Silva, Jorge Gracia, and Oscar Corcho. Associating semantics to multilingual tags in folksonomies. In *Proceedings of the 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2010)*, 2010.

Vijay Garla and Cynthia Brandt. Ontology-guided feature engineering for clinical text classification. *Journal of Biomedical Informatics*, 45(5):992–998, 2012.

Yegin Genc, Yasuaki Sakamoto, and Jeffrey V. Nickerson. Discovering context: classifying tweets through a semantic transform based on wikipedia. In *Proceedings of the 6th international conference on Foundations of augmented cognition: directing the future of adaptive systems*. Springer-Verlag, 2011.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. pages 1–6, 2009.

Andrew Goldberg and Xiaojin Zhu. Seeing stars when there aren?t many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52, New York City, June 2006. Association for Computational Linguistics.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In *Neural Information Processing Systems (NIPS)*, pages 537–544. 2004.

Tom Griffiths. Gibbs sampling in the generative model of Latent Dirichlet Allocation. Technical report, Stanford University, 2002.

Ralph Grishman and Beth Sundheim. Message understanding conference- 6: A brief history. In *International Conference on Computational Linguistics (COLING)*, pages 466–471, 1996.

Amit Gruber, Michal Rosen-zvi, and Yair Weiss. Hidden topic markov models. In *In Proceedings of Artificial Intelligence and Statistics*, 2007.

Honglei Guo, Li Zhang 0007, and Zhong Su. Empirical study on the performance stability of named entity recognition model across domains. In *EMNLP*, pages 509–516, 2006.

Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, Xian Wu, and Zhong Su. Domain adaptation with latent semantic association for named entity recognition. In *Proc. HTL-NAACL*, pages 281–289, June 2009.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Johan Hogberg, and Ulla Stenius. A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics*, 12:69, 2011a.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 273–283, Edinburgh, Scotland, UK., July 2011b. Association for Computational Linguistics.

Yufan Guo, Roi Reichart, and Anna Korhonen. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 928–937, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Rahul Gupta and Sunita Sarawagi. Domain adaptation of information extraction models. *SIGMOD Record*, 37(4):35–40, 2008.

Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In *ACL*, pages 368–378, 2011.

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis; an overview with application to learning methods. Technical report, Royal Holloway, University of London, May 2003.

Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, March 1997. ISSN 0891-2017.

Kenji Hirohata, Naoaki Okazaki, and Sophia Ananiadou. Identifying sections in scientific abstracts using conditional random fields. In *International Joint Conference on Natural Language Processing*, 2008.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. Yago2: a spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence Journal, Special Issue on Artificial Intelligence, Wikipedia and Semi-Structured Resources*, 2012.

Fei Huang and Alexander Yates. Exploring representation-learning approaches to domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 23–30, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, pages 601–608, 2007.

Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 465–474, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1869-3.

Stephanie Husby and Denilson Barbosa. Topic classification of blog posts using distant supervision. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 28–36, Avignon, France, April 2012. Association for Computational Linguistics.

IHTSDO. Snomed ct: Systematized nomenclature of medicine-clinical terms, 2010.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449, 2002.

Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pages 19–33, 1997.

Jing Jiang. A literature survey on domain adaptation of statistical classifiers. Technical report, 2008a.

Jing Jiang. *Domain Adatation in Natural Language Processing.* PhD thesis, University of Illinois at Urbana-Champaign, 2008b.

Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, 2007a.

Jing Jiang and ChengXiang Zhai. A two-stage approach to domain adaptation for statistical classifiers. In *CIKM*, pages 401–410, 2007b.

Thorsten Joachims. Estimating the generalization performance of a SVM efficiently. In Pat Langley, editor, *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 431–438, Stanford, US, 2000. Morgan Kaufmann Publishers, San Francisco, US.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 200–209, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-612-2.

Cristina Kadar and José Iria. Domain adaptation for text categorization by feature labeling. In *ECIR*, pages 424–435, 2011.

Eugen Kagan, Irad Ben-Gal, Natalya Sharkov, and Oded Maimon. Unsupervised zoning of scientific articles using huffman trees. In *Electrical and Electronics Engineers in Israel, 2008. IEEEI 2008. IEEE 25th Convention of*, pages 399 –402, dec. 2008.

Shiva Prasad Kasiviswanathan, Prem Melville, Arindam Banerjee, and Vikas Sindhwani. Emerging topic detection using dictionary learning. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 745–754, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0717-8.

Adam Kilgarriff. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1): 1–37, 2001.

J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19A:27–43, 1982.

Vasileios Lampos. Detecting events and patterns in large-scale user generated textual streams with statistical learning methods. CoRR, 2012.

Claudia Leacock and Martin Chodorow. *Combining Local Context and WordNet Similarity for Word Sense Identification*, chapter 11, pages 265–283. The MIT Press, May 1998.

Douglas B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, pages 33–38, 1995.

V.I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 1966.

Jianqiang Li, Yu Zhao, and Bo Liu. Fully automatic text categorization by exploiting wordnet. In *Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology*, AIRS '09, pages 1–12, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-04768-8.

Ying Li, Sharon Lipsky Gorman, and Noemie Elhadad. Section classification in clinical notes using supervised hidden markov model. In Tiffany C. Veinot, Ümit V. Çatalyürek, Gang Luo, Henrique Andrade, and Neil R. Smalheiser, editors, *IHI*, pages 744–750. ACM, 2010. ISBN 978-1-4503-0030-8.

Maria Liakata. Zones of conceptualisation in scientific papers: a window to negative and speculative statements. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 1–4, Uppsala, Sweden, July 2010. University of Antwerp.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Corpora for the conceptualisation and zoning of scientific papers. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 375–384, 2009.

Dekang Lin. An information-theoretic definition of similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998.

Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 65–72, New York, New York, June 2006. Association for Computational Linguistics.

Jimmy Lin, Rion Snow, and William Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011.

Jimmy J. Lin. Is searching full text more effective than searching abstracts? *BMC Bioinformatics*, 10, 2009.

Bing Liu, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. Text classification by labeling words. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, pages 425–430. AAAI Press, 2004. ISBN 0-262-51183-5.

LOD. http://lod-cloud.net/state/, 2011.

Julie Beth Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 1968.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8, 1988.

Tamara Martín-Wanton, Julio Gonzalo, and Enrique Amigó. An unsupervised transfer learning approach to discover topics for online reputation management. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pages 1565–1568, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8.

Larry McKnight and Padmini Srinivasan. Categorization of sentence types in medical abstracts. *AMIA Annu Symp Proc*, pages 440–444, 2003.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013.

Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12*, pages 563–572, Seattle, Washington, USA,, 2012.

Stephen Merity, Tara Murphy, and James R. Curran. Accurate argumentative zoning with maximum entropy models. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, pages 19–26, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-58-9.

Matthew Michelson and Sofus A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*. CEUR, 2010.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, pages 775–780, 2006.

Lilyana Simeonova Mihalkova. *Learning with Markov Logic Networks: Transfer Learning, Structure Learning, and an Application to Web Query Disambiguation*. PhD thesis, University of Texas at Austin, 2009.

Claudiu Mihăilă, Riza Theresa Batista-Navarro, and Sophia Ananiadou. Analysing entity type variation across biomedical subdomains. In Sophia Ananiadou, Kevin Cohen, Dina Demner-Fushman, and Paul Thompson, editors, *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012)*, May 2012.

Peter Mika, Massimiliano Ciaramita, Hugo Zaragoza, and Jordi Atserias. Learning to tag and tagging to learn: A case study on wikipedia. *IEEE Intelligent Systems*, 23(5):26–33, 2008.

D. Milne and I. H. Witten, editors. *Learning to link with Wikipedia.* 2008.

T. Mitchell. *Machine Learning.* McGraw-Hill Education (ISE Editions), 1st edition, October 1997. ISBN 0071154671.

Yusuke Miyao, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, and Takashi Ninomiya. Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *In Proceedings of Coling/ACL 2006*, pages 1017–1024, 2006.

Yoko Mizuta and Nigel Collier. Zone identification in biology articles as a basis for information extraction. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, JNLPBA '04, pages 29–35, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

Óscar Muñoz García, Andrés García-Silva, Óscar Corcho, Manuel de la Higuera Hernández, and Carlos Navarro. Identifying Topics in Social Media Posts using DBpedia. In *Proceedings of the NEM Summit.* Eurescom, the European Institute for Research and Strategic Studies in Telecommunications, 2011.

Tony Mullen, Yoko Mizuta, and Nigel Collier. A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *SIGKDD Explorations*, 7(1):52–58, 2005.

Ramesh Nallapati, Mihai Surdeanu, and Christopher D. Manning. Blind domain transfer for named entity recognition using generative latent topic models. In *Proceedings of the NIPS 2010 Workshop on Transfer Learning Via Rich Generative Models*, 2010.

Raheel Nawaz, Paul Thompson, John McNaught, and Sophia Ananiadou. Meta-knowledge annotation of bio-events. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Kamal Nigam, Andrew K. Mccallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2): 103–134, May 2000a. ISSN 08856125.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Mach. Learn.*, 39(2-3): 103–134, May 2000b. ISSN 0885-6125.

Pavan Kapanipathi Pablo Mendes, Alexandre Passant and Amit Sheth. Linked open social signals. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.* IEEE Computer Society, 2010.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010.

Sinno Jialin Pan, James T. Kwok, and Qiang Yang 0001. Transfer learning via dimensionality reduction. In Dieter Fox and Carla P. Gomes, editors, *AAAI*, pages 677–682. AAAI Press, 2008. ISBN 978-1-57735-368-3.

Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang 0001. Domain adaptation via transfer component analysis. In Craig Boutilier, editor, *IJCAI*, pages 1187–1192, 2009.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. Using measures of semantic relatedness for word sense disambiguation. In Alexander F. Gelbukh, editor, *CICLing*, Lecture Notes in Computer Science, pages 241–257. Springer, 2003.

Natalia Ponomareva and Mike Thelwall. Biographies or blenders: Which resource is best for cross-domain sentiment analysis? In *Conference on Intelligent Text Processing and Computational Linguistics*, pages 488–499, 2012a.

Natalia Ponomareva and Mike Thelwall. Do neighbours help? an exploration of graph-based algorithms for cross-domain sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 655–665, Jeju Island, Korea, July 2012b. Association for Computational Linguistics.

Piyush Rai, Avishek Saha, Hal Daume, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32, Los Angeles, California, June 2010. Association for Computational Linguistics.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. Self-taught learning: Transfer learning from unlabeled data. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, 2007.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. ACL, 2009.

Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing Microblogs with Topic Models. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. AAAI Press, 2010.

Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In Eric Brill and Kenneth Church, editors, *Proceedings of the Empirical Methods in Natural Language Processing*, pages 133–142, 1996.

Alan L. Rector, Sam Brandt, and Thomas Schneider. Getting the foot out of the pelvis: modeling problems affecting use of snomed ct hierarchies in practical applications. *JAMIA*, 18(4):432–440, 2011.

Roi Reichart and Anna Korhonen. Document and corpus level inference for unsupervised and transductive learning of information structure of scientific documents. In *Proceedings of*

COLING 2012: Posters, pages 995–1006, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. ACL, 2011.

Giuseppe Rizzo and Raphaël Troncy. NERD: evaluating named entity recognition tools in the web of data. In *ISWC 2011, Workshop on Web Scale Knowledge Extraction (WEKEX'11), October 23-27, 2011, Bonn, Germany*, Bonn, GERMANY, 10 2011.

Giuseppe Rizzo and Raphael Troncy. Nerd : a framework for evaluating named entity recognition tools in the web of data. In *Proceedings of the 10th International Semantic Web Conference*. Springer, 2011.

F B Rogers. Medical subject headings. *Bulletin of the Medical Library Association*, 51: 114–6, January 1963. ISSN 0025-7338.

Nicholas Roy and Andrew Mccallum. Toward optimal active learning through sampling estimation of error reduction. In *In Proc. 18th International Conf. on Machine Learning*, pages 441–448. Morgan Kaufmann, 2001.

Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Manfred Hauswirth, JosianeXavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein, and Eva Blomqvist, editors, *The Semantic Web ? ISWC 2012*, volume 7649 of *Lecture Notes in Computer Science*, pages 508–524. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-35175-4.

Manuel Salvadores, Paul R. Alexander, Mark A. Musen, and Natalya Fridman Noy. Bioportal as a dataset of linked biomedical ontologies and terminologies in rdf. *Semantic Web*, 4(3):277–284, 2013.

Yutaka Sasaki, Yoshimasa Tsuruoka, John McNaught, and Sophia Ananiadou. How to make the most of NE dictionaries in statistical NER. *BMC bioinformatics*, 9 Suppl 11, 2008. ISSN 1471-2105.

Sandeepkumar Satpal and Sunita Sarawagi. Domain adaptation of conditional probability models via feature subsetting. In *PKDD*, pages 224–235, 2007.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, 1st edition, December 2001. ISBN 0262194759.

Fabrizio Sebastiani. Text categorization. In *Encyclopedia of Database Technologies and Applications*, pages 683–687. 2005.

Burr Settles. Active learning literature survey. Technical report, Universityof Wisconsin-Madison, 2010.

Burr Settles and Mark Craven. Exploiting zone information, syntactic features, and informative terms in gene ontology annotation from biomedical documents. In *Proceedings of the Text Retrieval Conference (TREC)*, 2005.

Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W. John Wilbur. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093, 2008.

Xiaoxiao Shi, Wei Fan, and Jiangtao Ren. Actively transfer domain knowledge. In *ECML/PKDD (2)*, pages 342–357, 2008.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 2000.

Yongwook Shin, Chuhyeop Ryo, and Jonghun Park. Automatic extraction of persistent topics from social text streams. *World Wide Web*, 2013.

Vikas Sindhwani and Prem Melville. Document-word co-regularization for semi-supervised sentiment analysis. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 1025–1030, Washington, DC, USA, 2008. IEEE Computer Society.

Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. Short text conceptualization using a probabilistic knowledgebase. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. AAAI Press, 2011.

Manfred Stede. *Discourse Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2011.

Ke Tao, Fabian Abel, Claudia Hauff, and Geert-Jan Houben. What makes a tweet relevant for a topic? In *Making Sense of Microposts (#MSM2012)*. CEUR, 2012.

Doina Tatar. Word sense disambiguation by machine learning approach: A short survey. *Fundam. Inf.*, July 2004.

Imad Tbahriti, Christine Chichester, Frederique Lisacek, and Patrick Ruch. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the medline digital library. *International Journal of Medical Informatics*, 75, 2005.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

Simone Teufel, Advaith Siddharthan, and Colin R. Batchelor. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *EMNLP*, pages 1493–1502, 2009.

Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition*. 2009.

Bin Tong, Junbin Gao, Thach Nguyen Huy, Hao Shao, and Einoshin Suzuki. Transfer dimensionality reduction by gaussian process in parallel. *Knowledge and Information Systems*, pages 1–31, 2013. ISSN 0219-1377.

Anusua Trivedi, Piyush Rai, Hal Daumé III, and Scott L. DuVall. Leveraging social bookmarks from partially tagged corpus for improved webpage clustering. In *ACM Transactions on Intelligent Systems and Technology*, 2011.

Vincent Van Asch and Walter Daelemans. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 31–36, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Vladimir Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, 2nd edition, November 1999. ISBN 0387987800.

Daniele Vitale, Paolo Ferragina, and Ugo Scaiella. Classification of short texts by deploying topical annotations. In *Proceedings of the 34th European Conference on IR Research*. Springer, 2012.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1.

Pu Wang, Carlotta Domeniconi, and Jian Hu. Using wikipedia for co-clustering based cross-domain text classification. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, Washington, DC, USA, 2008. IEEE Computer Society.

Yefeng Wang. Annotating and recognising named entities in clinical notes. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pages 18–26, Suntec, Singapore, August 2009. Association for Computational Linguistics.

Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 486–497. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-24523-0.

James W. Woods, Charles Sneiderman, Kamran Hameed, Michael J. Ackerman, and Charlie Hatton. Using umls metathesaurus concepts to describe medical images: dermatology vocabulary. *Comp. in Bio. and Med.*, 36:89–100, 2006.

Dan Wu, Wee Sun Lee, Nan Ye, and Hai Leong Chieu. Domain adaptive bootstrapping for named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1523–1532, Singapore, August 2009a. Association for Computational Linguistics.

Qiong Wu, Songbo Tan, and Xueqi Cheng. Graph ranking for sentiment transfer. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 317–320, Suntec, Singapore, August 2009b. Association for Computational Linguistics.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 2012.

Zhibiao Wu and Martha Stone Palmer. Verb semantics and lexical selection. In *ACL*, pages 133–138, 1994.

Evan Wei Xiang, Bin Cao, Derek Hao Hu, and Qiang Yang. Bridging domains using world wide knowledge for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010.

Min Xiao and Yuhong Guo. Online active learning for cost sensitive domain adaptation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 1–9, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Tan Xu and Douglas W. Oard. Wikipedia-based topic clustering for microblogs. *Proceedings of the American Society for Information Science and Technology*, 2011.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *ICML*, 2004.

Ziqi Zhang. *Named Entity Recognition: Challenges in Document Annotation, Gazetteer Construction and Disambiguation*. PhD thesis, University of Sheffield, 2013.

Ziqi Zhang, Sam Chapman, and Fabio Ciravegna. A methodology towards effective and efficient manual document annotation: Addressing annotator discrepancy and annotation quality. In *EKAW*, pages 301–315, 2010.

Ziqi Zhang, Gentile Annalisa, Isabelle Augenstein, Eva Blomqvist, and Fabio Ciravegna. Mining equivalent relations from linked data. In *ACL*, 2013.

Wayne X. Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee P. Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European conference on Advances in information retrieval*. Springer-Verlag, 2011.

Xiaojin Zhu. Semi-supervised learning literature survey, 2006.

Xiaojin Zhu. Semi-Supervised Learning Literature Survey. Technical report, Computer Sciences, University of Wisconsin-Madison, 2005.

Appendices

Appendix A

Probabilistic Graphical Models

A.1 Background on Probabilistic Graphical Models

Probabilistic graphical models provide a powerful framework for content modelling which combines uncertainty (*probabilities*) and logical structure (*independence constraints*) to compactly encode and manipulate high dimensional distributions common to many text classification (TC) tasks. By recognising patterns of word use and grouping documents that are similar, probabilistic graphical models have emerged as a powerful technique for discovering useful structures in documents. One of the simplest and most widely used graphical model for content modelling was introduced by Blei et al. [2003a]: Latent Dirichlet Allocation (LDA) provides a new, semantically consistent model for representing the topics of a document. This model has since been extended and successfully applied to many other tasks, such as *sentiment analysis* [Lin and He, 2009], *topic modelling* [Zhao et al., 2011] and *entity resolution* [Dai and Storkey, 2011][1]. It is also widely accepted [Blei et al., 2003a] that *mixture models* are superior compared to simple clustering models (e.g. k-means) for modelling text documents, especially long documents such as research papers.

The basic idea behind these models is to efficiently represent the text as a *joint probability distribution* $P(\mathcal{X})$ over a set of *random variables* ($\mathcal{X} = (x_1, \ldots, x_n)$) (e.g. word), and to provide an elegant way of representing the interactions between these variables using a graphical structure (e.g. words belonging to different topics, or zone types). An illustration of the graphical representation of LDA using *plate notations* is given in Figure A.1, where each *node* refers to a random variable (the *hidden or unobservable variable* being shaded, and the *observable variable* being unshaded), each *edge* represents the direct probabilistic interactions between the nodes, and the replications of nodes are represented by boxes, called *plates*. For instance, the outer plate represents documents, while the inner plate denotes the repeated choices of words and topics within a document.

The underlying assumption behind these models is that the data is generated according to a generative process, which defines the *joint probability distribution* over both the observed and hidden random variables. That is, the documents come from a generative process that includes T hidden topics, allowing the documents to be represented as a probability distribution over hidden topics (θ), while the topics are represented as a probability

[1] A more complete list of applications is provided at: http://www.cs.princeton.edu/~mimno/topics.html.

distribution over words (ϕ).

The key strengths of these models are the *conditional independence* assumption and *exchangeability*. The *conditional independence* assumption ensures that the variables inside a plate are conditionally independent of all the other variables. For instance, the number of words in a document (N_w) is independent on all the other variables (t hidden topics, θ document-topic distributions). The *exchangeability* property holds for a sequence of random variables whose joint probability distribution is invariant to any permutation. This allows the words in the documents to be interchangeable with one another, which can be particularly useful when dealing with multi-genre text in large repositories.

Following the topic modelling terminology [Blei et al., 2003a], the following definitions can be introduced:

- T denotes the number of topics
- V denotes the size of vocabulary in the corpus or domain
- α is a positive vector, consisting of the *hyper-parameters* of the Dirichlet distribution ($Dir(\alpha)$)
- $Dir(\alpha)$ is a V-dimensional Dirichlet vector with parameter vector α
- β is a scalar, consisting of the *hyper-parameters* of the Dirichlet distribution ($Dir(\beta)$)
- $Dir(\beta)$ is a T-dimensional *symmetric* Dirichlet with scalar parameter β (each component of the parameter having the same value)
- θ^d represents the per-document topic proportions (topic mixture proportion)
- $t_{d,n}$ refers to the topic index
- $\phi^{t,w}$ represents the per-word topic assignments (topic mixture component)
- $w_{d,n}$ denotes the term indicator for the nth word in document d.

In light of these notations, the generative story for creating the data can be described as follows:

Algorithm 6 Generative process of original LDA model. T denotes the number of topics, N_d denotes the number of documents, α refers to a vector for Dirichlet prior for the document zone distributions, θ^d refers to the document zone distribution for document d, $t_{d,n}$ stands for the topic index for the word at the position n in document d, $w_{d,n}$ denotes the word at the position n in document d, β refers to the word probability vectors as $Z \times V$ for the Dirichlet prior for each zone.

1: **for all** topics $t = \{1, \ldots, T\}$ **do**
2: **draw** mixture components $\phi^t \sim \text{Dir}(\beta)$
3: **for all** document $d = \{1, \ldots, N_d\}$ **do**
4: **draw** mixture proportion $\theta^d \sim \text{Dir}(\alpha)$
5: **for all** word $w_{d,n}$ **do**
6: **draw** topic index $t_{d,n} \sim \text{Multinomial}(\theta^d)$
7: **draw** term for word $w_{d,n} \sim \text{Multinomial}(\phi^{t_{d,n}})$

As shown in Algorithm 6, the corpus-level parameters α, β are sampled once; the document level parameter θ^d is sampled once per document, and the $t_{d,n}$ and $w_{d,n}$ word-level parameters are sampled once for each word in the document.

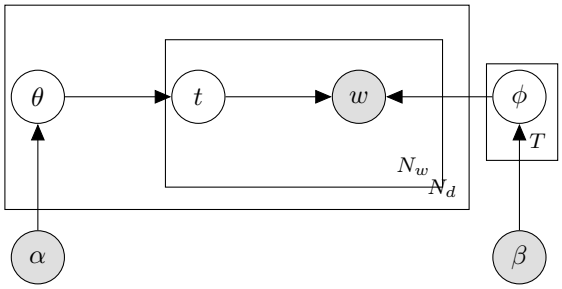

Figure A.1: Graphical representation of the Latent Dirichlet Allocation probabilistic graphical model: shaded nodes representing observed variables, and unshaded nodes referring to hidden variables, boxes (plates) representing replications of nodes, w corresponds to words, t refers to topics, α, β are hyperparameters, θ stands for the document-topic distribution, and Θ refers to the topic-word distribution.

During the training phase, the model is provided with a set of *unlabelled* training documents from which the posterior distribution of the hidden variable t needs to be computed. Given the complex nature of LDA model, however, this posterior distribution ($P(t_{(d,i)} = k|t_{-(d,i)}, w)$), the probability of assigning topic k to the word at position i in document d, given the word w), is intractable. For this reason an approximate inference algorithm called Gibbs sampling [Griffiths, 2002] is generally employed, which uses Markov chain, that upon convergence, approximately generates samples according to the posterior distribution [Wallach et al., 2009]. Using Gibbs sampling, the posterior distribution of topic mixtures, can be computed as follows [Griffiths and Steyvers, 2004]:

$$P(t_{(d,i)} = k|t_{-(d,i)}, w) \propto \frac{n^k_{d,-i} + \alpha_k}{n_{d,-i} + T\alpha_k} \frac{n^k_{v,-i} + \beta}{n^k_{\cdot,-i} + V\beta},$$

where $n^k_{d,-i}$ denotes the number of times topic k is assigned to document d, $n_{d,-i}$ denotes the total number of topics assigned to document d, $n^k_{v,-i}$ refers to the total number of times topic k is assigned to word v, and $n^k_{\cdot,-i}$ denotes the total number words assigned to topic k. (The notation $-i$ means that when computing these counts, the current topic assignment is disregarded.) During testing, the posterior distribution is used to assign topics on previously unseen (held-out) documents.

The core distribution behind the LDA model is the Dirichlet distribution, which is a member of the exponential family, providing useful properties for inference. One of these important properties is that the distribution has a conjugate prior, which is the multinomial distribution. And this conjugate prior can be used for factorising the posterior distribution.

The values of the Dirichlet parameter can significantly influence the performance of the LDA model. For instance, smaller values of α, β will result in sparser distributions for θ, ϕ. By controlling the document topic proportion θ to be sparse, the model prefers representing the documents by fewer topics. By having sparse distribution for the mixture components ϕ, the model tries to assign few terms to each topic. Very often a *symmetric* Dirichlet prior is employed in LDA, which assumes that all topics have the same chance of being assigned to a document and all words (frequent and infrequent ones). Alternatively, an *asymmetric*

Dirichlet prior can be employed, in which case only specific topics will have the chance of being assigned to each document (those with high α values), and also only specific words will have the same chance of being assigned to a topic (those with high β values). These values are typically set and evaluated experimentally.

As previously mentioned, LDA-based graphical models have gained popularity in many TC tasks such as sentiment analysis or topic modelling. However, the application of graphical models to other TC tasks such as document zoning has not yet been studied. In order to address this limitation, this thesis proposes two variants of the LDA model for the task of *within-document* zoning. This allows flexible modelling of the domain documents as a mixture of zone distributions, at the same time ignoring the order of the sentences in the documents.

Appendix B

Additional Experimental Results on Adaptive Document Zoning

This appendix contains additional experimental results obtained for the adaptive document zone classifiers introduced in Section 5.3.

B.1 Results Obtained using Semantic Class Features

B.1.1 Single-domain Scenario

This section presents the precision and recall values obtained for the single-domain (TGT_ONLY) classifier using semantic class (**Cls**) features. Table B.1 shows the results obtained in terms of precision, while Table B.2 shows the results obtained in terms of recall.

TGT Domain	Semantic				Baseline	
	sct+*msh* CCA	*sct*+*msh*	*sct*	*msh*	*BoW*	*BoE*
Biol	**0.727**	0.710	0.718	0.716	0.700	**0.704**
CellBiol	**0.823**	0.806	0.812	0.812	0.795	**0.801**
Communi	**0.666**	0.635	0.647	0.647	0.635	**0.640**
HealthS	**0.631**	0.598	0.611	0.612	**0.605**	0.599
Medicin	**0.713**	0.671	0.697	0.696	0.683	**0.693**
PublicH	**0.652**	0.624	0.635	0.639	0.616	**0.625**
Tropica	**0.657**	0.621	0.637	0.640	0.626	**0.630**

Table B.1: Precision results for the SVM TGT_ONLY classifier using semantic class (**Cls**) features extracted from two KS ontologies (*sct* and *msh*) and various baseline lexical features (BoW, BoE).

B.1.2 Cross-domain Scenario

This section presents the precision and recall values obtained for the OntoEA, SRC_TGT, EA and SRC_ONLY cross domain classifiers. The results for the precision are shown in Table B.3, while those for the recall are shown in Table B.4.

TGT Domain	Semantic				Baseline	
	sct+msh CCA	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*
Biol	**0.713**	0.698	0.703	0.701	0.700	**0.704**
CellBiol	**0.811**	0.797	0.800	0.801	0.795	**0.801**
Communi	**0.663**	0.628	0.642	0.642	0.635	**0.640**
HealthS	**0.617**	0.592	0.598	0.598	**0.605**	0.599
Medicin	**0.711**	0.674	0.693	0.693	0.683	**0.693**
PublicH	**0.641**	0.617	0.623	0.628	0.616	**0.625**
Tropica	**0.650**	0.620	0.630	0.630	0.626	**0.630**

Table B.2: Recall results for the SVM TGT_ONLY classifier using semantic class (**Cls**) features extracted from two KS ontologies (*sct* and *msh*) and various baseline lexical features (BoW, BoE).

B.2 Results Obtained using Semantic Upper-Class Features

B.2.1 Single-domain Scenario

This section summarises the results obtained for the single-domain (TGT_ONLY) classifier and the cross-domain classifiers using upper-class (**parent(Cls)**) features. Table B.5 shows the results obtained for the TGT_ONLY classifier in terms of F1 measure. Table B.6 presents the results in terms of precision. Table B.7 shows the results obtained in terms of recall. As it can be observed, the TGT_ONLY ($sct{+}msh_{CCA}$) performs the best overall results.

B.2.2 Cross-domain Scenario

This subsection presents the results obtained for the cross-domain classifiers using upper-class features (**parent(Cls)**). The results are displayed on Table B.8 for F1, on Table B.9 for precision and on Table B.10 for recall.

B.3 Results for the Adaptive Document Zone Classifiers

This section presents additional learning curves for the OntoEA, SRC_TGT, TGT_ONLY and EA classifiers in terms of F1 measure for all the seven biomedical sub-domains analysed. The results obtained for the Biology source domain are shown in Figure B.1, for the Cell Biology are shown in Figure B.2, for the Communicable Disease are presented in Figure B.3, for the Health Services are displayed in Figure B.4, for the Medicine are shown in Figure B.5 Public Health are shown in Figure B.6, for Tropical Medicine are displayed on Figure B.7.

SRC Domain	TGT Domain	OntoEA				SRC_TGT						EA		SRC_ONLY					
		sct+msh CCA	*sct+msh*	*sct*	*msh*	*sct+msh CCA*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*	*BoW*	*BoE*	*sct+msh CCA*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*
Biol	*CellBiol*	**0.836**	0.798	0.804	0.805	**0.807**	0.792	0.806	0.805	0.724	**0.795**	0.796	**0.807**	**0.773**	0.749	0.750	0.754	0.348	**0.750**
	Communi	**0.696**	0.653	0.654	0.657	**0.653**	0.636	0.646	0.647	0.530	**0.645**	0.612	**0.648**	**0.552**	0.519	0.524	0.527	0.104	**0.525**
	HealthS	**0.653**	0.605	0.617	0.619	**0.608**	0.584	0.605	0.607	0.526	**0.609**	0.574	**0.576**	**0.438**	0.427	0.412	0.413	0.111	**0.411**
	Medicin	**0.765**	0.701	0.729	0.727	**0.714**	0.671	0.705	0.705	0.567	**0.718**	0.677	**0.707**	**0.648**	0.588	0.615	0.615	0.133	**0.614**
	PublicH	**0.678**	0.609	0.643	0.640	**0.639**	0.609	0.629	0.631	0.558	**0.631**	0.598	**0.633**	**0.489**	0.461	0.468	0.465	0.108	**0.468**
	Tropica	**0.694**	0.626	0.661	0.659	**0.652**	0.609	0.644	0.645	0.498	**0.646**	0.607	**0.644**	**0.557**	0.536	0.533	0.528	0.097	**0.535**
CellBiol	*Biol*	**0.777**	0.740	0.747	0.748	**0.714**	0.694	0.713	0.714	0.667	**0.737**	0.707	**0.712**	**0.683**	0.647	0.661	0.664	0.413	**0.662**
	Communi	**0.691**	0.663	0.652	0.652	**0.651**	0.621	0.643	0.645	0.609	**0.644**	0.621	**0.644**	**0.457**	0.438	0.433	0.432	0.328	**0.433**
	HealthS	**0.653**	0.600	0.619	0.620	**0.609**	0.577	0.603	0.603	0.590	**0.608**	0.587	**0.604**	**0.375**	0.354	0.354	0.359	0.270	**0.352**
	Medicin	**0.751**	0.702	0.715	0.715	**0.710**	0.658	0.702	0.705	0.670	**0.703**	0.686	**0.701**	**0.558**	0.510	0.529	0.529	0.387	**0.528**
	PublicH	**0.684**	0.621	0.648	0.649	**0.633**	0.609	0.627	0.632	0.613	**0.639**	0.604	**0.628**	**0.398**	0.381	0.375	0.376	0.280	**0.372**
	Tropica	**0.699**	0.644	0.660	0.658	**0.650**	0.602	0.641	0.642	0.606	**0.646**	0.618	**0.638**	**0.473**	0.447	0.444	0.444	0.336	**0.443**
Communi	*Biol*	**0.759**	0.690	0.729	0.729	**0.720**	0.694	0.716	0.718	0.650	**0.719**	0.706	**0.717**	**0.624**	0.579	0.598	0.604	0.091	**0.598**
	CellBiol	**0.826**	0.764	0.762	0.764	**0.810**	0.761	0.771	0.773	0.739	**0.750**	0.775	**0.772**	**0.654**	0.622	0.631	0.637	0.386	**0.632**
	HealthS	**0.681**	0.626	0.645	0.642	**0.616**	0.583	0.607	0.609	0.624	**0.635**	0.604	**0.608**	**0.557**	0.512	0.528	0.530	0.552	**0.527**
	Medicin	**0.767**	0.715	0.731	0.728	**0.705**	0.665	0.700	0.699	0.715	**0.722**	0.693	**0.700**	**0.661**	0.597	0.622	0.624	0.609	**0.622**
	PublicH	**0.701**	0.650	0.663	0.665	**0.639**	0.612	0.633	0.632	0.646	**0.655**	0.621	**0.635**	**0.620**	0.563	0.590	0.591	0.590	**0.590**
	Tropica	**0.712**	0.660	0.668	0.667	**0.650**	0.616	0.640	0.641	0.593	**0.662**	0.628	**0.640**	**0.653**	0.613	0.631	0.630	0.464	**0.630**
HealthS	*Biol*	**0.730**	0.669	0.682	0.681	**0.665**	0.661	0.663	0.664	0.637	**0.672**	0.673	**0.686**	**0.543**	0.510	0.515	0.515	0.127	**0.518**
	CellBiology	**0.824**	0.742	0.754	0.760	**0.774**	0.761	0.773	0.774	**0.754**	0.746	0.774	**0.773**	**0.512**	0.500	0.486	0.490	0.412	**0.488**
	Communi	**0.705**	0.651	0.664	0.667	**0.661**	0.625	0.645	0.643	**0.658**	0.651	**0.650**	0.646	**0.632**	0.574	0.596	0.602	0.572	**0.595**
	Medicin	**0.743**	0.682	0.699	0.701	**0.702**	0.664	0.696	0.693	**0.691**	0.689	0.690	**0.695**	**0.594**	0.524	0.548	0.552	0.536	**0.550**
	PublicH	**0.711**	0.652	0.672	0.675	**0.640**	0.612	0.637	0.633	0.658	**0.661**	0.623	**0.636**	**0.647**	0.604	0.617	0.618	0.615	**0.619**
	Tropica	**0.694**	0.648	0.655	0.655	**0.652**	0.612	0.642	0.642	0.599	**0.645**	0.629	**0.641**	**0.609**	0.567	0.581	0.577	0.448	**0.580**
Medicin	*Biol*	**0.774**	0.726	0.740	0.740	**0.720**	0.693	0.718	0.718	0.616	**0.731**	0.704	**0.718**	**0.711**	0.642	0.662	0.661	0.093	**0.662**
	CellBiol	**0.839**	0.803	0.804	0.805	**0.810**	0.758	0.807	0.810	0.774	**0.793**	0.801	**0.809**	**0.725**	0.666	0.683	0.684	0.434	**0.680**
	Communi	**0.723**	0.677	0.685	0.681	**0.663**	0.634	0.655	0.657	0.668	**0.674**	0.643	**0.655**	**0.655**	0.600	0.627	0.625	0.625	**0.625**
	HealthS	**0.679**	0.621	0.636	0.639	**0.617**	0.581	0.609	0.610	0.615	**0.625**	0.599	**0.608**	**0.548**	0.503	0.520	0.519	0.528	**0.519**
	PublicH	**0.691**	0.643	0.651	0.651	**0.637**	0.605	0.631	0.629	**0.646**	0.640	0.621	**0.632**	**0.609**	0.561	0.583	0.584	0.571	**0.580**
	Tropica	**0.701**	0.659	0.660	0.656	**0.646**	0.612	0.639	0.640	0.602	**0.652**	0.631	**0.638**	**0.648**	0.599	0.616	0.609	0.472	**0.611**
PublicH	*Biol*	**0.731**	0.682	0.687	0.687	**0.685**	0.664	0.684	0.682	0.624	**0.673**	0.699	**0.684**	**0.560**	0.523	0.536	0.539	0.135	**0.539**
	CellBiol	**0.824**	0.734	0.752	0.754	**0.773**	0.760	0.772	0.773	0.737	**0.743**	0.776	**0.772**	**0.561**	0.529	0.542	0.542	0.404	**0.542**
	Communi	**0.710**	0.658	0.666	0.668	**0.660**	0.625	0.652	0.650	0.665	**0.654**	0.652	**0.648**	**0.647**	0.590	0.616	0.612	0.628	**0.614**
	HealthS	**0.691**	0.640	0.657	0.653	**0.625**	0.592	0.614	0.615	0.640	**0.647**	0.610	**0.613**	**0.634**	0.581	0.610	0.606	0.618	**0.610**
	Medicin	**0.751**	0.691	0.710	0.711	**0.701**	0.660	0.695	0.692	0.696	**0.701**	0.691	**0.694**	**0.607**	0.553	0.571	0.565	0.570	**0.566**
	Tropica	**0.696**	0.650	0.657	0.655	**0.651**	0.610	0.638	0.641	0.599	**0.647**	0.630	**0.648**	**0.625**	0.572	0.596	0.595	0.450	**0.593**
Tropica	*Biol*	**0.761**	0.685	0.732	0.728	**0.721**	0.696	0.714	0.717	0.643	**0.726**	0.706	**0.715**	**0.640**	0.603	0.615	0.622	0.102	**0.617**
	CellBiol	**0.837**	0.765	0.761	0.759	**0.808**	0.759	0.807	0.807	0.779	**0.752**	0.770	**0.807**	**0.658**	0.630	0.634	0.638	0.414	**0.637**
	Communi	**0.723**	0.660	0.685	0.685	**0.664**	0.625	0.657	0.657	0.598	**0.676**	0.639	**0.657**	**0.658**	0.604	0.625	0.631	0.468	**0.629**
	HealthS	**0.687**	0.621	0.647	0.648	**0.617**	0.587	0.610	0.612	0.588	**0.639**	0.602	**0.609**	**0.560**	0.520	0.533	0.529	0.411	**0.533**
	Medicin	**0.766**	0.705	0.724	0.725	**0.711**	0.671	0.699	0.701	0.648	**0.715**	0.679	**0.699**	**0.642**	0.577	0.604	0.602	0.415	**0.605**
	PublicH	**0.698**	0.642	0.662	0.657	**0.639**	0.615	0.633	0.633	0.603	**0.652**	0.616	**0.632**	**0.595**	0.551	0.574	0.570	0.434	**0.574**

Table B.3: Precision results for the OntoEA, SRC_TGT, EA and SRC_ONLY models using semantic class (**Cls**) features from two KS ontologies: *sct* and *msh*.

SRC Domain	TGT Domain	OntoEA				SRC_TGT						EA		SRC_ONLY					
		sct+msh CCA	*sct+msh*	*sct*	*msh*	*sct+msh CCA*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*	*BoW*	*BoE*	*sct+msh CCA*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*
Biol	*CellBiol*	**0.825**	0.786	0.790	0.792	**0.778**	0.782	0.795	0.761	0.680	**0.781**	0.782	**0.796**	**0.755**	0.729	0.731	0.734	0.367	**0.731**
	Communi	**0.685**	0.630	0.641	0.644	**0.636**	0.608	0.628	0.628	0.511	**0.631**	0.599	**0.630**	**0.533**	0.475	0.509	0.509	0.092	**0.510**
	HealthS	**0.625**	0.572	0.584	0.585	**0.587**	0.548	0.567	0.568	0.499	**0.577**	0.546	**0.606**	**0.420**	0.380	0.393	0.397	0.100	**0.393**
	Medicin	**0.759**	0.692	0.720	0.720	**0.705**	0.659	0.694	0.694	0.538	**0.710**	0.672	**0.696**	**0.650**	0.579	0.618	0.618	0.102	**0.618**
	PublicH	**0.655**	0.583	0.617	0.615	**0.615**	0.580	0.603	0.604	0.540	**0.605**	0.579	**0.606**	**0.468**	0.409	0.444	0.443	0.102	**0.444**
	Tropica	**0.679**	0.608	0.644	0.641	**0.632**	0.589	0.620	0.619	0.481	**0.628**	0.599	**0.619**	**0.547**	0.499	0.528	0.520	0.080	**0.530**
CellBiol	*Biol*	**0.766**	0.731	0.736	0.736	**0.698**	0.684	0.696	0.698	0.644	**0.726**	0.689	**0.696**	**0.643**	0.625	0.621	0.623	0.314	**0.622**
	Communi	**0.677**	0.639	0.635	0.635	**0.635**	0.612	0.624	0.626	0.593	**0.626**	0.606	**0.625**	**0.451**	0.418	0.429	0.429	0.351	**0.429**
	HealthS	**0.623**	0.593	0.590	0.590	**0.575**	0.568	0.566	0.565	0.569	**0.579**	**0.573**	0.566	**0.372**	0.344	0.347	0.351	0.309	**0.346**
	Medicin	**0.746**	0.694	0.708	0.709	**0.701**	0.660	0.691	0.694	0.662	**0.697**	0.679	**0.689**	**0.568**	0.523	0.540	0.542	0.397	**0.540**
	PublicH	**0.662**	0.600	0.621	0.622	**0.611**	0.599	0.601	0.604	0.597	**0.612**	0.593	**0.602**	**0.408**	0.373	0.383	0.384	0.327	**0.384**
	Tropica	**0.682**	0.628	0.640	0.638	**0.627**	0.598	0.617	0.616	0.603	**0.626**	0.610	**0.614**	**0.473**	0.441	0.448	0.447	0.362	**0.447**
Communi	*Biol*	**0.743**	0.700	0.713	0.712	**0.704**	0.688	0.699	0.702	0.624	**0.703**	0.691	**0.700**	**0.567**	0.541	0.536	0.542	0.063	**0.535**
	CellBiol	**0.834**	0.778	0.770	0.773	**0.802**	0.776	0.783	0.787	0.732	**0.759**	0.779	**0.784**	**0.610**	0.582	0.574	0.585	0.393	**0.575**
	HealthS	**0.673**	0.626	0.636	0.634	**0.599**	0.576	0.591	0.591	0.622	**0.627**	**0.597**	0.592	**0.537**	0.511	0.512	0.513	**0.544**	0.511
	Medicin	**0.765**	0.717	0.728	0.726	**0.703**	0.668	0.698	0.697	0.714	**0.719**	0.693	**0.698**	**0.653**	0.604	0.617	0.621	0.616	**0.618**
	PublicH	**0.694**	0.647	0.657	0.659	**0.628**	0.605	0.621	0.619	0.641	**0.649**	0.611	**0.622**	**0.610**	0.564	0.581	0.583	**0.584**	0.582
	Tropica	**0.709**	0.659	0.664	0.662	**0.644**	0.613	0.633	0.633	0.579	**0.657**	0.627	**0.632**	**0.635**	0.605	0.612	0.612	0.445	**0.611**
HealthS	*Biol*	**0.734**	0.684	0.686	0.686	**0.681**	0.676	0.660	0.664	0.613	**0.677**	0.659	**0.688**	**0.509**	0.498	0.487	0.488	0.082	**0.490**
	CellBiol	**0.836**	0.757	0.767	0.772	**0.784**	0.772	0.783	0.784	0.752	**0.758**	0.775	**0.784**	**0.511**	0.495	0.483	0.487	0.324	**0.485**
	Communi	**0.704**	0.644	0.662	0.664	**0.658**	0.617	0.641	0.638	**0.656**	**0.649**	**0.644**	0.641	**0.630**	0.572	0.599	0.604	0.573	**0.599**
	Medicin	**0.743**	0.686	0.699	0.700	**0.701**	0.668	0.693	0.691	**0.691**	0.688	0.690	**0.693**	**0.597**	0.546	0.562	0.566	0.551	**0.563**
	PublicH	**0.706**	0.648	0.667	0.672	**0.760**	0.606	0.625	0.621	0.654	**0.656**	0.613	**0.624**	**0.644**	0.602	0.618	0.621	0.613	**0.621**
	Tropica	**0.691**	0.646	0.651	0.650	**0.647**	0.611	0.636	0.635	0.580	**0.641**	0.627	**0.634**	**0.617**	0.578	0.591	0.590	0.430	**0.591**
Medicin	*Biol*	**0.758**	0.711	0.723	0.722	**0.704**	0.683	0.700	0.700	0.585	**0.714**	0.690	**0.701**	**0.654**	0.616	0.614	0.612	0.068	**0.613**
	CellBiol	**0.821**	0.790	0.786	0.786	**0.798**	0.773	0.794	0.798	0.743	**0.775**	0.789	**0.797**	**0.637**	0.611	0.595	0.600	0.409	**0.596**
	Communi	**0.721**	0.671	0.680	0.677	**0.657**	0.627	0.649	0.652	0.666	**0.670**	0.639	**0.649**	**0.651**	0.594	0.627	0.625	0.623	**0.625**
	HealthS	**0.670**	0.615	0.626	0.629	**0.600**	0.574	0.593	0.593	0.607	**0.616**	0.591	**0.593**	**0.498**	0.484	0.480	0.479	**0.500**	0.478
	PublicH	**0.679**	0.635	0.637	0.637	**0.625**	0.597	0.618	0.615	**0.637**	0.627	0.611	**0.620**	**0.581**	0.543	0.558	0.558	0.546	**0.554**
	Tropica	**0.695**	0.657	0.654	0.650	**0.638**	0.611	0.631	0.630	0.589	**0.645**	**0.630**	0.629	**0.629**	0.593	0.604	0.598	0.438	**0.599**
PublicH	*Biol*	**0.733**	0.697	0.688	0.689	**0.683**	0.678	0.683	0.681	0.601	**0.674**	0.683	**0.683**	**0.522**	0.505	0.498	0.503	0.078	**0.498**
	CellBiol	**0.834**	0.750	0.762	0.764	**0.783**	0.771	0.772	0.783	0.731	**0.753**	0.777	**0.782**	**0.547**	0.522	0.526	0.527	0.375	**0.525**
	Communi	**0.708**	0.654	0.661	0.664	**0.656**	0.619	0.648	0.646	**0.664**	0.648	**0.646**	0.644	**0.653**	0.600	0.626	0.623	**0.636**	0.623
	HealthS	**0.685**	0.640	0.650	0.646	**0.609**	0.585	0.598	0.597	0.637	**0.639**	0.604	**0.597**	**0.622**	0.581	0.600	0.596	0.617	**0.598**
	Medicin	**0.750**	0.695	0.709	0.709	**0.699**	0.663	0.692	0.690	0.696	**0.700**	0.691	**0.691**	**0.604**	0.567	0.576	0.573	**0.582**	0.573
	Tropica	**0.693**	0.649	0.652	0.651	**0.646**	0.610	0.631	0.634	0.583	**0.643**	0.628	**0.641**	**0.625**	0.581	0.598	0.597	0.453	**0.595**
Tropica	*Biol*	**0.745**	0.696	0.717	0.712	**0.706**	0.687	0.698	0.700	0.600	**0.711**	0.687	**0.699**	**0.611**	0.593	0.585	0.594	0.035	**0.586**
	CellBiol	**0.824**	0.780	0.770	0.767	**0.797**	0.771	0.796	0.797	0.752	**0.762**	0.777	**0.795**	**0.641**	0.612	0.603	0.610	0.375	**0.606**
	Communi	**0.724**	0.657	0.684	0.684	**0.660**	0.619	0.651	0.651	0.565	**0.675**	0.631	**0.651**	**0.667**	0.606	0.635	0.640	0.451	**0.639**
	HealthS	**0.679**	0.619	0.641	0.641	**0.599**	0.580	0.595	0.595	0.563	**0.633**	0.591	**0.593**	**0.560**	0.527	0.535	0.532	0.386	**0.537**
	Medicin	**0.765**	0.708	0.722	0.724	**0.710**	0.674	0.697	0.699	0.635	**0.713**	0.678	**0.697**	**0.653**	0.598	0.620	0.620	0.435	**0.621**
	PublicH	**0.692**	0.639	0.653	0.650	**0.628**	0.606	0.620	0.619	0.577	**0.644**	0.605	**0.619**	**0.601**	0.554	0.581	0.577	0.395	**0.581**

Table B.4: Recall results for the OntoEA, SRC_TGT, EA and SRC_ONLY models using semantic class (**Cls**) features from two KS ontologies: *sct* and *msh*.

TGT Domain	Semantic				Baseline	
	sct+msh $_{CCA}$	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*
Biol	**0.717**	0.700	0.706	0.703	0.705	**0.709**
CellBiol	**0.814**	0.799	0.805	0.804	0.799	**0.805**
Communi	**0.661**	0.630	0.646	0.645	0.636	**0.641**
HealthS	**0.624**	0.595	0.604	0.606	**0.606**	0.604
Medicin	**0.711**	0.672	0.689	0.695	0.683	**0.693**
PublicH	**0.641**	0.621	0.631	0.631	0.619	**0.629**
Tropica	**0.655**	0.617	0.632	0.633	0.625	**0.633**

Table B.5: F1 results for the SVM TGT_ONLY classifier using upper-class (**parent(Cls)**) features extracted from two KS ontologies (*sct* and *msh*) and various baseline lexical features (BoW, BoE).

TGT Domain	Semantic				Baseline	
	sct+msh $_{CCA}$	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*
Biol	**0.718**	0.700	0.708	0.706	0.700	**0.704**
CellBiol	**0.815**	0.800	0.804	0.804	0.795	**0.801**
Communi	**0.663**	0.629	0.643	0.643	0.635	**0.640**
HealthS	**0.622**	0.592	0.603	0.603	**0.605**	0.599
Medicin	**0.711**	0.671	0.694	0.693	0.683	**0.693**
PublicH	**0.644**	0.619	0.627	0.632	0.616	**0.625**
Tropica	**0.653**	0.619	0.633	0.634	0.626	**0.630**

Table B.6: Precision results for the SVM TGT_ONLY classifier using upper-class (**parent(Cls)**) features extracted from two KS ontologies (*sct* and *msh*) and various baseline lexical features (BoW, BoE).

TGT Domain	Semantic				Baseline	
	sct+msh $_{CCA}$	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*
Biol	**0.713**	0.698	0.700	0.698	0.700	**0.704**
CellBiol	**0.810**	0.797	0.801	0.800	0.795	**0.801**
Communi	**0.660**	0.628	0.645	0.644	0.635	**0.640**
HealthS	**0.619**	0.595	0.600	0.600	**0.605**	0.599
Medicin	**0.711**	0.675	0.689	0.695	0.683	**0.693**
PublicH	**0.637**	0.619	0.626	0.627	0.616	**0.625**
Tropica	**0.652**	0.618	0.629	0.629	0.626	**0.630**

Table B.7: Precision results for the SVM TGT_ONLY classifier using upper-class (**parent(Cls)**) features extracted from two KS ontologies (*sct* and *msh*) and various baseline lexical features (BoW, BoE).

SRC Domain	TGT Domain	OntoEA				SRC_TGT						EA		SRC_ONLY					
		sct+msh CCA	*sct+msh*	*sct*	*msh*	*sct+msh CCA*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*	*BoW*	*BoE*	*sct+msh CCA*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*
Biol	*CellBiol*	**0.822**	0.783	0.789	0.790	**0.782**	0.784	0.793	0.785	0.695	**0.786**	0.787	**0.800**	**0.755**	0.727	0.731	0.734	0.348	**0.731**
	Communi	**0.687**	0.636	0.644	0.646	**0.640**	0.618	0.632	0.634	0.516	**0.636**	0.602	**0.637**	**0.522**	0.463	0.495	0.496	0.089	**0.496**
	HealthS	**0.629**	0.578	0.591	0.592	**0.589**	0.561	0.576	0.577	0.507	**0.588**	0.555	**0.598**	**0.420**	0.384	0.393	0.396	0.090	**0.392**
	Medicin	**0.755**	0.688	0.718	0.716	**0.702**	0.662	0.692	0.692	0.546	**0.712**	0.672	**0.699**	**0.635**	0.564	0.602	0.602	0.109	**0.602**
	PublicH	**0.656**	0.585	0.620	0.618	**0.617**	0.591	0.606	0.608	0.545	**0.614**	0.585	**0.616**	**0.461**	0.403	0.437	0.435	0.092	**0.437**
	Tropica	**0.684**	0.613	0.650	0.648	**0.639**	0.595	0.629	0.628	0.485	**0.635**	0.599	**0.628**	**0.543**	0.494	0.521	0.514	0.072	**0.524**
CellBiol	*Biol*	**0.770**	0.733	0.740	0.740	**0.704**	0.685	0.702	0.703	0.652	**0.730**	0.694	**0.702**	**0.644**	0.623	0.622	0.624	0.352	**0.623**
	Communi	**0.673**	0.639	0.632	0.633	**0.633**	0.614	0.623	0.625	0.599	**0.632**	0.610	**0.632**	**0.416**	0.384	0.394	0.393	0.321	**0.393**
	HealthS	**0.626**	0.587	0.594	0.593	**0.580**	0.570	0.573	0.572	0.577	**0.589**	0.578	**0.580**	**0.343**	0.326	0.320	0.322	0.274	**0.319**
	Medicin	**0.743**	0.690	0.705	0.706	**0.699**	0.658	0.690	0.693	0.665	**0.698**	0.679	**0.692**	**0.538**	0.490	0.508	0.509	0.372	**0.507**
	PublicH	**0.666**	0.604	0.627	0.627	**0.615**	0.602	0.607	0.610	0.603	**0.621**	0.596	**0.611**	**0.377**	0.351	0.353	0.354	0.289	**0.354**
	Tropica	**0.688**	0.633	0.647	0.646	**0.635**	0.599	0.626	0.626	0.602	**0.634**	0.611	**0.623**	**0.445**	0.414	0.420	0.420	0.329	**0.419**
Communi	*Biol*	**0.742**	0.687	0.713	0.712	**0.704**	0.688	0.699	0.701	0.632	**0.708**	0.695	**0.706**	**0.560**	0.536	0.529	0.536	0.055	**0.529**
	CellBiol	**0.828**	0.769	0.765	0.767	**0.804**	0.765	0.775	0.778	0.732	**0.753**	0.772	**0.777**	**0.607**	0.587	0.574	0.585	0.354	**0.575**
	HealthS	**0.674**	0.622	0.637	0.634	**0.603**	0.577	0.595	0.596	0.622	**0.629**	**0.599**	0.598	**0.529**	0.501	0.502	0.503	0.538	**0.501**
	Medicin	**0.759**	0.710	0.723	0.721	**0.698**	0.665	0.693	0.692	0.714	**0.719**	0.693	**0.698**	**0.640**	0.587	0.604	0.607	0.607	**0.604**
	PublicH	**0.692**	0.643	0.655	0.657	**0.628**	0.607	0.622	0.620	0.643	**0.651**	0.615	**0.627**	**0.602**	0.555	0.573	0.574	0.580	**0.574**
	Tropica	**0.710**	0.658	0.666	0.664	**0.646**	0.613	0.636	0.636	0.585	**0.659**	0.626	**0.635**	**0.635**	0.605	0.613	0.612	0.443	**0.612**
HealthS	*Biol*	**0.722**	0.666	0.674	0.673	**0.668**	0.666	0.651	0.652	0.622	**0.673**	0.664	**0.685**	**0.488**	0.479	0.468	0.466	0.072	**0.469**
	CellBiol	**0.826**	0.745	0.757	0.762	**0.774**	0.762	0.773	0.776	0.749	**0.750**	0.770	**0.775**	**0.498**	0.490	0.474	0.476	0.297	**0.475**
	Communi	**0.704**	0.645	0.662	0.665	**0.659**	0.619	0.642	0.640	0.655	**0.650**	**0.646**	0.643	**0.622**	0.561	0.589	0.593	0.560	**0.589**
	Medicin	**0.738**	0.679	0.694	0.696	**0.697**	0.665	0.689	0.687	0.690	**0.688**	0.689	**0.693**	**0.572**	0.514	0.532	0.535	0.527	**0.534**
	PublicH	**0.705**	0.645	0.666	0.670	**0.665**	0.607	0.627	0.623	0.655	**0.658**	0.616	**0.628**	**0.640**	0.595	0.613	0.615	0.610	**0.615**
	Tropica	**0.692**	0.645	0.653	0.652	**0.649**	0.610	0.639	0.638	0.588	**0.642**	0.627	**0.637**	**0.609**	0.567	0.582	0.578	0.421	**0.581**
Medicin	*Biol*	**0.759**	0.710	0.725	0.724	**0.706**	0.685	0.703	0.702	0.595	**0.719**	0.694	**0.707**	**0.657**	0.615	0.614	0.613	0.059	**0.614**
	CellBiol	**0.818**	0.785	0.783	0.783	**0.793**	0.763	0.789	0.793	0.755	**0.781**	0.793	**0.801**	**0.640**	0.614	0.601	0.606	0.378	**0.602**
	Communi	**0.714**	0.664	0.674	0.671	**0.652**	0.628	0.644	0.647	0.666	**0.671**	0.640	**0.651**	**0.645**	0.584	0.618	0.616	0.620	**0.617**
	HealthS	**0.673**	0.615	0.630	0.632	**0.606**	0.575	0.599	0.600	0.610	**0.619**	0.594	**0.599**	**0.503**	0.483	0.482	0.483	0.505	**0.482**
	PublicH	**0.675**	0.628	0.634	0.634	**0.621**	0.600	0.615	0.612	**0.640**	0.632	0.615	**0.624**	**0.580**	0.538	0.555	0.556	0.552	**0.551**
	Tropica	**0.697**	0.656	0.656	0.652	**0.641**	0.611	0.634	0.634	0.594	**0.648**	0.629	**0.633**	**0.633**	0.592	0.606	0.600	0.447	**0.601**
PublicH	*Biol*	**0.726**	0.683	0.681	0.681	**0.677**	0.669	0.677	0.675	0.610	**0.672**	0.688	**0.682**	**0.508**	0.491	0.483	0.487	0.075	**0.483**
	CellBiol	**0.823**	0.735	0.751	0.753	**0.770**	0.760	0.766	0.771	0.730	**0.746**	0.772	**0.773**	**0.532**	0.511	0.513	0.511	0.339	**0.510**
	Communi	**0.701**	0.646	0.655	0.658	**0.650**	0.620	0.642	0.640	0.664	**0.650**	**0.648**	0.646	**0.638**	0.582	0.609	0.606	0.627	**0.607**
	HealthS	**0.681**	0.632	0.646	0.642	**0.608**	0.586	0.598	0.598	0.637	**0.641**	0.606	**0.603**	**0.615**	0.570	0.591	0.588	**0.610**	0.590
	Medicin	**0.750**	0.692	0.709	0.709	**0.699**	0.661	0.692	0.690	0.696	**0.700**	0.691	**0.692**	**0.586**	0.547	0.555	0.551	0.560	**0.552**
	Tropica	**0.694**	0.648	0.654	0.653	**0.648**	0.609	0.634	0.637	0.590	**0.645**	0.628	**0.644**	**0.619**	0.572	0.592	0.591	0.437	**0.589**
Tropica	*Biol*	**0.749**	0.687	0.721	0.717	**0.711**	0.688	0.703	0.705	0.615	**0.716**	0.693	**0.704**	**0.608**	0.588	0.583	0.592	0.036	**0.585**
	CellBiol	**0.824**	0.766	0.761	0.758	**0.797**	0.761	0.795	0.796	0.761	**0.756**	0.771	**0.799**	**0.636**	0.610	0.603	0.609	0.358	**0.606**
	Communi	**0.720**	0.654	0.681	0.680	**0.658**	0.620	0.650	0.650	0.577	**0.675**	0.633	**0.653**	**0.656**	0.593	0.624	0.629	0.449	**0.628**
	HealthS	**0.674**	0.611	0.636	0.636	**0.599**	0.580	0.593	0.594	0.574	**0.635**	0.595	**0.599**	**0.545**	0.508	0.519	0.516	0.374	**0.520**
	Medicin	**0.763**	0.704	0.721	0.722	**0.709**	0.671	0.696	0.698	0.639	**0.713**	0.678	**0.697**	**0.636**	0.575	0.601	0.601	0.414	**0.603**
	PublicH	**0.694**	0.639	0.656	0.652	**0.631**	0.609	0.625	0.624	0.588	**0.647**	0.609	**0.624**	**0.590**	0.541	0.571	0.566	0.392	**0.571**

Table B.8: F1 results for the OntoEA, SRC_TGT, EA and SRC_ONLY models using semantic upper-class (**parent(Cls)**) features from two KS ontologies: *sct* and *msh*.

SRC Domain	TGT Domain	OntoEA				SRC_TGT						EA		SRC_ONLY					
		sct+msh	*sct+msh*	*sct*	*msh*	*sct+msh*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*	*BoW*	*BoE*	*sct+msh*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*
		CCA				*CCA*								*CCA*					
Biol	*CellBiol*	**0.830**	0.792	0.798	0.799	**0.807**	0.792	0.800	0.805	0.724	**0.795**	0.796	**0.807**	**0.767**	0.743	0.744	0.748	0.348	**0.744**
	Communi	**0.694**	0.651	0.652	0.655	**0.651**	0.636	0.644	0.645	0.530	**0.645**	0.612	**0.648**	**0.550**	0.517	0.522	0.525	0.104	**0.523**
	HealthS	**0.648**	0.600	0.612	0.614	**0.608**	0.584	0.600	0.602	0.526	**0.609**	0.574	**0.609**	**0.433**	0.422	0.407	0.408	0.111	**0.406**
	Medicin	**0.760**	0.696	0.724	0.722	**0.709**	0.671	0.700	0.700	0.567	**0.718**	0.677	**0.707**	**0.643**	0.583	0.610	0.610	0.133	**0.609**
	PublicH	**0.672**	0.603	0.637	0.634	**0.633**	0.609	0.623	0.625	0.558	**0.631**	0.598	**0.633**	**0.483**	0.455	0.462	0.459	0.108	**0.462**
	Tropica	**0.694**	0.626	0.661	0.659	**0.652**	0.609	0.644	0.645	0.498	**0.646**	0.607	**0.644**	**0.557**	0.536	0.533	0.528	0.097	**0.535**
CellBiol	*Biol*	**0.777**	0.740	0.747	0.748	**0.714**	0.694	0.713	0.714	0.667	**0.737**	0.707	**0.712**	**0.683**	0.647	0.661	0.664	0.413	**0.662**
	Communi	**0.683**	0.655	0.644	0.644	**0.643**	0.621	0.635	0.637	0.609	**0.644**	0.621	**0.644**	**0.449**	0.430	0.425	0.424	0.328	**0.425**
	HealthS	**0.646**	0.593	0.612	0.613	**0.602**	0.577	0.596	0.596	0.590	**0.608**	0.587	**0.604**	**0.368**	0.347	0.347	0.352	0.270	**0.345**
	Medicin	**0.747**	0.698	0.711	0.711	**0.706**	0.658	0.698	0.701	0.670	**0.703**	0.686	**0.701**	**0.554**	0.506	0.525	0.525	0.387	**0.524**
	PublicH	**0.680**	0.617	0.644	0.645	**0.629**	0.609	0.623	0.628	0.613	**0.639**	0.604	**0.628**	**0.394**	0.377	0.371	0.372	0.280	**0.368**
	Tropica	**0.699**	0.644	0.660	0.658	**0.650**	0.602	0.641	0.642	0.606	**0.646**	0.618	**0.638**	**0.473**	0.447	0.444	0.444	0.336	**0.443**
Communi	*Biol*	**0.753**	0.684	0.723	0.723	**0.714**	0.694	0.710	0.712	0.650	**0.719**	0.706	**0.717**	**0.618**	0.573	0.592	0.598	0.091	**0.592**
	CellBiol	**0.826**	0.764	0.762	0.764	**0.810**	0.761	0.771	0.773	0.739	**0.750**	0.775	**0.772**	**0.654**	0.622	0.631	0.637	0.386	**0.632**
	HealthS	**0.679**	0.624	0.643	0.640	**0.614**	0.583	0.605	0.607	0.624	**0.635**	0.604	**0.608**	**0.555**	0.510	0.526	0.528	0.552	**0.525**
	Medicin	**0.762**	0.710	0.726	0.723	**0.700**	0.665	0.695	0.694	0.715	**0.722**	0.693	**0.700**	**0.656**	0.592	0.617	0.619	0.609	**0.617**
	PublicH	**0.697**	0.646	0.659	0.661	**0.635**	0.612	0.629	0.628	0.646	**0.655**	0.621	**0.635**	**0.616**	0.559	0.586	0.587	0.590	**0.586**
	Tropica	**0.712**	0.660	0.668	0.667	**0.650**	0.616	0.640	0.641	0.593	**0.662**	0.628	**0.640**	**0.653**	0.613	0.631	0.630	0.464	**0.630**
HealthS	*Biol*	**0.721**	0.660	0.673	0.672	**0.665**	0.661	0.663	0.664	0.637	**0.672**	0.673	**0.686**	**0.534**	0.501	0.506	0.506	0.127	**0.509**
	CellBiol	**0.822**	0.740	0.752	0.758	**0.772**	0.761	0.771	0.772	**0.754**	0.746	0.774	**0.773**	**0.510**	0.498	0.484	0.488	0.412	**0.486**
	Communi	**0.705**	0.651	0.664	0.667	**0.661**	0.625	0.645	0.643	**0.658**	0.651	**0.650**	0.646	**0.632**	0.574	0.596	0.602	0.572	**0.595**
	Medicin	**0.739**	0.678	0.695	0.697	**0.698**	0.664	0.692	0.689	**0.691**	0.689	0.690	**0.695**	**0.590**	0.520	0.544	0.548	0.536	**0.546**
	PublicH	**0.708**	0.649	0.669	0.672	**0.640**	0.612	0.634	0.630	0.658	**0.661**	0.623	**0.636**	**0.644**	0.601	0.614	0.615	0.615	**0.616**
	Tropica	**0.694**	0.648	0.655	0.655	**0.652**	0.612	0.642	0.642	0.599	**0.645**	0.629	**0.641**	**0.609**	0.567	0.581	0.577	0.448	**0.580**
Medicin	*Biol*	**0.770**	0.722	0.736	0.736	**0.716**	0.693	0.714	0.714	0.616	**0.731**	0.704	**0.718**	**0.707**	0.638	0.658	0.657	0.093	**0.658**
	CellBiol	**0.830**	0.794	0.795	0.796	**0.801**	0.758	0.798	0.801	0.774	**0.793**	0.801	**0.809**	**0.716**	0.657	0.674	0.675	0.434	**0.671**
	Communi	**0.716**	0.670	0.678	0.674	**0.656**	0.634	0.648	0.650	0.668	**0.674**	0.643	**0.655**	**0.648**	0.593	0.620	0.618	0.625	**0.618**
	HealthS	**0.679**	0.621	0.636	0.639	**0.617**	0.581	0.609	0.610	0.615	**0.625**	0.599	**0.608**	**0.548**	0.503	0.520	0.519	0.528	**0.519**
	PublicH	**0.683**	0.635	0.643	0.643	**0.629**	0.605	0.623	0.621	**0.646**	0.640	0.621	**0.632**	**0.601**	0.553	0.575	0.576	0.571	**0.572**
	Tropica	**0.701**	0.659	0.660	0.656	**0.646**	0.612	0.639	0.640	0.602	**0.652**	0.631	**0.638**	**0.648**	0.599	0.616	0.609	0.472	**0.611**
PublicH	*Biol*	**0.726**	0.677	0.682	0.682	**0.680**	0.664	0.679	0.677	0.624	**0.673**	0.699	**0.684**	**0.555**	0.518	0.531	0.534	0.135	**0.534**
	CellBiol	**0.819**	0.729	0.747	0.749	**0.768**	0.760	0.772	0.768	0.737	**0.743**	0.776	**0.772**	**0.556**	0.524	0.537	0.537	0.404	**0.537**
	Communi	**0.702**	0.650	0.658	0.660	**0.652**	0.625	0.644	0.642	0.665	**0.654**	0.652	**0.648**	**0.639**	0.582	0.608	0.604	0.628	**0.606**
	HealthS	**0.685**	0.634	0.651	0.647	**0.619**	0.592	0.608	0.609	0.640	**0.647**	0.610	**0.613**	**0.628**	0.575	0.604	0.600	0.618	**0.604**
	Medicin	**0.751**	0.691	0.710	0.711	**0.701**	0.660	0.695	0.692	0.696	**0.701**	0.691	**0.694**	**0.607**	0.553	0.571	0.565	0.570	**0.566**
	Tropica	**0.696**	0.650	0.657	0.655	**0.651**	0.610	0.638	0.641	0.599	**0.647**	0.630	**0.648**	**0.625**	0.572	0.596	0.595	0.450	**0.593**
Tropica	Biology	**0.760**	0.684	0.731	0.727	**0.720**	0.696	0.713	0.716	0.643	**0.726**	0.706	**0.715**	**0.639**	0.602	0.614	0.621	0.102	**0.616**
	CellBiol	**0.833**	0.761	0.757	0.755	**0.804**	0.759	0.803	0.803	0.779	**0.752**	0.770	**0.807**	**0.654**	0.626	0.630	0.634	0.414	**0.633**
	Communi	**0.720**	0.657	0.682	0.682	**0.661**	0.625	0.654	0.654	0.598	**0.676**	0.639	**0.657**	**0.655**	0.601	0.622	0.628	0.468	**0.626**
	HealthS	**0.680**	0.614	0.640	0.641	**0.610**	0.587	0.603	0.605	0.588	**0.639**	0.602	**0.609**	**0.553**	0.513	0.526	0.522	0.411	**0.526**
	Medicin	**0.765**	0.704	0.723	0.724	**0.710**	0.671	0.698	0.700	0.648	**0.715**	0.679	**0.699**	**0.641**	0.576	0.603	0.601	0.415	**0.604**
	PublicH	**0.698**	0.642	0.662	0.657	**0.639**	0.615	0.633	0.633	0.603	**0.652**	0.616	**0.632**	**0.595**	0.551	0.574	0.570	0.434	**0.574**

Table B.9: Precision results for the OntoEA, SRC_TGT, EA and SRC_ONLY models using semantic upper-class (**parent(Cls)**) features from two KS ontologies: *sct* and *msh*.

SRC Domain	TGT Domain	OntoEA				SRC_TGT						EA		SRC_ONLY					
		sct+msh CCA	*sct+msh*	*sct*	*msh*	*sct+msh CCA*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*	*BoW*	*BoE*	*sct+msh CCA*	*sct+msh*	*sct*	*msh*	*BoW*	*BoE*
Biol	*CellBiol*	**0.819**	0.780	0.784	0.786	**0.778**	0.782	0.789	0.761	0.680	**0.781**	0.782	**0.796**	**0.749**	0.723	0.725	0.728	0.367	**0.725**
	Communi	**0.683**	0.628	0.639	0.642	**0.634**	0.608	0.626	0.626	0.511	**0.631**	0.599	**0.630**	**0.531**	0.473	0.507	0.507	0.092	**0.508**
	HealthS	**0.620**	0.567	0.579	0.580	**0.587**	0.548	0.562	0.563	0.499	**0.577**	0.546	**0.577**	**0.415**	0.375	0.388	0.392	0.100	**0.388**
	Medicin	**0.754**	0.687	0.715	0.715	**0.700**	0.659	0.689	0.689	0.538	**0.710**	0.672	**0.696**	**0.645**	0.574	0.613	0.613	0.102	**0.613**
	PublicH	**0.649**	0.577	0.611	0.609	**0.609**	0.580	0.597	0.598	0.540	**0.605**	0.579	**0.606**	**0.462**	0.403	0.438	0.437	0.102	**0.438**
	Tropica	**0.679**	0.608	0.644	0.641	**0.632**	0.589	0.620	0.619	0.481	**0.628**	0.599	**0.619**	**0.547**	0.499	0.528	0.520	0.080	**0.530**
CellBiol	*Biol*	**0.766**	0.731	0.736	0.736	**0.698**	0.684	0.696	0.698	0.644	**0.726**	0.689	**0.696**	**0.643**	0.625	0.621	0.623	0.314	**0.622**
	Communi	**0.669**	0.631	0.627	0.627	**0.627**	0.612	0.616	0.618	0.593	**0.626**	0.606	**0.625**	**0.443**	0.410	0.421	0.421	0.351	**0.421**
	HealthS	**0.616**	0.586	0.583	0.583	**0.568**	0.568	0.559	0.558	0.569	**0.579**	**0.573**	0.566	**0.365**	0.337	0.340	0.344	0.309	**0.339**
	Medicin	**0.742**	0.690	0.704	0.705	**0.697**	0.660	0.687	0.690	0.662	**0.697**	0.679	**0.689**	**0.564**	0.519	0.536	0.538	0.397	**0.536**
	PublicH	**0.658**	0.596	0.617	0.618	**0.607**	0.599	0.597	0.600	0.597	**0.612**	0.593	**0.602**	**0.404**	0.369	0.379	0.380	0.327	**0.380**
	Tropica	**0.682**	0.628	0.640	0.638	**0.627**	0.598	0.617	0.616	0.603	**0.626**	0.610	**0.614**	**0.473**	0.441	0.448	0.447	0.362	**0.447**
Communi	*Biol*	**0.737**	0.694	0.707	0.706	**0.698**	0.688	0.693	0.696	0.624	**0.703**	0.691	**0.700**	**0.561**	0.535	0.530	0.536	0.063	**0.529**
	CellBiol	**0.834**	0.778	0.770	0.773	**0.802**	0.776	0.783	0.787	0.732	**0.759**	0.779	**0.784**	**0.610**	0.582	0.574	0.585	0.393	**0.575**
	HealthS	**0.671**	0.624	0.634	0.632	**0.597**	0.576	0.589	0.589	0.622	**0.627**	**0.597**	0.592	**0.535**	0.509	0.510	0.511	**0.544**	0.509
	Medicin	**0.760**	0.712	0.723	0.721	**0.698**	0.668	0.693	0.692	0.714	**0.719**	0.693	**0.698**	**0.648**	0.599	0.612	0.616	0.616	**0.613**
	PublicH	**0.690**	0.643	0.653	0.655	**0.624**	0.605	0.617	0.615	0.641	**0.649**	0.611	**0.622**	**0.606**	0.560	0.577	0.579	**0.584**	0.578
	Tropica	**0.709**	0.659	0.664	0.662	**0.644**	0.613	0.633	0.633	0.579	**0.657**	0.627	**0.632**	**0.635**	0.605	0.612	0.612	0.445	**0.611**
HealthS	*Biol*	**0.725**	0.675	0.677	0.677	**0.681**	0.676	0.660	0.664	0.613	**0.677**	0.659	**0.688**	**0.500**	0.489	0.478	0.479	0.082	**0.481**
	CellBiol	**0.834**	0.755	0.765	0.770	**0.782**	0.772	0.781	0.782	0.752	**0.758**	0.775	**0.784**	**0.509**	0.493	0.481	0.485	0.324	**0.483**
	Communi	**0.704**	0.644	0.662	0.664	**0.658**	0.617	0.641	0.638	**0.656**	0.649	**0.644**	0.641	**0.630**	0.572	0.599	0.604	0.573	**0.599**
	Medicin	**0.739**	0.682	0.695	0.696	**0.697**	0.668	0.689	0.687	**0.691**	0.688	0.690	**0.693**	**0.593**	0.542	0.558	0.562	0.551	**0.559**
	PublicH	**0.703**	0.645	0.664	0.669	**0.760**	0.606	0.622	0.618	0.654	**0.656**	0.613	**0.624**	**0.641**	0.599	0.615	0.618	0.613	**0.618**
	Tropica	**0.691**	0.646	0.651	0.650	**0.647**	0.611	0.636	0.635	0.580	**0.641**	0.627	**0.634**	**0.617**	0.578	0.591	0.590	0.430	**0.591**
Medicin	*Biol*	**0.754**	0.707	0.719	0.718	**0.700**	0.683	0.696	0.696	0.585	**0.714**	0.690	**0.701**	**0.650**	0.612	0.610	0.608	0.068	**0.609**
	CellBiol	**0.812**	0.781	0.777	0.777	**0.789**	0.773	0.785	0.789	0.743	**0.775**	0.789	**0.797**	**0.628**	0.602	0.586	0.591	0.409	**0.587**
	Communi	**0.714**	0.664	0.673	0.670	**0.650**	0.627	0.642	0.645	0.666	**0.670**	0.639	**0.649**	**0.644**	0.587	0.620	0.618	0.623	**0.618**
	HealthS	**0.670**	0.615	0.626	0.629	**0.600**	0.574	0.593	0.593	0.607	**0.616**	0.591	**0.593**	**0.498**	0.484	0.480	0.479	**0.500**	0.478
	PublicH	**0.671**	0.627	0.629	0.629	**0.617**	0.597	0.610	0.607	**0.637**	0.627	0.611	**0.620**	**0.573**	0.535	0.550	0.550	0.546	**0.546**
	Tropica	**0.695**	0.657	0.654	0.650	**0.638**	0.611	0.631	0.630	0.589	**0.645**	**0.630**	0.629	**0.629**	0.593	0.604	0.598	0.438	**0.599**
PublicH	*Biol*	**0.728**	0.692	0.683	0.684	**0.678**	0.678	0.678	0.676	0.601	**0.674**	0.683	**0.683**	**0.517**	0.500	0.493	0.498	0.078	**0.493**
	CellBiol	**0.829**	0.745	0.757	0.759	**0.778**	0.771	0.772	0.780	0.731	**0.753**	0.777	**0.782**	**0.542**	0.517	0.521	0.522	0.375	**0.520**
	Communi	**0.700**	0.646	0.653	0.656	**0.648**	0.619	0.640	0.638	**0.664**	0.648	**0.646**	0.644	**0.645**	0.592	0.618	0.615	**0.636**	0.615
	HealthS	**0.679**	0.634	0.644	0.640	**0.603**	0.585	0.592	0.591	0.637	**0.639**	0.604	**0.597**	**0.616**	0.575	0.594	0.590	0.617	**0.592**
	Medicin	**0.750**	0.695	0.709	0.709	**0.699**	0.663	0.692	0.690	0.696	**0.700**	0.691	**0.691**	**0.604**	0.567	0.576	0.573	**0.582**	0.573
	Tropica	**0.693**	0.649	0.652	0.651	**0.646**	0.610	0.631	0.634	0.583	**0.643**	0.628	**0.641**	**0.625**	0.581	0.598	0.597	0.453	**0.595**
Tropica	*Biol*	**0.744**	0.695	0.716	0.711	**0.705**	0.687	0.697	0.699	0.600	**0.711**	0.687	**0.699**	**0.610**	0.592	0.584	0.593	0.035	**0.585**
	CellBiol	**0.820**	0.776	0.766	0.763	**0.793**	0.771	0.792	0.793	0.752	**0.762**	0.777	**0.795**	**0.637**	0.608	0.599	0.606	0.375	**0.602**
	Communi	**0.721**	0.654	0.681	0.681	**0.657**	0.619	0.648	0.648	0.565	**0.675**	0.631	**0.651**	**0.664**	0.603	0.632	0.637	0.451	**0.636**
	HealthS	**0.672**	0.612	0.634	0.634	**0.592**	0.580	0.588	0.588	0.563	**0.633**	0.591	**0.593**	**0.553**	0.520	0.528	0.525	0.386	**0.530**
	Medicin	**0.764**	0.707	0.721	0.723	**0.709**	0.674	0.696	0.698	0.635	**0.713**	0.678	**0.697**	**0.652**	0.597	0.619	0.619	0.435	**0.620**
	PublicH	**0.692**	0.639	0.653	0.650	**0.628**	0.606	0.620	0.619	0.577	**0.644**	0.605	**0.619**	**0.601**	0.554	0.581	0.577	0.395	**0.581**

Table B.10: Recall results for the OntoEA, SRC_TGT, EA and SRC_ONLY models using semantic upper-class (**parent(Cls)**) features from two KS ontologies: *sct* and *msh*.

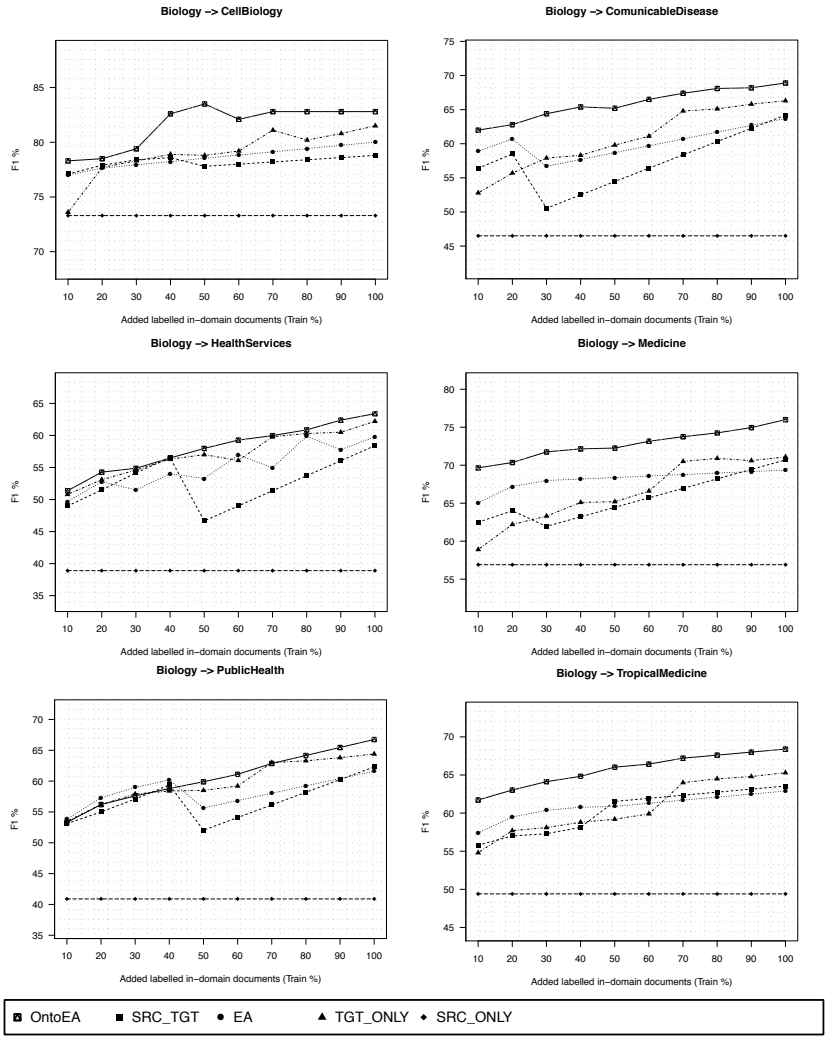

Figure B.1: F1 curves for the OntoEA, SRC_TGT, EA and SRC_ONLY classifiers, having Biology as source domain.

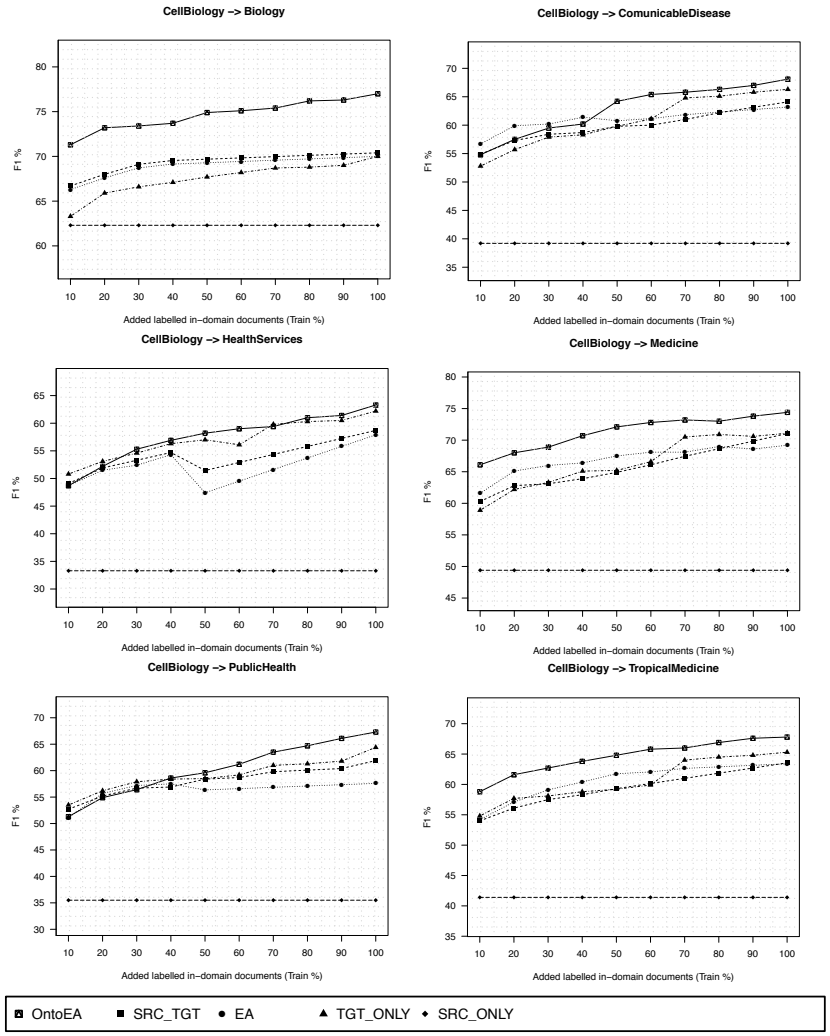

Figure B.2: F1 curves for the OntoEA, SRC_TGT, EA and SRC_ONLY classifiers, having Cell Biology as source domain.

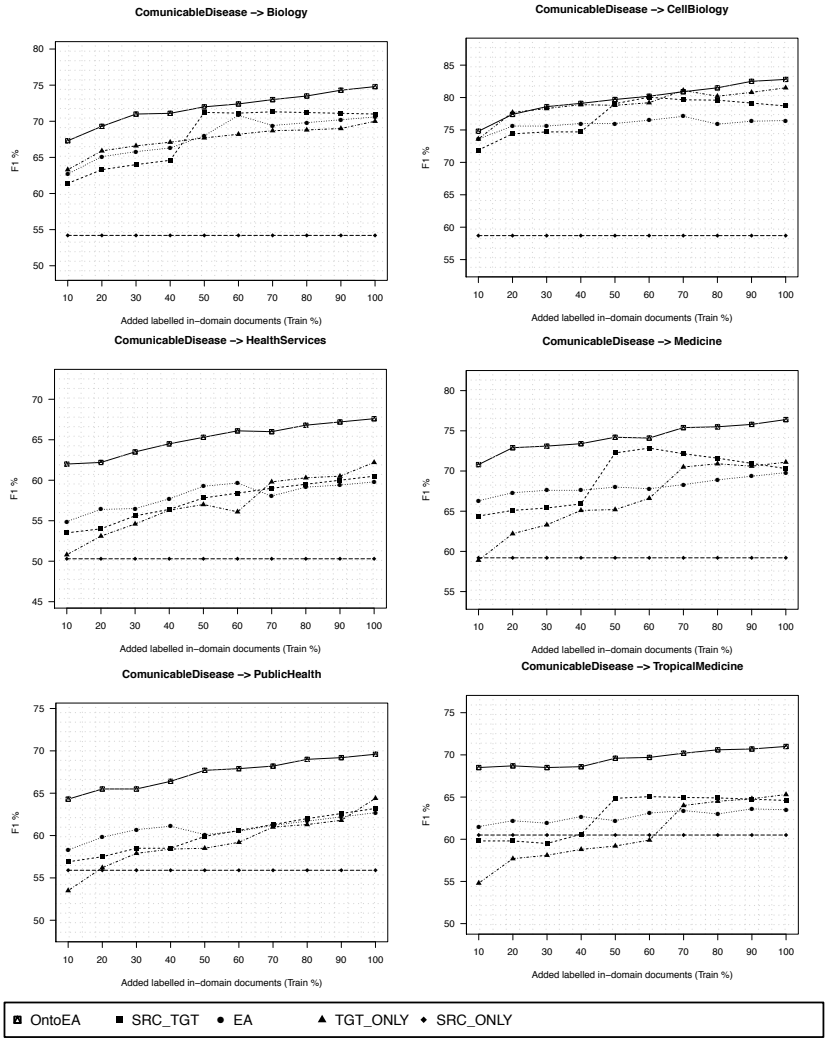

Figure B.3: F1 curves for the OntoEA, SRC_TGT, EA and SRC_ONLY classifiers, having Comunicable Disease as source domain.

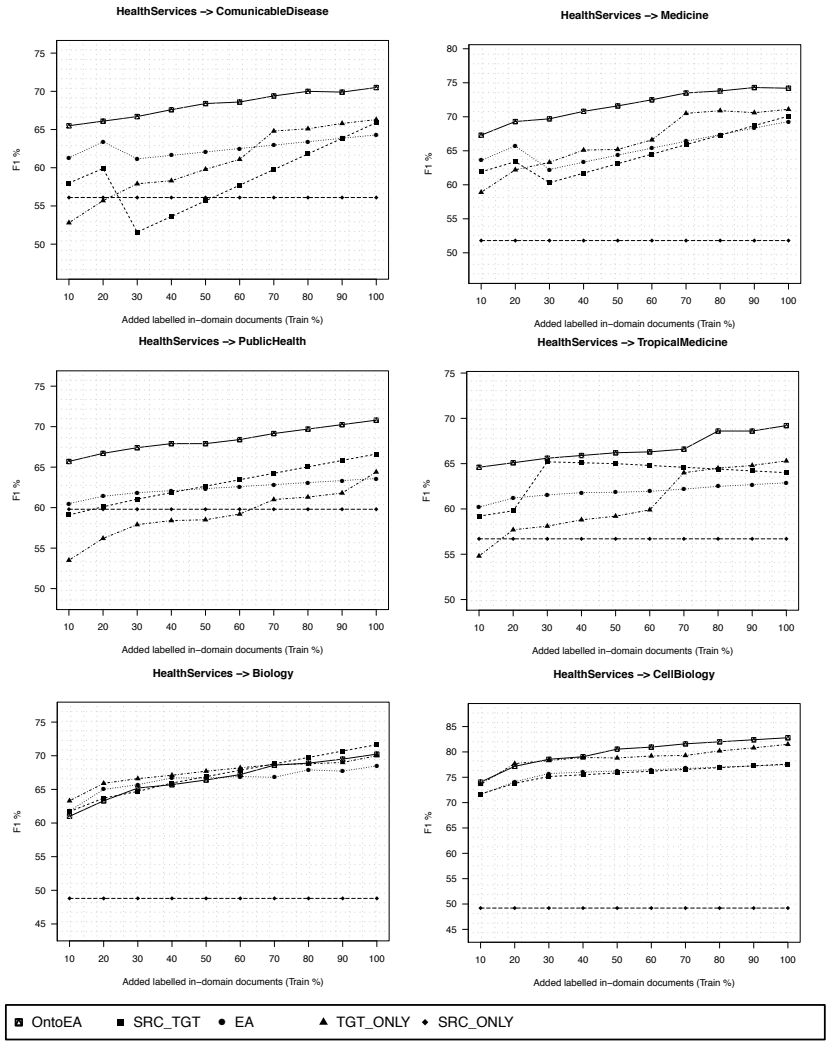

Figure B.4: F1 curves for the OntoEA, SRC_TGT, EA and SRC_ONLY classifiers, having Health Services as source domain.

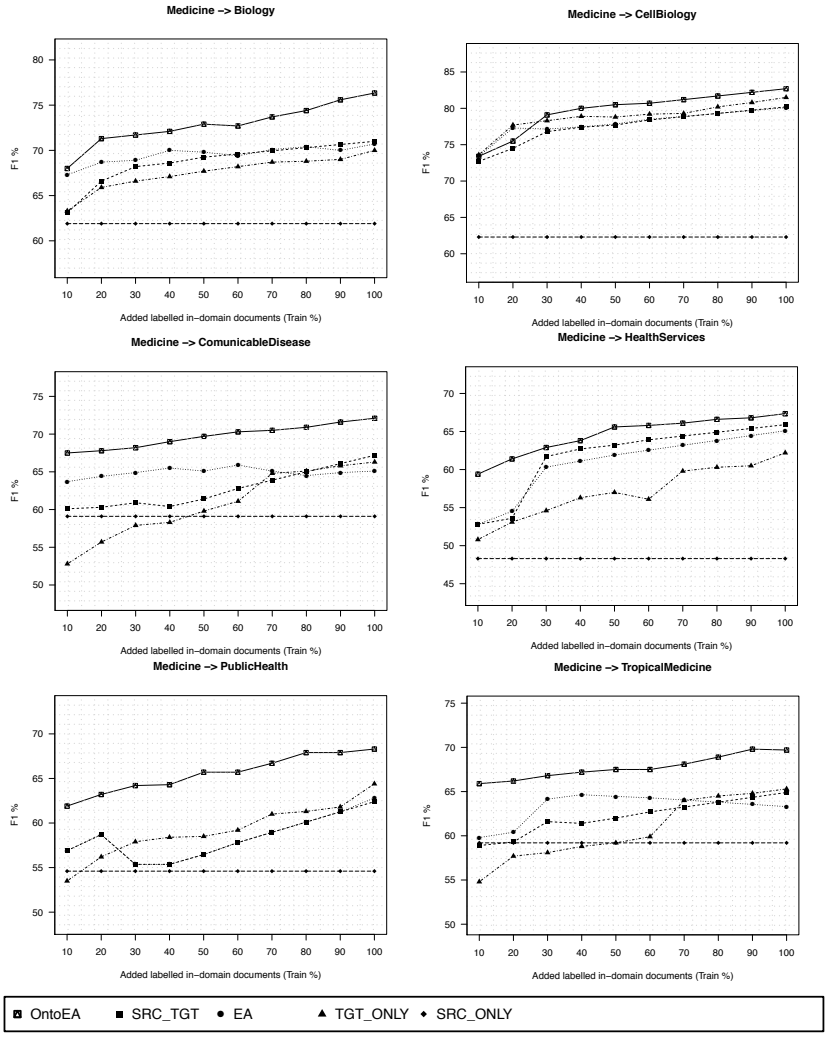

Figure B.5: F1 curves for the OntoEA, SRC_TGT, EA and SRC_ONLY classifiers, having Medicine as source domain.

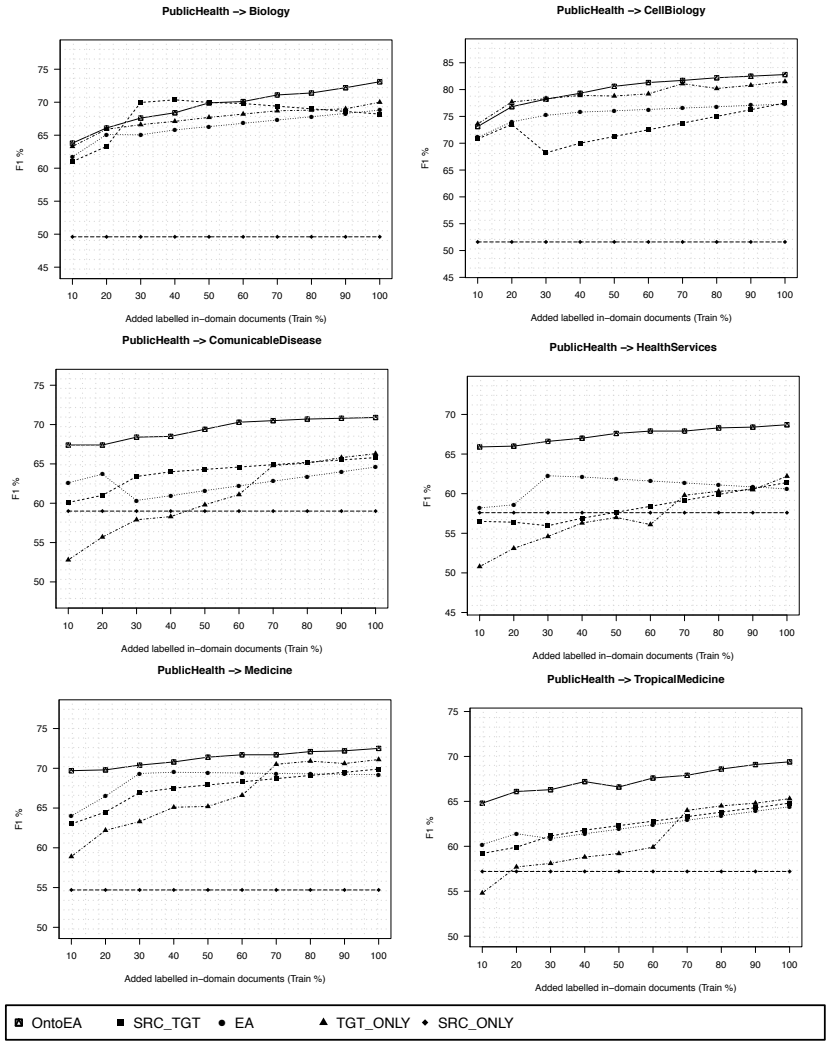

Figure B.6: F1 curves for the OntoEA, SRC_TGT, EA and SRC_ONLY classifiers, having Public Health as source domain.

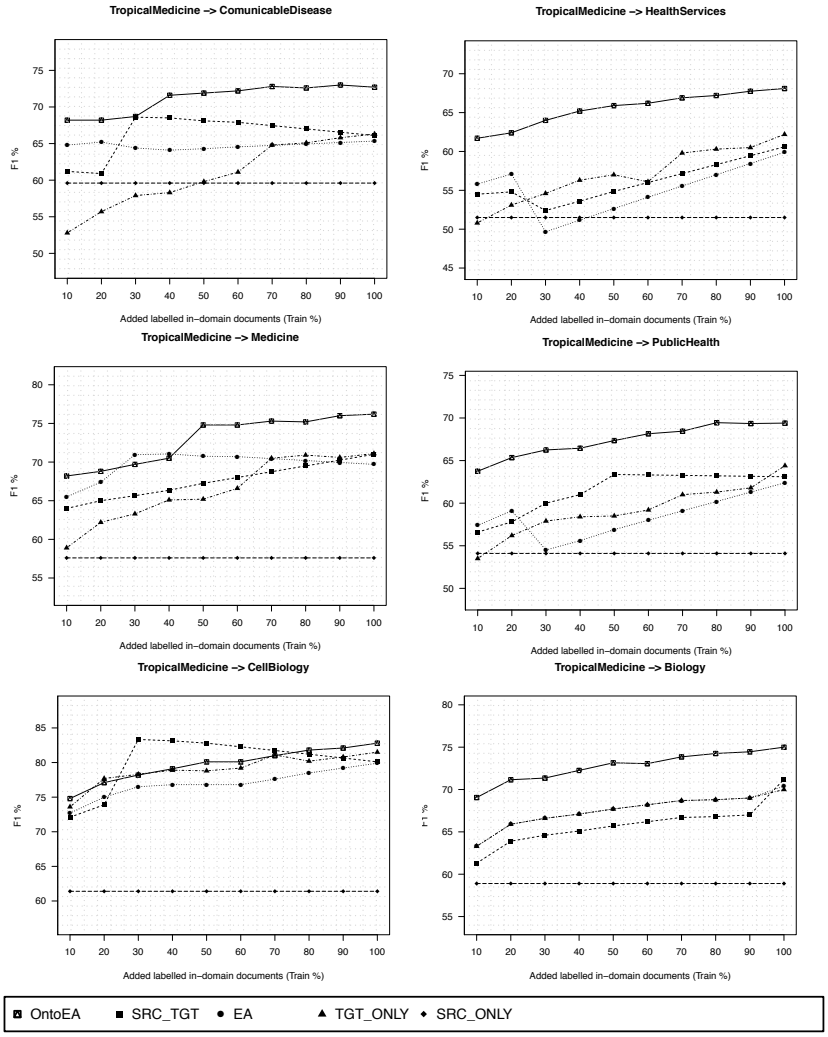

Figure B.7: F1 curves for the OntoEA, SRC_TGT, EA and SRC_ONLY classifiers, having Tropical Medicine as source domain.

Appendix C

Additional Experimental Results on Adaptive Topic Classification

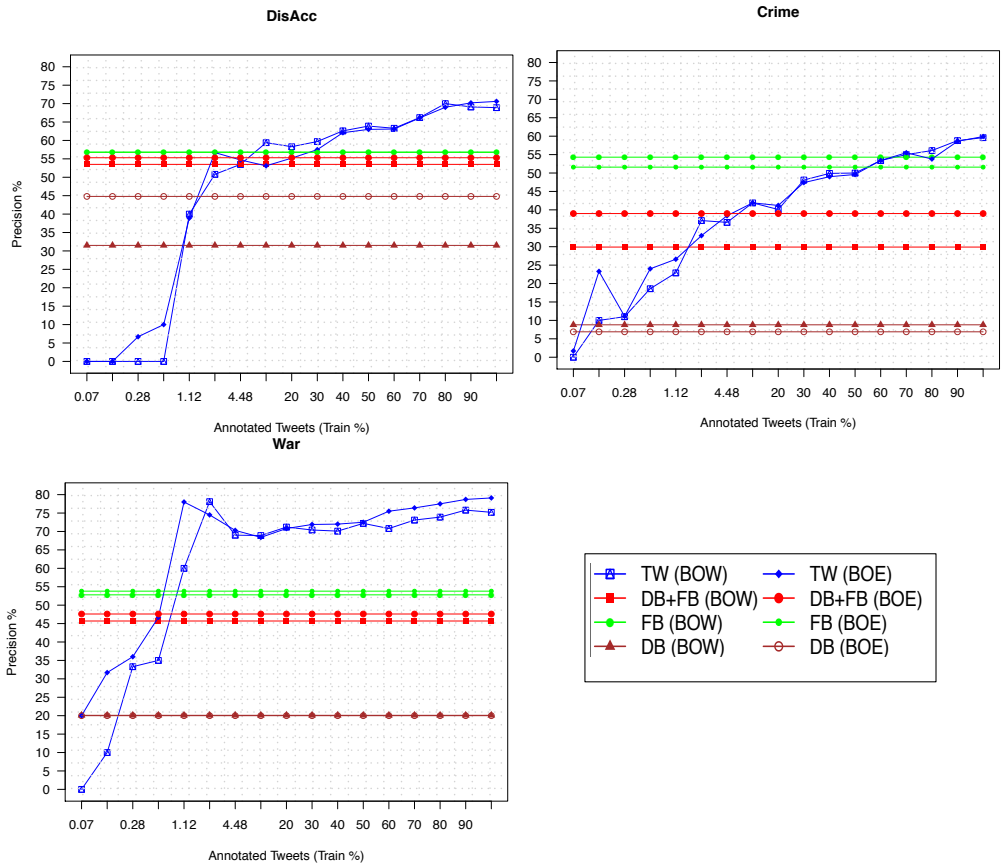

Figure C.1: Precision results for the single-source TW classifier and cross-source DB, FB and DB+FB classifiers over the full learning curve using lexical features.

This appendix contains additional experimental results obtained for the adaptive topic classifiers introduced in Chapter 6. It presents the results obtained for the single-source TW classifier and cross-source DB, FB and DB+FB classifiers over the full learning curve using lexical features ($BoW$and BoE). Figure C.1 shows the results in terms of precision,

and Figure C.2 shows the results for recall.

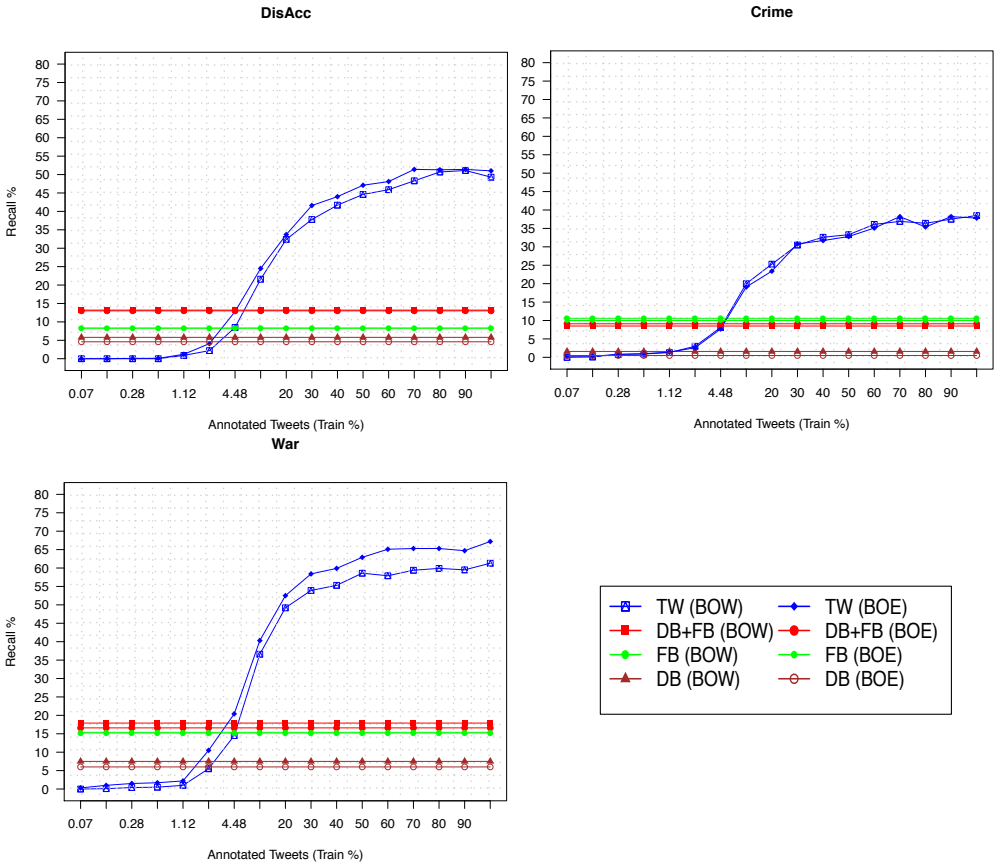

Figure C.2: Recall results for the single-source TW classifier and cross-source DB, FB and DB+FB classifiers over the full learning curve using lexical features.

C.1 Results Obtained using Semantic Meta-graph Features in Single-domain Scenario

This subsection presents the results obtained using Semantic Meta-graph Features in single-domain scenario. The results in terms of precision can be seen in Figure C.3 and in terms of recall in Figure C.4.

C.2 Results Obtained using Semantic Meta-graph Features in Cross-domain Scenario

This subsection presents the results obtained using Semantic Meta-graph Features in cingle-domain scenario. The results in terms of precision can be seen in Table C.1 and in terms of recall in Table C.2.

Dataset	Semantic graph	Features	DB+FB	DB+FB+TW	DB	DB+TW	FB	FB+TW
			P	*P*	*P*	*P*	*P*	*P*
War	Baseline	BOW	0.420	0.955	0.208	0.877	0.678	0.851
		POS	0.217	0.952	0.258	0.859	0.597	0.809
		BOE	0.903	0.842	0.490	0.856	0.767	0.753
		BOC	0.038	0.748	0.176	0.837	0.486	0.807
	Resource	Cls(W-Freq)	0.370	0.957	0.221	0.881	0.678	0.844
		parent(Cls)(W-Freq)	0.426	0.957	0.206	0.877	0.678	0.846
		P(W-Freq/Cls)	0.364	0.956	0.222	0.876	0.683	0.845
		Cls+P(W-SG)	0.422	0.956	0.195	0.878	0.673	0.844
		parent(Cls)+P(W-SG)	0.406	0.955	0.244	0.874	0.683	0.844
		P(W-SG)	0.902	0.967	0.303	0.874	0.670	0.850
	Category	Cat(W-Freq)	0.395	0.959	0.302	0.878	0.693	0.854
		parent(Cat)(W-Freq)	0.418	0.960	0.326	0.875	0.679	0.852
		P(W-Freq/Cat)	0.401	0.959	0.225	0.878	0.679	0.850
		Cat+P(W-SG)	0.431	0.960	0.221	0.880	NA	NA
		parent(Cat)+P(W-SG)	0.387	0.961	0.230	0.878	NA	NA
		P(W-SG/Cat)	0.441	0.960	0.341	0.877	0.679	0.852
Cri	Baseline	BOW	0.489	0.944	0.071	0.718	0.747	0.723
		POS	0.448	0.950	0.069	0.676	0.695	0.667
		BOE	0.353	0.814	0.049	0.744	0.656	0.733
		BOC	0.037	0.655	0.051	0.700	0.638	0.692
	Resource	Cls(W-Freq)	0.616	0.944	0.083	0.702	0.691	0.722
		parent(Cls)(W-Freq)	0.586	0.944	0.082	0.705	0.740	0.728
		P(W-Freq/Cls)	0.628	0.944	0.096	0.705	0.710	0.724
		Cls+P(W-SG)	0.663	0.945	0.062	0.703	0.726	0.728
		parent(Cls)+P(W-SG)	0.617	0.945	0.067	0.706	0.738	0.716
		P(W-SG/Cls)	0.666	0.944	0.126	0.699	0.713	0.739
	Category	Cat(W-Freq)	0.606	0.948	0.061	0.674	0.758	0.689
		parent(Cat)(W-Freq)	0.472	0.947	0.088	0.690	0.763	0.690
		P(W-Freq/Cat)	0.458	0.947	0.069	0.690	0.785	0.688
		Cat+P(W-SG)	0.461	0.948	0.057	0.684	NA	NA
		parent(Cat)+P(W-SG)	0.457	0.946	0.086	0.702	NA	NA
		P(W-SG/Cat)	0.606	0.947	0.090	0.695	0.740	0.699
DisAcc	Baseline	BOW	0.216	0.955	0.584	0.782	0.835	0.819
		POS	0.322	0.951	0.273	0.746	0.719	0.744
		BOE	0.875	0.810	0.494	0.806	0.909	0.744
		BOC	0.059	0.743	0.366	0.785	0.626	0.740
	Resource	Cls(W-Freq/Cls)	0.293	0.951	0.553	0.783	0.835	0.805
		parent(Cls)(W-Freq)	0.267	0.953	0.568	0.789	0.835	0.814
		P(W-Freq/Cls)	0.238	0.953	0.519	0.777	0.835	0.805
		Cls+P(W-SG)	0.237	0.953	0.570	0.786	0.835	0.812
		parent(Cls)+P(W-SG)	0.268	0.953	0.578	0.785	0.835	0.816
		P(W-SG/Cls)	0.248	0.954	0.643	0.800	0.835	0.815
	Category	Cat(W-Freq)	0.233	0.957	0.521	0.801	0.787	0.815
		parent(Cat)(W-Freq)	0.271	0.957	0.546	0.792	0.775	0.802
		P(W-Freq/Cat)	0.266	0.956	0.562	0.770	0.775	0.801
		Cat+P(W-SG)	0.261	0.958	0.555	0.761	NA	NA
		parent(Cat)+P(W-SG)	0.327	0.959	0.552	0.782	NA	NA
		P(W-SG/Cat)	0.312	0.956	0.661	0.787	0.775	0.790

Table C.1: Precision results for the DB, FB and $DB+FB$ cross-domain SVM topic classifiers using different KSs ontologies (DB -using *dbKS*'s ontologies, FB -using *fbKS*'s ontology) and two semantic meta-graphs derived from these KSs (*resource meta-graph* (Resource) and *category meta-graph* (Category)).

Dataset	Semantic graph	Features	DB+FB	DB+FB+TW	DB	DB+TW	FB	FB+TW
			R	*R*	*R*	*R*	*R*	*R*
War	Baseline	BOW	0.011	0.861	0.049	0.723	0.136	0.722
		POS	0.006	0.880	0.034	0.744	0.148	0.746
		BOE	0.007	0.761	0.040	0.754	0.12	0.801
		BOC	0.013	0.658	0.061	0.698	0.125	0.704
	Resource	Cls(W-Freq)	0.009	0.878	0.045	0.720	0.136	0.699
		parent(Cls)(W-Freq)	0.011	0.880	0.047	0.727	0.136	0.718
		P(W-Freq/Cls)	0.009	0.871	0.054	0.717	0.136	0.712
		Cls+P(W-SG)	0.011	0.864	0.043	0.726	0.136	0.723
		parent(Cls)+P(W-SG)	0.011	0.863	0.040	0.716	0.136	0.715
		P(W-SG)	0.006	0.879	0.062	0.731	0.136	0.732
	Category	Cat(W-Freq)	0.011	0.872	0.046	0.745	0.147	0.741
		parent(Cat)(W-Freq)	0.013	0.874	0.064	0.734	0.147	0.736
		P(W-Freq/Cat)	0.012	0.865	0.054	0.723	0.147	0.723
		Cat+P(W-SG)	0.013	0.867	0.054	0.724	NA	NA
		parent(Cat)+P(W-SG)	0.011	0.869	0.056	0.730	NA	NA
		P(W-SG/Cat)	0.013	0.878	0.055	0.742	0.147	0.744
Cri	Baseline	BOW	0.013	0.857	0.006	0.477	0.143	0.489
		POS	0.013	0.860	0.005	0.527	0.150	0.517
		BOE	0.028	0.626	0.004	0.502	0.108	0.498
		BOC	0.014	0.446	0.014	0.457	0.155	0.469
	Resource	Cls(W-Freq)	0.011	0.873	0.009	0.471	0.140	0.486
		parent(Cls)(W-Freq)	0.010	0.873	0.008	0.477	0.141	0.489
		P(W-Freq/Cls)	0.011	0.866	0.011	0.473	0.145	0.484
		Cls+P(W-SG)	0.013	0.858	0.006	0.464	0.140	0.490
		parent(Cls)+P(W-SG)	0.012	0.858	0.006	0.469	0.143	0.490
		P(W-SG/Cls)	0.014	0.864	0.009	0.468	0.138	0.496
	Category	Cat(W-Freq)	0.011	0.860	0.007	0.502	0.156	0.509
		parent(Cat)(W-Freq)	0.011	0.858	0.010	0.512	0.145	0.512
		P(W-Freq/Cat)	0.011	0.855	0.007	0.499	0.157	0.505
		Cat+P(W-SG)	0.011	0.856	0.005	0.486	NA	NA
		parent(Cat)+P(W-SG)	0.012	0.856	0.009	0.507	NA	NA
		P(W-SG/Cat)	0.012	0.859	0.007	0.515	0.150	0.517
DisAcc	Baseline	BOW	0.002	0.869	0.059	0.608	0.090	0.605
		POS	0.009	0.860	0.029	0.630	0.090	0.625
		BOE	0.040	0.629	0.043	0.653	0.048	0.648
		BOC	0.014	0.506	0.098	0.564	0.103	0.544
	Resource	Cls(W-Freq/Cls)	0.002	0.881	0.070	0.599	0.090	0.605
		parent(Cls)(W-Freq)	0.002	0.883	0.060	0.611	0.090	0.601
		P(W-Freq/Cls)	0.002	0.871	0.070	0.600	0.090	0.591
		Cls+P(W-SG)	0.002	0.866	0.067	0.606	0.090	0.598
		parent(Cls)+P(W-SG)	0.002	0.866	0.062	0.615	0.090	0.602
		P(W-SG/Cls)	0.002	0.873	0.059	0.603	0.090	0.607
	Category	Cat(W-Freq)	0.002	0.869	0.057	0.625	0.093	0.621
		parent(Cat)(W-Freq)	0.002	0.869	0.048	0.617	0.091	0.602
		P(W-Freq/Cat)	0.002	0.863	0.052	0.611	0.091	0.612
		Cat+P(W-SG)	0.002	0.862	0.052	0.629	NA	NA
		parent(Cat)+P(W-SG)	0.003	0.865	0.054	0.613	NA	NA
		P(W-SG/Cat)	0.002	0.874	0.047	0.609	0.091	0.601

Table C.2: Recall results for the DB, FB and $DB + FB$ cross-domain SVM topic classifiers using different KSs ontologies (DB -using *dbKS*'s ontologies, FB -using *fbKS*'s ontology) and two semantic meta-graphs derived from these KSs (*resource meta-graph* (Resource) and *category meta-graph* (Category)).

Dataset	Semantic graph	Features	TW(*dbKS*+*fbKS*)	TW(*dbKS*	TW(*fbKS*)
			P	*P*	*P*
War	Baseline	BOW	0.867	0.867	0.867
		POS	0.844	0.844	0.844
		BOE	0.857	0.857	0.857
		BOC	0.839	0.839	0.839
	Resource	Cls(W-Freq)	0.864	0.867	0.873
		parent(Cls)(W-Freq)	0.859	0.862	0.874
		P(W-Freq/Cls)	0.874	0.872	0.869
		Cls+P(W-SG)	0.869	0.880	0.868
		parent(Cls)+P(W-SG)	0.871	0.868	0.873
		P(W-SG/Cls)	0.885	0.885	0.881
	Category	Cat(W-Freq)	0.882	0.882	0.879
		parent(Cat)(W-Freq)	0.887	0.876	0.885
		P(W-Freq/Cat)	0.871	0.871	0.871
		Cat+P(W-SG)	0.871	0.871	NA
		parent(Cat)+P(W-SG)	0.879	0.879	NA
		P(W-SG/Cat)	0.880	0.877	0.878
Cri	Baseline	BOW	0.715	0.715	0.715
		POS	0.667	0.667	0.667
		BOE	0.736	0.736	0.736
		BOC	0.677	0.677	0.677
	Resource	Cls(W-Freq)	0.705	0.714	0.715
		parent(Cls)(W-Freq)	0.716	0.723	0.724
		P(W-Freq/Cls)	0.711	0.712	0.718
		Cls+P(W-SG)	0.709	0.712	0.717
		parent(Cls)+P(W-SG)	0.716	0.709	0.716
		P(W-SG/Cls)	0.729	0.716	0.731
	Category	Cat(W-Freq)	0.694	0.700	0.702
		parent(Cat)(W-Freq)	0.698	0.698	0.693
		P(W-Freq/Cat)	0.701	0.701	0.704
		Cat+P(W-SG)	0.701	0.701	NA
		parent(Cat)+P(W-SG)	0.710	0.710	NA
		P(W-SG/Cat)	0.690	0.686	0.691
DisAcc	Baseline	BOW	0.800	0.800	0.800
		POS	0.746	0.746	0.746
		BOE	0.798	0.798	0.798
		BOC	0.772	0.772	0.798
	Resource	Cls(W-Freq)	0.790	0.800	0.792
		parent(Cls)(W-Freq)	0.793	0.799	0.795
		P(W-Freq/Cls)	0.779	0.793	0.797
		Cls+P(W-SG)	0.799	0.804	0.797
		parent(Cls)+P(W-SG)	0.810	0.804	0.797
		P(W-SG/Cls)	0.808	0.811	0.800
	Category	Cat(W-Freq)	0.786	0.798	0.800
		parent(Cat)(W-Freq)	0.788	0.788	0.788
		P(W-Freq/Cat)	0.796	0.796	0.796
		Cat+P(W-SG)	0.796	0.796	NA
		parent(Cat)+P(W-SG)	0.805	0.805	NA
		P(W-SG/Cat)	0.777	0.795	0.786

Table C.3: Precision results for the single-domain *TW* topic classifiers using different KSs ontologies (DBpedia *dbKS*'s ontologies, and Freebase *fbKS*'s ontology) and two semantic meta-graphs derived from these KSs (*resource meta-graph* (Resource) and *category meta-graph* (Category)).

C.2.1 Results Obtained using Twitter Indicators in Single-domain Classification

This subsection presents the results obtained using Twitter indicator features in single-domain classification. The results in terms of precision can be seen in Figure C.5 and in terms of recall in Figure C.6.

Dataset	Semantic graph	Features	TW(*dbKS+fbKS*)	TW(*dbKS*	TW(*fbKS*)
			R	*R*	*R*
War	Baseline	BOW	0.743	0.743	0.743
		POS	0.757	0.757	0.757
		BOE	0.761	0.761	0.761
		BOC	0.735	0.735	0.735
	Resource	Cls(W-Freq)	0.727	0.736	0.744
		parent(Cls)(W-Freq)	0.734	0.730	0.743
		P(W-Freq/Cls)	0.743	0.739	0.742
		Cls+P(W-SG)	0.746	0.748	0.749
		parent(Cls)+P(W-SG)	0.745	0.745	0.754
		P(W-SG/Cls)	0.777	0.759	0.759
	Category	Cat(W-Freq)	0.763	0.767	0.763
		parent(Cat)(W-Freq)	0.770	0.775	0.764
		P(W-Freq/Cat)	0.759	0.759	0.756
		Cat+P(W-SG)	0.759	0.759	NA
		parent(Cat)+P(W-SG)	0.762	0.762	NA
		P(W-SG/Cat)	0.773	0.767	0.771
Cri	Baseline	BOW	0.521	0.521	0.521
		POS	0.541	0.541	0.541
		BOE	0.534	0.534	0.534
		BOC	0.523	0.523	0.523
	Resource	Cls(W-Freq)	0.518	0.516	0.525
		parent(Cls)(W-Freq)	0.523	0.518	0.523
		P(W-Freq/Cls)	0.525	0.524	0.524
		Cls+P(W-SG)	0.521	0.517	0.522
		parent(Cls)+P(W-SG)	0.522	0.521	0.526
		P(W-SG/Cls)	0.547	0.534	0.532
	Category	Cat(W-Freq)	0.549	0.545	0.538
		parent(Cat)(W-Freq)	0.547	0.547	0.536
		P(W-Freq/Cat)	0.541	0.541	0.535
		Cat+P(W-SG)	0.541	0.541	NA
		parent(Cat)+P(W-SG)	0.543	0.543	NA
		P(W-SG/Cat)	0.551	0.542	0.553
DisAcc	Baseline	BOW	0.637	0.637	0.637
		POS	0.652	0.652	0.652
		BOE	0.670	0.670	0.670
		BOC	0.608	0.608	0.644
	Resource	Cls(W-Freq)	0.636	0.632	0.631
		parent(Cls)(W-Freq)	0.634	0.632	0.635
		P(W-Freq/Cls)	0.620	0.636	0.628
		Cls+P(W-SG)	0.629	0.636	0.637
		parent(Cls)+P(W-SG)	0.629	0.635	0.630
		P(W-SG/Cls)	0.656	0.644	0.646
	Category	Cat(W-Freq)	0.651	0.646	0.639
		parent(Cat)(W-Freq)	0.655	0.655	0.655
		P(W-Freq/Cat)	0.649	0.649	0.642
		Cat+P(W-SG)	0.649	0.649	NA
		parent(Cat)+P(W-SG)	0.650	0.650	NA
		P(W-SG/Cat)	0.662	0.655	0.647

Table C.4: Recall results for the single-domain *TW* topic classifiers using different KSs ontologies (DBpedia *dbKS*'s ontologies, and Freebase *fbKS*'s ontology) and two semantic meta-graphs derived from these KSs (*resource meta-graph* (Resource) and *category meta-graph* (Category)).

C.2.2 Results Obtained using Twitter Indicators in Cross-domain Classification

This subsection presents the results obtained using Twitter indicator features in cross-domain classification. The results in terms of precision can be seen in Figure C.7 and in terms of recall in Figure C.8.

Case	Semantic graph	Features	*DisAcc* P	*Cri* P	*War* P
Full	Baseline	BOW	0.800	0.715	0.867
		BoL(1)	0.801	0.738	0.880
		BoL(L)	0.806	0.734	0.975
		BoL(T)	0.788	0.741	0.881
	Resource	BoH(Cls)	0.783	0.705	0.870
		BoH(P/Cls)	0.793	0.713	0.891
		Cls+BoH(Cls)	0.785	0.707	0.869
		P(SG/Cls)+BoH(P/Cls)	0.847	0.721	0.907
	Category	BoH(Cat)	0.777	0.700	0.872
		BoH(P/Cat)	0.806	0.718	0.882
		Cat+BoH(Cat)	0.783	0.702	0.882
		P(SG/Cat)+BoH(P/Cat)	0.812	0.718	0.892
Filt	Baseline	BOW-Filt	0.877	0.749	0.955
		BoL(1-Filt)	0.801	0.725	0.839
		BoL(L-Filt)	0.801	0.727	0.839
		BoL(T-Filt)	0.813	0.766	0.874
	Resource	BoH(Cls-Filt)	0.810	0.733	0.868
		BoH(P-Filt/Cls)	0.796	0.747	0.892
		Cls+BoH(Cls-Filt)	0.811	0.733	0.868
		P(SG/Cls)+BoH(P-Filt/Cls)	0.817	0.769	0.887
	Category	BoH(Cat-Filt)	0.755	0.664	0.840
		BoH(P-Filt/Cat)	0.754	0.695	0.856
		Cat+BoH(Cat-Filt)	0.835	0.719	0.867
		P(SG/Cat)+BoH(P-Filt/Cat)	0.824	0.755	0.866

Table C.5: Precision results for the single-domain SVM *TW* topic classifier using external *data source indicators*.

Case	Semantic graph	Features	*DisAcc* R	*Cri* R	*War* R
Full	Baseline	BOW	0.637	0.521	0.743
		BoL(1)	0.654	0.540	0.737
		BoL(L)	0.650	0.543	0.829
		BoL(T)	0.636	0.525	0.741
	Resource	BoH(Cls)	0.654	0.558	0.754
		BoH(P/Cls)	0.659	0.562	0.761
		Cls+BoH(Cls)	0.656	0.561	0.761
		P(SG/Cls)+BoH(P/Cls)	0.665	0.561	0.795
	Category	BoH(Cat)	0.657	0.555	0.753
		BoH(P/Cat)	0.654	0.557	0.766
		Cat+BoH(Cat)	0.663	0.557	0.763
		P(SG/Cat)+BoH(P/Cat)	0.660	0.558	0.776
Filt	Baseline	BOW-Filt	0.498	0.400	0.624
		BoL(1-Filt)	0.509	0.474	0.698
		BoL(L-Filt)	0.509	0.474	0.698
		BoL(T-Filt)	0.497	0.488	0.714
	Resource	BoH(Cls-Filt)	0.523	0.488	0.724
		BoH(P-Filt/Cls)	0.515	0.526	0.746
		Cls+BoH(Cls-Filt)	0.526	0.488	0.724
		P(SG/Cls)+BoH(P-Filt/Cls)	0.538	0.517	0.756
	Category	BoH(Cat-Filt)	0.528	0.435	0.688
		BoH(P-Filt/Cat)	0.570	0.516	0.680
		Cat+BoH(Cat-Filt)	0.523	0.482	0.727
		P(SG/Cat)+BoH(P-Filt/Cat)	0.543	0.512	0.749

Table C.6: Recall results for the single-domain SVM *TW* topic classifier using external *data source indicators*.

Case	Semantic graph	Features	*DisAcc* *P*	*Cri* *P*	*War* *P*
Full	Baseline	BOW	0.955	0.944	0.955
	Resource	BoL(1)	0.955	0.943	0.958
		BoL(L)	0.955	0.945	0.958
		BoL(T)	0.953	0.944	0.959
		BoH(Cls)	0.959	0.946	0.964
		BoH(P/Cls)	0.955	0.947	0.958
		Cls+BoH(Cls)	0.960	0.947	0.964
		P(SG/Cls)+BoH(P/Cls)	0.976	0.973	0.989
	Category	BoH(Cat)	0.958	0.946	0.962
		BoH(P/Cat)	0.959	0.948	0.962
		Cat+BoH(Cat)	0.959	0.946	0.962
		P(SG/Cat)+BoH(P/Cat)	0.959	0.948	0.962
Filt	Baseline	BOW-Filt	0.842	0.711	0.956
	Resource	BoL(1-Filt)	0.766	0.917	0.956
		BoL(L-Filt)	0.957	0.914	0.958
		BoL(T-Filt)	0.958	0.917	0.961
		BoH(Cls-Filt)	0.953	0.920	0.964
		BoH(P-Filt)	0.956	0.919	0.960
		BoH(Cat-Filt)	0.954	0.918	0.964
		BoH(P-Filt/Cat)	0.955	0.918	0.962
	Category	Cls+BoH(Cls-Filt)	0.956	0.935	0.964
		P(SG/Cls)+BoH(P-Filt/Cls)	0.956	0.923	0.962
		Cat+BoH(Cat-Filt)	0.954	0.920	0.957
		P(SG/Cat)+BoH(P-Filt/Cat)	0.947	0.918	0.945

Table C.7: Precision results for the DB+FB+TW cross-domain SVM topic classifier using various external *datasource indicators*.

Case	Semantic graph	Features	*DisAcc* R	*Cri* R	*War* R
Full	Baseline	BOW	0.869	0.857	0.861
	Resource	BoL(1)	0.867	0.857	0.871
		BoL(L)	0.866	0.859	0.868
		BoL(T)	0.862	0.854	0.868
		BoH(Cls)	0.979	0.974	0.984
		BoH(P/Cls)	0.900	0.895	0.902
		Cls+BoH(Cls)	0.979	0.973	0.984
		P(SG/Cls)+BoH(P/Cls)	0.885	0.882	0.896
	Category	BoH(Cat)	0.978	0.975	0.984
		BoH(P/Cat)	0.980	0.974	0.986
		Cat+BoH(Cat)	0.979	0.975	0.984
		P(SG/Cat)+BoH(P/Cat)	0.980	0.975	0.986
Filt	Baseline	BOW-Filt	0.409	0.386	0.823
	Resource	BoL(1-Filt)	0.673	0.813	0.827
		BoL(L-Filt)	0.841	0.805	0.824
		BoL(T-Filt)	0.834	0.814	0.822
		BoH(Cls-Filt)	0.986	0.972	0.981
		BoH(P-Filt)	0.876	0.842	0.868
		BoH(Cat-Filt)	0.985	0.965	0.979
		BoH(P-Filt/Cat)	0.986	0.966	0.982
	Category	Cls+BoH(Cls-Filt)	0.985	0.972	0.981
		P(SG/Cls)+BoH(P-Filt/Cls)	0.820	0.801	0.813
		Cat+BoH(Cat-Filt)	0.983	0.956	0.980
		P(SG/Cat)+BoH(P-Filt/Cat)	0.978	0.966	0.980

Table C.8: Recall results for the DB+FB+TW cross-domain SVM topic classifier using various external *datasource indicators*.