Language-Independent Methods for Identifying Cross-Lingual Similarity in Wikipedia

Monica Lestari Paramita

Information School

University of Sheffield

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

February 2019

Declaration

I hereby declare that this thesis is a presentation of my original research work. The contents of this thesis have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions. A number of the chapters in this thesis have been published in academic papers or journals; these publications are specifically indicated at the end of each relevant chapter.

Monica Lestari Paramita
February 2019

Abstract

The diversity and richness of multilingual information available in Wikipedia have increased its significance as a language resource. The information extracted from Wikipedia has been utilised for many tasks, such as Statistical Machine Translation (SMT) and supporting multilingual information access. These tasks often rely on gathering data from articles that describe the same topic in different languages with the assumption that the contents are equivalent to each other. However, studies have shown that this might not be the case.

Given the scale and use of Wikipedia, there is a need to develop an approach to measure cross-lingual similarity across Wikipedia. Many existing similarity measures, however, require the availability of 'language-dependent' resources, such as dictionaries or Machine Translation (MT) systems, to translate documents into the same language prior to comparison. This presents some challenges for some language pairs, particularly those involving 'under-resourced' languages where the required linguistic resources are not widely available. This study aims to present a solution to this problem by first, investigating cross-lingual similarity in Wikipedia, and secondly, developing 'language-independent' approaches to measure cross-lingual similarity in Wikipedia.

Two main contributions were provided in this work to identify cross-lingual similarity in Wikipedia. The first key contribution of this work is the development of a Wikipedia similarity corpus to understand the similarity characteristics of Wikipedia articles and to evaluate and compare various approaches for measuring cross-lingual similarity. The author elicited manual judgments from people with the appropriate language skills to assess similarities between a set of 800 pairs of interlanguage-linked articles. This corpus contains Wikipedia articles for eight language pairs (all pairs involving English and including well-resourced and under-resourced languages) of varying degrees of similarity.

The second contribution of this work is the development of language-independent approaches to measure cross-lingual similarity in Wikipedia. The author investigated the utility of a number of "lightweight" language-independent features in four different experiments. The first experiment investigated the use of Wikipedia links to identify and align similar sentences, prior to aggregating the scores of the aligned sentences to represent the similarity of the document pair. The second experiment investigated the usefulness of content similarity features (such as char-n-gram overlap, links overlap, word overlap and word length ratio). The third experiment focused on analysing the use of structure similarity features (such as the ratio of section length, and similarity between the section headings). And finally, the fourth experiment investigates a combination of these features in a classification and a regression approach.

Most of these features are language-independent whilst others utilised freely available resources (Wikipedia and Wiktionary) to assist in identifying overlapping information across languages. The approaches proposed are lightweight and can be applied to any languages written in Latin script; non-Latin script languages need to be transliterated prior to using these approaches. The performances of these approaches were evaluated against the human judgments in the similarity corpus.

Overall, the proposed language-independent approaches achieved promising results. The best performance is achieved with the combination of all features in a classification and a regression approach. The results show that the Random Forest classifier was able to classify 81.38% document pairs correctly (F_1 score=0.79) in a binary classification problem, 50.88% document pairs correctly (F_1 score=0.71) in a 5-class classification problem, and RMSE of 0.73 in a regression approach. These results are significantly higher compared to a classifier utilising machine translation and cosine similarity of the tf-idf scores. These findings showed that language-independent approaches can be used to measure cross-lingual similarity between Wikipedia articles. Future work is needed to evaluate these approaches in more languages and to incorporate more features.

Acknowledgements

Firstly, I would like to thank my first supervisor, Professor Paul Clough, for his expertise and continuous support throughout my entire study. Thank you for always finding time to provide valuable feedback on my research and for encouraging me to be a better researcher. I would also like to thank my second supervisor, Professor Rob Gaizauskas, for providing insightful comments on my study.

My PhD was partially funded by the ACCURAT project supported by the European Commission and I gratefully acknowledge this support. Thank you also to my colleagues in the ACCURAT project who have participated in the evaluation task.

To my colleagues, Ahmet, Emina, Emma, Mengdie, Paula, Sophie, Alberto, Nikos, Simon, and all the members of the IR and NLP group in the University of Sheffield, and to my wonderful friends from the writing club, Akanimo and Anna, thank you very much for the motivation, encouragement, peer support and all the feedback you have given me. I would also like to thank Kay Guccione, Sarah Barnes (my career mentor) and, especially Hannah Roddie (my thesis mentor), whose advice and encouragement have built up my confidence to finish my thesis.

Finally, my profound gratitude goes to my family and friends for their endless support and motivation during my years of study. To my wonderful husband, Paul, who has never stopped believing in me. To my son, Oliver, who always gives me a reason to laugh and be grateful every day, and to my baby bump, whose kicks have kept me accompanied (and awake) during the late night writing sessions. Finally, to my wonderful parents and parents in law who have loved, supported and encouraged me throughout this process. This thesis would not have been possible without their support.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Wikipedia is a significant language resource due to the richness of information of diverse topics in different languages. Wikipedia articles that discuss the same topics in different languages have been assumed to be equivalent and utilised for many linguistic and translation tasks, such as Statistical Machine Translation (SMT), text classification, etc. Previous studies, however, have identified that the contents of these Wikipedia articles can vary significantly. This thesis aims to investigate the degree of similarity in Wikipedia articles and to develop language-independent approaches for measuring cross-lingual similarity in Wikipedia.

In this chapter, the author describes the motivation for measuring cross-lingual similarity in Wikipedia (Section 1.1), defines the aims and objectives of the thesis and lists the research questions (Section 1.2), and identifies main contributions of this thesis (Section 1.3). An outline of this thesis is provided in Section 1.4.

1.1 Motivation

Measuring content similarity between texts is an important task in various fields. For example, in information retrieval (Manning, Raghavan, & Schütze, 2008), similarity measures are applied to identify relevant documents for a user-given query. Similarity measures are also used in identifying text reuse (Clough, Gaizauskas, Piao, & Wilks, 2002), text classification (Y. Lin, Jiang, & Lee, 2014), plagiarism detection (Maurer, Kappe, & Zaka, 2006), and identifying partially or entirely duplicated documents on the Web (Brin, Davis, & Garcia-Molina, 1995). Another task relying on similarity measures is clustering or categorising documents (Bigi, 2003), in which similarity measures are used to identify documents that describe the same topic in order to cluster them together.

To measure *monolingual similarity* (i.e., between texts written in the same language), a simple approach is to calculate the proportion of words or word sequences that are shared by both documents.[1] However, when measuring similarity between documents written in different languages (also referred to as *cross-lingual similarity*), a different approach is required to identify the shared content across the different languages. One approach is to translate the documents into the same language prior to identifying the overlap of words.[2] To perform this process, linguistic resources, such as bilingual dictionaries or Machine Translation (MT) systems, are required. However, these requirements present a problem for some languages, for which the availability of these linguistic resources are limited; these languages are further referred to as *under-resourced languages*. Additionally, when an MT system is available, the translation qualities for these under-resourced languages are generally poorer than those for high-resourced languages (Skadiņa et al., 2012) and, therefore, are expected to reduce the accuracy of the monolingual similarity methods used on the translated contents (Gao et al., 2001).

Measuring cross-lingual similarity is important for a number of tasks, such as performing cross-language information retrieval (Peters, Braschler, & Clough, 2012; Vulić & Moens, 2015), multilingual document clustering (Yapomo, Bernhard, & Gançarski, 2015) and automatic classification of multilingual articles (Gupta, Barrón-Cedeno, & Rosso, 2012; Mogadala & Rettinger, 2016; Steinberger, Pouliquen, & Hagman, 2002). Cross-lingual

[1] This approach can be used to identify syntactical similarity, meaning the similarity of strings. Semantic relatedness, on the other hand, defines the likeness of meaning rather than the syntactical similarity and therefore cannot be easily identified by using a simple word overlap. Instead, it requires linguistic knowledge for the language, such as hypernym and hyponym (i.e. words with an is-a relation, e.g. "car" and "vehicle").

[2] Methods do exist whereby consecutive sequences of characters or words (referred to as n-grams) are used to perform cross-lingual comparison without the need for translating the contents. However, these approaches can only be used with languages with similar writing systems (e.g. alphabets), such as English and French, and would not work with languages using different alphabets, such as English and Arabic.

similarity approaches have also been utilised for building multilingual linguistic resources from the Web, such as bilingual lexicons and dictionaries (Erdmann, Nakayama, Hara, & Nishio, 2008; Sadat, 2010; Štajner & Mladenić, 2018), and more recently, creating multilingual corpora (Koehn, 2009; Munteanu & Marcu, 2005; Saad, Langlois, & Smaïli, 2014; Skadiņa et al., 2012).

Multilingual corpora are significant resources for many linguistic tasks. *Parallel corpora* are defined as collections of multilingual documents that are translated sentence by sentence. They are valuable resources for building a Statistical Machine Translation system (Koehn, 2009). However, these corpora are often only available for limited languages and domains (mostly legal documents) (Koehn, 2005; Skadiņa et al., 2012). Previous studies have therefore utilised 'similar' or 'comparable' (instead of translated) documents as a solution to enhance linguistic resources for languages with limited numbers of parallel documents (Maia, 2003; McEnery & Xiao, 2007; Munteanu, Fraser, & Marcu, 2004; Skadiņa et al., 2012). These multilingual texts are not translations of each other. However, their contents are related and may contain overlapping information that could be valuable for building multilingual resources (Munteanu & Marcu, 2005). These collections are referred to as *comparable corpora* (Otero & López, 2010; Skadiņa et al., 2012; Tomás, Bataller, Casacuberta, & Lloret, 2008).

A variety of web sources have been previously mined for building comparable corpora (Skadiņa et al., 2012), including Wikipedia. Wikipedia is one of the most popular web sources for extracting multilingual articles, due to its diversity and richness of information provided in a large number of languages (Tomás et al., 2008). Wikipedia has seen a drastic growth over the past decade,[3] containing at the time of writing over 44.5 million articles written in 298 languages[4] and covering a wide variety of topics and domains. As the largest online encyclopaedia, it is one of the most valuable linguistic resources on the Web. Another advantage of Wikipedia is that multilingual articles describing the same topic are linked to each other using Wikipedia *interlanguage links* (these linked articles

[3] `http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth` (accessed on 7 August 2013)

[4] Data collected in March 2018 from `http://meta.wikimedia.org/wiki/List_of_Wikipedias`

are further referred to as *interlanguage-linked articles*). These links provide an alignment at the document level of Wikipedia articles written on different languages on the same topic. These document-level alignment information is not often available, nor easily produced, when collecting multilingual documents from other web sources, such as news sites or blogs. These advantages have made Wikipedia a significant source for various IR and NLP tasks, such as building multilingual corpora (Ion, Tufiş, Boroş, Ceauşu, & Ştefănescu, 2010; Sadat, 2010), extracting multilingual resources such as sentences or terms (Adafre & de Rijke, 2006; Erdmann, Nakayama, Hara, & Nishio, 2009; Smith, Quirk, & Toutanova, 2010; Tomás et al., 2008; Yasuda & Sumita, 2008), and to support multilingual tasks such as computing semantic relatedness (Milne, 2007; Potthast, Stein, & Anderka, 2008).

Since interlanguage-linked Wikipedia articles describe the same topic in different languages, various studies (Mohammadi & GhasemAghaee, 2010; Otero & López, 2010; Sadat, 2010) have assumed that the content of these interlanguage-linked articles are the same. These articles have been used to create bilingual corpora for supporting tasks, such as computing semantic relatedness between documents (Potthast et al., 2008) or terms (Milne, 2007). Other studies, however, have found that the degrees of similarity between these articles may vary widely, and in some cases, may even contain contradictory information (Filatova, 2009; Patry & Langlais, 2011). It is important to be able to compute the similarity of all interlanguage-linked articles in Wikipedia, in order to support future applications that rely on these resources.

Methods to measure cross-lingual similarity often rely on some language-specific linguistic resources, such as dictionaries or translation systems resources (Agirre et al., 2009; Munteanu & Marcu, 2005; Uszkoreit, Ponte, Popat, & Dubiner, 2010; Yasuda & Sumita, 2008). Although these resources are widely available for highly-resourced languages, the availability for these resources are scarce for some languages, particularly under-resourced languages (Argaw & Asker, 2005; Skadiņa et al., 2012). Given the large number of languages in Wikipedia, there is a need for methods to measure similarity without requiring any resources. These methods are further referred to as *language-independent* *methods*. The development of language-independent methods to accurately quantify the similarity between interlanguage-linked articles is the focus of this study.

1.2 Research questions and objectives

To date, there is a lack of studies that have identified the similarity characteristics between Wikipedia articles. Understanding these characteristics is essential in developing suitable methods for identifying similarity in Wikipedia. Moreover, there is a lack of an evaluation benchmark that allows the different techniques to measure cross-lingual similarity to be evaluated automatically. Furthermore, little work has also been undertaken to develop language-independent methods to compute similarity between Wikipedia articles written in different languages. Therefore, this study intends to address these gaps by first identifying the different characteristics of documents that contributes to their similarity in Wikipedia, and to create an evaluation benchmark to enable the automatic evaluation of the methods. This study also intends to develop language-independent approaches for measuring similarity in Wikipedia.

This study aims to answer three research questions:

RQ1. What are the characteristics of similar interlanguage-linked articles in Wikipedia?

RQ2. Can we create an evaluation benchmark for Wikipedia? I.e., do human assessors agree on Wikipedia similarity?

RQ3. Can language-independent approaches be used to identify cross-lingual similarity in Wikipedia? This thesis aims to investigate four different approaches. For each experiment, the following sub-questions are investigated:

(a) How does the method compare to approaches using linguistic resources, such as MT systems?

(b) How does the performance for the approach vary for different language pairs?

(c) What language-independent features are best for measuring cross-lingual similarity in Wikipedia?

There are two research objectives in this work:

1. To develop an evaluation benchmark containing human judgments on the similarity of interlanguage-linked Wikipedia articles.
2. To develop language-independent techniques to measure cross-lingual similarity across Wikipedias.

1.2.1 Scope of the research

This work aims to investigate methods to compute cross-lingual similarity between pairs of interlanguage-linked Wikipedia articles. Approaches for measuring cross-lingual similarity without using linguistic resources are investigated in this work. Their performances are compared to those using language-dependent resources, specifically Machine Translation (MT). These approaches are evaluated against a newly created evaluation benchmark containing human judgments regarding cross-lingual similarity in Wikipedia. In this section, the author specify the scope of this research study.

Firstly, most of the features investigated in this study are limited to languages based on the Latin script. Languages containing different script, such as Greek, need to be transliterated if such tools exist prior to using these approaches.

To assist with identifying overlapping information across languages, the author also investigated approaches that utilised Wikipedia and Wiktionary. These sources were selected because they were available in a large number of languages. The study also focused on proposing measures that are lightweight and easy to extract.

The focus of this study is to develop approaches for measuring cross-lingual similarity in Wikipedia. The use of these approaches for identifying cross-lingual similarity in other sources is not investigated in this study. Furthermore, the approaches proposed in this study focus on measuring similarity at the document level. They do not aim to identify specific fragments that were similar within the document pairs.

Finally, Wikipedia is utilised for many linguistic tasks, such as extracting bilingual terms and building comparable corpora. Cross-lingual similarity approaches to measure similarity in Wikipedia can therefore be used to identify similar articles and to filter out dissimilar articles in order to more efficiently support these tasks. Analysing the improvement of these tasks is not in the scope of this work. However, previous studies have shown that using a higher comparable corpus (i.e., corpus with highly similar documents) improve the performance of these tasks (Li & Gaussier, 2010; Skadiņa et al., 2012).

1.3 Main contributions to the research field

This work contributes to the research area in two ways. Firstly, this work creates *an evaluation benchmark of Wikipedia similarity*. At the time of writing, the author is not aware of any existing sets of Wikipedia articles that have been manually evaluated based in their content similarities. The evaluation benchmark created in this work, therefore, provides a useful resource for research, especially to further identify the characteristics of similar documents in Wikipedia and to evaluate new similarity approaches.

Secondly, this work investigates *the development of language-independent approaches to identify cross-lingual similarity in a variety of languages*.[5] Numerous methods which have been performed in identifying cross-lingual similar fragments, such as translated sentences, in Wikipedia articles often require the use of linguistic resources (such as MT systems or dictionaries) to translate them to a language (e.g., English) prior to measuring their similarities monolingually. These methods, therefore, cannot be applied to languages with no available translation resources. Most of the language-independent techniques developed in this work do not require any linguistic resources and therefore do not have these limitations. A small number of features developed in this work utilise widely available multilingual sources, i.e., Wikipedia and Wiktionary, as translation resources. The approaches proposed in this study, therefore, can be applied to a large number of languages.

[5] The methods developed in this study are able to compute similarity in all languages that used Latin-based characters. Meanwhile, languages with other characters, such as Arabic, Chinese, or Greek, will require a tool to transliterate them into Latin-based characters prior to measuring similarity using these methods.

1.4 Thesis outline

This thesis contains 11 chapters which describe the work involved in this study. A brief overview summarising each chapter is shown below:

Chapter 1: Introduction

In this chapter, the author describes the motivation of the work followed by the research objectives and research questions. The main contributions of this work are also listed in this chapter.

Chapter 2: Related Work

This chapter reports the literature review in this area, which includes previous works in defining similarity and the development of techniques to identify similarity, both monolingually and cross-lingually. Previous work in extracting Wikipedia documents and investigating similarity of Wikipedia are also described in this section, followed by a summary identifying the gap in the literatures which this work aims to fill.

Chapter 3: Methodology

This chapter reports the methodology of this work and describes the selection of languages used in this research, the corpus pre-processing phase and the statistics of Wikipedia dataset for each language. A summary of the methods proposed in this study are also described in this chapter.

Chapter 4: Identifying Similarity Features in Wikipedia

In this chapter, the author describes the initial study of analysing Wikipedia similarity. A small number of documents are analysed to identify the type of similarity and dissimilarity that occur in Wikipedia articles.

Chapter 5: Evaluation Corpus

In this chapter, the author describes the creation of evaluation benchmark, which includes the selection of documents and the creation of evaluation guidelines. Human judgments gathered in this task are reported and analysed to further understand the characteristics of similarity of Wikipedia interlanguage-linked articles.

Chapter 6: Anchor Text and Word Overlap Method

This chapter describes the first approach, which identifies similar sentences using an adaptation of the link-based bilingual lexicon approach (Adafre & de Rijke, 2006). Similarity is measured at the sentence level, prior to aggregating at the document level.

Chapter 7: Content Similarity Features

In this chapter, a number of similarity features are extracted from the main contents of the articles. Evaluation is performed by measuring the correlation between each feature to human judgment of similarity.

Chapter 8: Structure Similarity Features

A different approach is investigated in this chapter, by identifying similarity between Wikipedia articles by analysing only the structure (i.e., section titles) of the articles. Similarly, the evaluation corpus is used to report the correlation between these features to human judgments.

Chapter 9: Classification of Similar Documents

In the final experiment, all the features identified in Chapter 6 to Chapter 8 are combined into a classification approach. These features are used to build a binary similarity classifier, and a multi-class similarity classifier. Using these features in a regression approach is also investigated in this chapter. The evaluation corpus is used as a gold-standard, and a 10-fold cross-validation is reported.

Chapter 10: Discussion

In this chapter, the author reflects on the contributions of this study and how they relate to existing literature. Possible applications of the methods proposed in this study are also identified in this chapter.

Chapter 11: Conclusion and Future Work

Finally, the author summarises the contributions of this study and answers the research questions in this chapter. Limitations of this study and recommendations on future avenues of research in this area are also described.

1.5 Structure of the work

Figure 1.1 shows the different components investigated in this study and how they relate to each other. These components can be categorised into two main tasks. The first one is to *understand the similarity in Wikipedia*. In this first task, the author reviewed related work in the area (Chapter 2) and carried out an initial study on Wikipedia similarity (Chapter 4). Based on these findings, the author created an annotation task to gather human judgments on similarity (Chapter 5). A pilot study was conducted prior to carrying out the final study, in which human judgments for 800 document pairs in 8 language pairs were gathered. These annotations are referred to as the evaluation corpus.

The second task in this study is to *develop approaches to measure cross-lingual similarity*. The findings from the first task were used to inform a set of features that are valuable for measuring cross-lingual similarity in Wikipedia. Four different experiments were then carried out to develop approaches using these different selection of features (Chapter 6-9). The performance of these approaches were evaluated against the evaluation corpus. The remainder of the thesis discusses these findings and how they relate to the existing literature (Chapter 10). Finally, the last chapter concludes the work (Chapter 11).

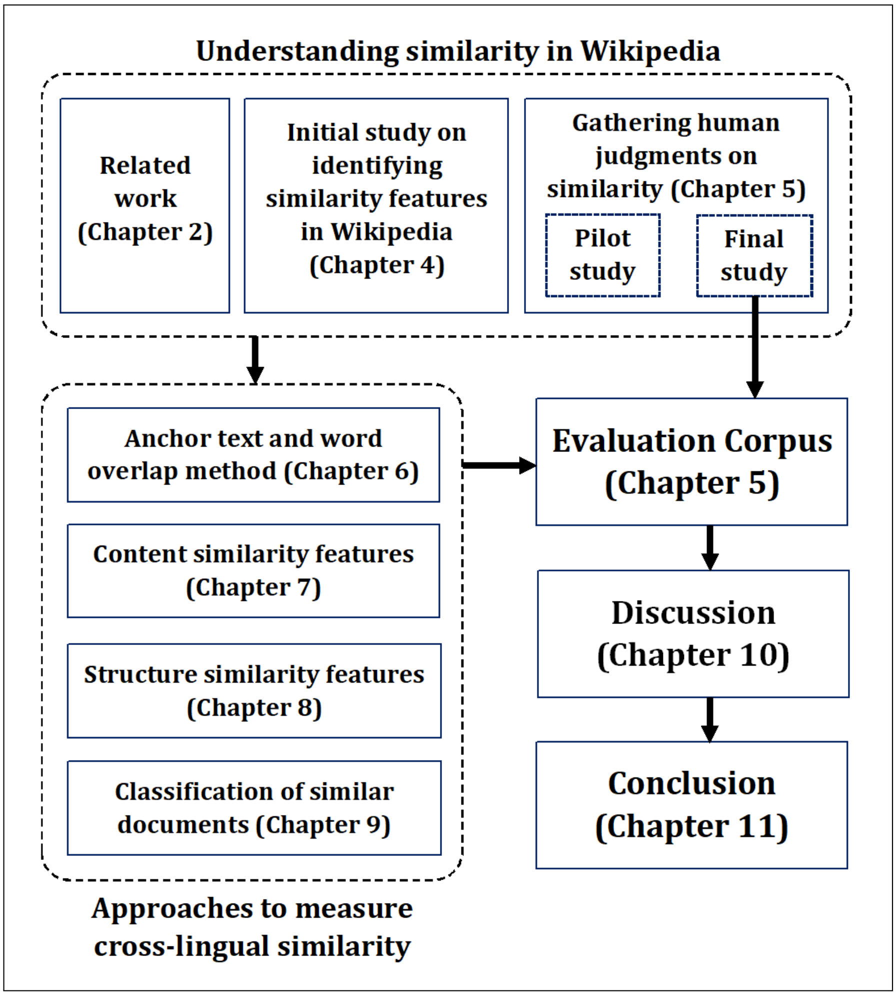

Understanding similarity in Wikipedia
Related work (Chapter 2)
Initial study on identifying similarity features in Wikipedia (Chapter 4)
Gathering human judgments on similarity (Chapter 5)
Pilot study
Final study
Anchor text and word overlap method (Chapter 6)
Content similarity features (Chapter 7)
Structure similarity features (Chapter 8)
Classification of similar documents (Chapter 9)
Approaches to measure cross-lingual similarity
Evaluation Corpus (Chapter 5)
Discussion (Chapter 10)
Conclusion (Chapter 11)

Fig. 1.1 Structure of work

Chapter 2

Related Work

Chapter 1 has described the motivation of developing methods to compute cross-lingual similarity in Wikipedia. In this chapter, the author reviews previous studies that have been conducted in this area.

Firstly, the author reviewed studies that were aimed at measuring similarity to identify the different tasks that relied on measuring similarity (Section 2.1) and how similarity was defined in previous work (Section 2.2). Previous literature has also developed approaches to measure similarity between texts written in the same language (*monolingual similarity*) and texts between different languages (*cross-lingual similarity*). The author reviewed these monolingual similarity and cross-lingual similarity approaches in Section 2.3 and Section 2.4, respectively.

The aim of this thesis is to measure similarity in Wikipedia. To further identify applications that benefit from measuring similarity in Wikipedia, the author reviewed previous work that utilised Wikipedia as a linguistic resource in Section 2.5. Previous studies that specifically analysed the degree of similarity (and dissimilarity) in Wikipedia are then highlighted in Section 2.6. Finally, the gap in literature that this work aims to fill is identified in Section 2.7.

2.1 Measuring similarity

Measuring similarity between texts is an important task for many fields, such as information retrieval (Manning et al., 2008), plagiarism detection (Maurer et al., 2006), clustering (Bigi, 2003; A. Huang, 2008) and text classification (Wu et al., 2017). These different tasks, however, require similarity to be measured at different granularities. In information retrieval, for example, similarity measures are used to compute the relevance between a query (usually a few words) and a collection of documents in order to retrieve the most relevant documents to the query (Manning et al., 2008). Similarity between sentences are investigated for the purpose of identifying text reuse, both to identify monolingual text reuse (Clough et al., 2002; Hoad & Zobel, 2003; Maurer et al., 2006; Shivakumar & Garcia-Molina, 1995) and cross-lingual text reuse (Potthast, Barrón-Cedeño, Stein, & Rosso, 2011). Sentence similarity methods have also been investigated for the purpose of identifying paraphrases (Mihalcea, Corley, & Strapparava, 2006), semantic textual similarity between a pair of sentences (Agirre et al., 2012; Bär, Biemann, Gurevych, & Zesch, 2012) and textual entailment (Dagan, Glickman, & Magnini, 2006). Textual entailment task aims to identify whether information in one sentence can be inferred by the information in another sentence and has been utilised for the purpose of content synchronisation between two document versions (Mehdad, Negri, & Federico, 2010; Vilarino, Pinto, Tovar, León, & Castillo, 2012; Wäschle & Fendrich, 2012). Identifying similar sentences is also a valuable task for summarising text, in order to identify and include diverse information into the summary (Do, Roth, Sammons, Tu, & Vydiswaran, 2009).

Meanwhile, tasks such as clustering or classification tasks, often measure similarity between documents instead (A. Huang, 2008; Wu et al., 2017). Measuring similarity at the document level has also been carried out in the news domains, such as for tracking related news articles, both monolingually (M. D. Lee, Pincombe, & Welsh, 2005) and cross-lingually (Pouliquen, Steinberger, Ignat, Käsper, & Temnikova, 2004). Document similarity has also been measured specifically for academic publications in previous work (Elsayed, Lin, & Oard, 2008; Lakkaraju, Gauch, & Speretta, 2008; Trivison, 1987) in order to suggest similar publications to readers and to investigate relations between cited and citing articles. Its application in the Web domain has also been researched for the purpose of finding similar documents (Cooper, Coden, & Brown, 2002), near-duplicate documents (Shivakumar & Garcia-Molina, 1995) and finding translated documents in the Web (Resnik & Smith, 2003).

This last work was aimed for creating bilingual *parallel corpora* to be used as translation resources. However, in the past decades, the research has further progressed to measuring cross-lingual similarity for the purpose of finding similar (yet non-parallel) articles across languages for building a corpus of comparable documents, or more frequently referred to as *comparable corpora* (Maia, 2003; Skadiņa et al., 2012). Similar to parallel corpora, comparable corpora have also been utilised as translation resources because they have wider availability than parallel corpora for languages and domains that are under-resourced.

Approaches to measure similarity (which are further reviewed in Section 2.3 and Section 2.4) can rely on measuring syntactical similarity between two texts, such as by measuring the overlap of words between the texts. However, many have identified limitations of these methods since similar texts may not use the same words. A large number of studies have aimed at identifying semantic similarity (i.e., similarity of meanings) between words or concepts (J. J. Jiang & Conrath, 1997; Y. Jiang, Zhang, Tang, & Nie, 2015; Kandola, Cristianini, & Shawe-Taylor, 2003; Lakkaraju et al., 2008; Pedersen, Patwardhan, & Michelizzi, 2004; Taieb, Aouicha, & Hamadou, 2014). These approaches often require the use of a lexical database, such as WordNet (Miller, 1995), or a large corpus to learn the co-occurrence between related words or concepts (Agirre et al., 2009). These approaches have been further utilised in measuring the semantic similarity between two texts (Wan & Peng, 2005).

2.2 Defining the concept of similarity

Although a large number of studies have aimed to measure similarity between texts (previously described in Section 2.1), there is no consensus on the definition of similarity.

Instead, various studies have defined similarity differently based on the tasks. Furthermore, the different degrees of similarity have also been defined differently depending on various factors, such as the tasks (e.g., search tasks, semantic similarity task, etc.), the domains (e.g., Web articles, news articles, etc.), and the granularity of the applications (e.g., at the document level, sentence level, etc.). In this section, the author reviewed literature that have described the concept 'similarity' and how the different levels of similarity have been defined.

In information retrieval tasks, similarity is often measured between a query and a document (Manning et al., 2008). Similarity is therefore defined as the *relevance* between a document to the given query. Human annotations on relevance have often used binary judgments (e.g., "not relevant" and "relevant") or graded relevance judgments, such as a 3-point relevance used in the Web Track TREC (Voorhees, 2001). Milios, Zhang, He, and Dong (2003) also used a 3-point relevance for identifying the relevance of a document to a query: 0 ("not related"), 0.5 ("somewhat related") and 1 ("related"). In contrast, Paepcke, Garcia-Molina, Rodriguez-Mula, and Cho (2000) argued that the relevance of a document to a query is not the same to the similarity of the document to the given query. Instead, it should be more linked to the "information value" that was given by the document to the users.

The study of defining similarity has also been carried out extensively in the news domains. E.g., in the work in topic detection or tracking, news articles were annotated based on their similarity to a particular topic (Allan, Carbonell, Doddington, Yamron, & Yang, 1998). In developing the TDT3 (Topic Detection and Tracking) corpus for this work, annotators were asked to annotate the similarity between 9,000 news articles and 120 topics, specifying whether each article was "on-topic" or "off-topic" to a given topic. News stories that were annotated to be within the same topic were further assumed to be similar and relevant, whilst the rest were considered to be irrelevant and dissimilar.

Braschler and Schäuble (1998), on the other hand, argued that a more fine-grained scheme is required to identify the similarity between two news articles and proposed five similarity classes to describe the different alignments of retrieved multilingual news documents. The first class, '*Same story*', represents two documents covering exactly the same story or event (e.g., the presidential election results for the same candidate). Documents covering different yet related events (e.g., election results for two different candidates) are categorised into the second class, '*Related story*'. The third class, '*Shared aspect*', represents two documents addressing multiple topics but sharing at least one of them (e.g., one document about updates on US politics, and another about the upcoming presidential election). Unrelated documents which share a large number of terms (e.g., one document about the US presidential election and another document about the French presidential election) are categorised into the fourth class, '*Common terminology*'. Lastly, the class '*Unrelated*' represents two documents with no apparent relation (e.g., one about the presidential election and one about vacation traffic in Germany).

Similar to Braschler and Schäuble (1998), M. D. Lee et al. (2005) also used a 5-point Likert Scale (1=highly unrelated, 5=highly related) to gather human annotations on the similarity of news articles. However, no definition was provided to define the different levels. In the work of tracking similar news, Pouliquen et al. (2004) proposed a four-point scheme for defining similarity between news articles: "same news story", "interlinked news story" (e.g., Madrid bombing vs Spanish decision to pull troops out of Iraq), "loosely connected story" (e.g., documentary on drinking vs alcohol policy), and "wrong link".

Meanwhile, Barker and Gaizauskas (2012), whose work focused in identifying cross-lingual information between news articles, argued that news articles describing the *same event* could differ widely in content if they had different *focal events* (i.e., focus of the story). For example, articles describing a particular flood (the same news event) may have different focal points, such as the flood victims, the rescue efforts, or the disaster aid information. These differences will directly affect the amount of shared contents across the multiple news texts. To accommodate these issues, Barker and Gaizauskas (2012) created a two-level news relatedness scheme that categorised articles based on both the news events *and* the focal events. A comparison between this scheme and the previous two literature is summarised in Table 2.1.

The scheme proposed in Barker and Gaizauskas (2012) is more fine-grained than oth-

Table 2.1 A comparison of similarity in the news domains

Allan et al. (1998)	**Braschler and Schäuble (1998)**	**Barker and Gaizauskas (2012)**
Similar and relevant represent documents that discuss the same topic.	*Same story* represents documents covering exactly the same story or event.	*Same news events - same focal events* represents documents covering the same news event and the same focus of the story.
		Same news events - different focal events represents documents covering the same news event but different focus of the story.
Dissimilar and irrelevant represent documents that discuss different topics.	*Related story* represents documents covering different yet related events.	*Different news events (same type) - focal events (same type)* represents documents describing different topics of the same type (e.g., news about different hurricanes) and the same type of focus of the story.
		Different news events (same type) - focal events (different type) represents documents describing different topics of the same type but having different focus of the story.
	Shared aspects represents documents addressing multiple topics but sharing at least one of them.	*Different news events (different type) - related via background* represents documents describing different news events but share the same background (e.g., the same previous events, people or places).
	Common terminology represents unrelated documents that still share a large number of terms.	
	Unrelated represents two documents with different topics and no apparent relation.	*Different news events, different type - other* represents articles with no similar contents.

ers as it focused specifically on identifying shared content across languages rather than the similarity of the topic discussed in the articles in general. Similar tasks have also been performed to identify similar Web articles for the purpose of building comparable corpora for enhancing resources for under-resourced languages (Maia, 2003; McEnery & Xiao, 2007; Munteanu et al., 2004; Skadiņa et al., 2012). In these tasks, similarity is measured cross-lingually for the purpose of retrieving alignable fragments from bilingual documents (e.g., such as translated sentences or words). This specific aspect of similarity is often referred to as *comparability* (Fung & Cheung, 2004; Tomás et al., 2008).[1]

In assessing comparability between two documents, terms such as '*parallel*' and '*comparable*' have been used to represent the different proportion of translated sentences found in the document pair. 'Parallel' documents are defined to be a pair of documents which have been translated sentence-by-sentence (Fung & Cheung, 2004; Skadiņa et al., 2012; Tomás et al., 2008). Fung and Cheung (2004) also used a comparability level named '*noisy parallel*', to represent parallel documents with insertion or deletion which resulted in non-aligned sentences.

Documents which are similar yet do not correspond in a sentence-by-sentence translation, meanwhile, are often referred to as 'comparable' documents. The definitions of comparable documents, however, vary in different studies. Tomás et al. (2008) described comparable documents as a pair of documents which were not parallel but contained some translated sentences. Meanwhile, Fung and Cheung (2004) defined comparable documents as documents with no aligned sentences but containing the same topic. Meanwhile, the ACCURAT (Analysis of Comparable Corpora for Under Resourced Languages for machine Translation) project[2] (Skadiņa et al., 2012) noted that comparable documents could be further categorised into two classes, namely '*strongly comparable*', and '*weakly comparable*'. Strongly comparable documents were texts containing the same subject and having the same source, while weakly comparable documents repre-

[1] Although the term 'comparability' has mostly been used for cross-lingual tasks, it has also been used to represent monolingually similar documents. E.g., in their work, Barzilay and Elhadad (2003) referred to a corpus of rewriting examples in the same language as a monolingual comparable corpus.

[2] http://www.accurat-project.eu

Table 2.2 A comparison of comparability in Web articles

Fung and Cheung (2004)	**Tomás et al. (2008)**	**Skadiņa et al. (2012)**
Parallel represents texts which are translated sentence by sentence.	*Parallel* represents texts which are translated sentence by sentence (preserving the sentence order).	*Parallel* represents texts which are accurate translations, or approximate translations with some addition or omissions.
Noisy parallel represents texts which are mostly parallel but contain non-aligned sentences which may be caused by paragraph insertions or deletions.	*Comparable* describes texts that contain a noticeable number of translated sentences.	
Comparable describes texts which do not contain aligned sentences but are about the same topic.	*Unspecified*	*Strongly comparable* represents texts coming from the same source or containing the same subject.
Non parallel represents disparate bilingual documents which may or may not be in the same topic.		*Weakly comparable* represents texts in the same domain but different events.
		Not comparable

sented texts describing different events but still in the same domain.

Different terms have been used to categorise the least similar documents. Skadiņa et al. (2012) proposed a class named ‘not comparable’ to classify documents with no similarity. Meanwhile, Fung and Cheung (2004) named this category ‘non-parallel‘ and defined it as dissimilar bilingual documents which may or may not be in the same topic. A comparison of these different comparability levels is shown in Table 2.2.

The terms ‘parallel’ and ‘comparable’ have also been used in representing degrees of similarity of *sets of documents in a corpus (or corpora)*, rather than between two documents. Parallel corpora are identified as sets of parallel texts, i.e., bilingual texts which are translated sentence by sentence (Fung & Cheung, 2004; Skadiņa et al., 2012). Comparable corpora, on the other hand, have been defined differently in various studies. Zanettin (1998) defined comparable corpora as sets of bilingual texts which shared similar criteria of composition, genre and topic; meanwhile, Munteanu and Marcu (2005) defined comparable corpora not by the similarity of topics, but instead as “bilingual texts that, while not parallel in the strict sense, are somewhat related and convey overlapping information” (Munteanu & Marcu, 2005, p. 477).

Identifying similarity based on the proportion of similar sentences between the documents has also been explored in the context of identifying near-duplicate monolingual articles. In this work, Cooper et al. (2002)[p. 246] defined similar articles as “those in which a large percentage of the sentences, or words in the sentences, are the same”. They also defined duplicate documents as “ones that have essentially the same words in the same sentences and paragraphs” although they were allowed to be “in a somewhat different order” (Cooper et al., 2002, p. 246).

Similar scheme was proposed by Brants and Stolle (2002) who also differentiated the degrees of similarity between two texts based on the amount of syntactical similarity (i.e., overlap of words or sentences) shared between the texts. They referred to this concept as *surface similarity*. Different to previous works, their work focused on measuring similarity between troubleshooting manuals for photocopiers for the purpose of reducing redundant information found in search. In this work, a three-point scale was used to describe the different degrees of surface similarity between two documents: ‘same’ to represent identical or almost identical documents, ‘similar’ to represent cases where one document may use different words or synonyms and different order of sentences, and ‘different’ to represent cases where the texts were different. In the same work, they also proposed another dimension of similarity which took into account the semantic similarity within the documents; they referred to this dimension as *conceptual similarity*. Four-point scale was used in this work: ‘same’ (i.e., documents with (almost) the same contents and may include paraphrasing), ‘similar’ (i.e., documents with significant overlap of conceptual contents, e.g., those offering different solutions for the same problem), ‘subset’ (i.e., where the content of one document is a subset of the other) and ‘different’ (i.e., conceptually different documents).

Similarity has also been measured at the sub-document level, such as between sentences (Agirre et al., 2012; Negri, Marchetti, Mehdad, Bentivogli, & Giampiccolo, 2012)

Table 2.3 Semantic textual similarity levels (Agirre et al., 2012)

Class	Definition	Example
5	The two sentences are completely equivalent, as they mean the same thing.	1) The bird is bathing in the sink. 2) Birdie is washing itself in the water basin.
4	The two sentences are mostly equivalent, but some unimportant details differ.	1) In May 2010, the troops attempted to invade Kabul. 2) The US army invaded Kabul on May 7th last year, 2010.
3	The two sentences are roughly equivalent, but some important information differs/missing.	1) John said he is considered a witness but not a suspect. 2) “He is not a suspect anymore.” John said.
2	The two sentences are not equivalent, but share some details.	1) They flew out of the nest in groups. 2) They flew into the nest together.
1	The two sentences are not equivalent, but are on the same topic.	1) The woman is playing the violin. 2) The young lady enjoys listening to the guitar.
0	The two sentences are on different topics.	1) John went horseback riding at dawn with a whole group of friends. 2) Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

and between words (Camacho-Collados et al., 2017). Similarity between sentences has been measured to represent the degree of semantic equivalence between the two sentences; this is represented using the term ‘*semantic textual similarity*’ (STS). Agirre et al. (2012) defined six different classes to represent the different levels of STS; each class and its example of sentence pair is shown in Table 2.3. Another work, however, focused on identifying the similarity at the sentence level from the perspective of ‘*textual entailment*’ (Negri et al., 2012). In this case, the aim is to define the directional relationship between two sentences, i.e., whether one text entails the other. In this case, four different relations were used: “forward”, “backward”, “bidirectional” and “no entailment”.

A 5-point Likert scale has been used to describe the semantic similarity at the word level. This is for the purpose of a cross-lingual and multilingual semantic word similarity tasks across 5 languages (English, Farsi, German, Italian and Spanish) (Camacho-

Table 2.4 Word similarity levels (Camacho-Collados et al., 2017)

Class	**Definition**
Very similar	The two words are synonyms.
Similar	The two words share many of the important ideas of their meaning but include slightly different details (e.g., "lion-zebra" or "firefighter-policeman").
Slightly similar	The two words do not have a very similar meaning but shar ea common domain (e.g., "house-window" or "airplane-pilot").
Dissimilar	They describe clearly dissimilar concepts but may share some small details, a far relationship or a domain in common. These words are also likely to be found together in a longer document in the same topic (e.g., "software-keyboard" or "driver-suspension").
Totally dissimilar and unrelated	The words do not mean the same thing and are not on the same topic (e.g., "Playstation-monarchy").

Collados et al., 2017). The different similarity levels proposed in this work are shown in Table 2.4.

The literature described in this section has illustrated that there is no consensus on the definition of similarity. Instead, various studies have defined different classes or categories to represent the varying degrees of similarity based on their research aims (Barker & Gaizauskas, 2012; Braschler & Schäuble, 1998; Fung & Cheung, 2004; Skadiņa et al., 2012; Tomás et al., 2008). Although a large number of work have proposed different similarity schemes for news domains and Web articles, no available schemes have been specifically developed for Wikipedia articles.

2.3 Identifying monolingual similarity

A number of approaches have been investigated in previous studies to measure similarity of texts written in the same language (also referred to as *monolingual similarity*). In general, similarity can be measured lexically or semantically. Lexical similarity computes similarity based on the number of overlapping words or terms that the two texts share. Semantic similarity, on the other hand, measures the similarity between the meaning of the two texts; this often requires knowledge-based data, such as a thesaurus or a corpus. Previous studies that aimed to measure lexical and semantic similarity in monolingual texts are reviewed in Section 2.3.1 and 2.3.2, respectively.

2.3.1 Lexical similarity

Lexical similarity represents the similarity between two documents based on the words or terms they share. It does not take into account the meaning of the words. The most well-known method to measure lexical similarity is by representing each document as its index terms (e.g., using a bag-of-words method) and measuring similarity between these representations (e.g., using Jaccard similarity or cosine similarity) (Manning et al., 2008).

Pre-processing method

To determine a set of index terms for each document, a set of pre-processing tasks is often applied to remove irrelevant features from the text. In this section, the author describes a set of pre-processing methods that are frequently used.

A *case-folding* is carried out to remove capitalisation from the texts. E.g., the following words 'Country", "country" and "COUNTRY" are all represented as "country" after case-folding. In similar languages, similar words may be represented slightly differently with the use of diacritics. Therefore, a *diacritics removal* has also been used to normalise the characters. Punctuation marks are also often discarded in this stage.

A *tokenisation* process is then required to split the document into tokens by identifying the word and sentence boundaries (Grefenstette & Tapanainen, 1994). The simplest method is to split the document into words using whitespace characters (e.g., " "). Similarly, sentence boundaries can be identified using a set of rules (e.g. "." followed by a capital word can be used to represent the end of a sentence in English). However, different tokenisation process may be required for processing different languages.

In many languages, words may be modified to represent different grammatical variations, such as "revive", "revives", and "reviving". Two different processes can be used to identify that these word variations are similar by reducing the inflected words. A *stem-*

ming process uses rules such as suffix-stripping to reduce the different word variations into its stem (root word). For example, “revive”, “revives” and “reviving” are all stemmed into “reviv”. A *lemmatisation* process, on the other hand, analyses the context of the word and performs a dictionary look-up to determine the lemma, i.e. the morphological root form of the word. In the examples above, the words are lemmatised into “revive”. Note that a stem word is not necessarily the same as the lemma. Both a stemmer and a lemmatiser requires knowledge for a particular language, however the latter is more difficult to develop and is often unavailable for under-resourced languages.

Another process that is often used in the pre-processing task is a *stop-words removal.* Stop-words are high-frequency words that appear in the documents, such as prepositions (e.g., “in”, “on”, “at”) or articles (e.g., “the”). These words have little use in representing the document content and, therefore, are often removed at the pre-processing stage (Jurafsky & Martin, 2000),

Index terms

The most often used representation is a *bag-of-words* model, which assumes that a document is a bag containing all the words in it and discards any syntactical information such as the order of the words (Jurafsky & Martin, 2000). The simplicity of this representation loses the semantical meaning of the words, however, has been shown to offer good results in information retrieval.

Instead of using words as index terms, character-n-grams have also been used. Character-n-grams are defined as sequences of character of the length n. This representation is especially useful when neither stemming and lemmatisation tools are available, as char-n-grams can be used to solve the word inflection problems (McNamee & Mayfield, 2004). Consider the example in the previous section: “revive” is represented as the following 4-grams: “revi”, “eviv” and “vive”, whilst “reviving” is represented as: “revi”, “eviv”, “vivi”, “ivin” and “ving”. These representations enable both words to be identified as similar as they contain two overlapping index terms (i.e. “revi” and “eviv”). In tasks, such as identifying text reuse, n-grams have been applied at the level of word (i.e., n sequences of words) and part-of-speech (i.e., n sequences of part-of-speech) (Barrón-Cedeño, Rosso, Agirre, & Labaka, 2010; Clough et al., 2002).

Document representations

After determining a set of index terms for the documents, different models can be used to represent the document. A representation of text frequently used in information retrieval, document classification and clustering, is the Vector Space Model (VSM) (Manning et al., 2008). VSM represents each document as a vector of index terms found in the corpus (the entire document collection). The vector v for a document d is shown as:

$$\vec{v}(d) = (w_1, w_2, w_3, ..., w_n) \tag{2.1}$$

where n represents the number of unique index terms in the corpus and w_i represents the weight score of i-th term.

The weight for each index term can be represented in different ways. The simplest weighting model is a *Boolean (binary) weighting*. In this model, the weight of a term t in a document d is defined to be 1 if the term t is found in the document d, and 0 otherwise. This model, however, does not take into account that more frequent words may be more relevant as document representations compared to less frequent words.

The next model, term frequency (tf) utilises the number of times (frequency) a particular term occurs within the document as the term weight (Manning et al., 2008). The tf is also often normalised by the frequency of all terms in the document:

$$tf_t = \frac{f_{t,df}}{\sum_{i=1}^{n} f_{t_i,df}} \tag{2.2}$$

where $f_{t,df}$ defines the frequency of term t in document df, and n represents the number of unique terms in document df.

The effectiveness of tf, however, can be reduced by non-distinctive or unimportant words that appear many times in the documents in the corpus. Inverse document frequency (idf), therefore, has been used as a solution to this problem, as it is able to differentiate between common and rare terms. The idf score of a term t is calculated as:

$$idf_t = \log \frac{N}{df_t} \tag{2.3}$$

where N defines the total number of documents in the corpus and df_t represents number of documents in the corpus which contain the term t. Common words which appear in many documents will have low idf scores, whilst those which appear in fewer documents will have high idf scores. A combination of term frequency and inverse document frequency (tf-idf) is often used to represent an index term weight; the weight w of a term t is shown as:

$$w_t = tf\text{-}idf_t = tf_t \times idf_t \tag{2.4}$$

Therefore, the vector representation of document d which uses tf-idf weighting is shown as:

$$\vec{v}(d) = (tf\text{-}idf_1, tf\text{-}idf_2, tf\text{-}idf_3, ..., tf\text{-}idf_n) \tag{2.5}$$

where n represents the number of unique index terms in the corpus and tf-idf_i, represents the tf-idf score of the i-th term.

Similarity measures

Similarity between two documents can be measured using a number of approaches. When Boolean or binary representation is used, *Jaccard Coefficient* (Jaccard, 1912) can be applied to measure similarity by computing the number of overlapping words between two documents d_1 and d_2 normalised by the number of all words found in both documents:

$$sim(d_1, d_2) = J(d_1, d_2) = \frac{|\vec{v}(d_1) \bigcap \vec{v}(d_2)|}{|\vec{v}(d_1) \bigcup \vec{v}(d_2)} \tag{2.6}$$

In the Vector Space Model (VSM), a similarity score $sim(d_1, d_2)$ between two documents d_1 and d_2 is computed by carrying out a comparison between the vectors representing both documents, notated as $\vec{v}(d_1)$ and $\vec{v}(d_2)$, respectively. This approach, referred to as *cosine similarity*, measures similarity by calculating cosine value of the angle of the two vectors:

$$sim(d_1, d_2) = \frac{\vec{v}(d_1).\vec{v}(d_2)}{|\vec{v}(d_1)||\vec{v}(d_2)|} \tag{2.7}$$

The use of lexical similarity approaches has been combined with other features in identifying similar content. Barzilay and Elhadad (2003) identified similar sentences in a monolingual corpus using cosine similarity to measure the lexical similarity of the sentences. However, they also combined this with a proximity feature that takes other similar sentences into account, i.e., a proximity to sentence pairs with high similarity is considered in the alignment.

The list of approaches reported in this chapter is by no means complete as the author only selected the most frequently used methods in measuring lexical similarity between texts that are relevant for the aim of this study. Other approaches which have also been developed are available in other literature, such as Manning et al. (2008).

2.3.2 Semantic similarity

The approaches described in the previous section measure similarity by computing literal overlap between index terms (e.g., words) of the two documents. However, they are not able to recognise cases where a variety of words are used to describe the same entity or concept, i.e., synonyms. Additionally, morphologically-related words, such as ‘physician’ and ‘doctor’, are also not recognised to be similar using the literal overlap approaches. In this case, the author reviewed a list of approaches that have been used to measure semantic similarity (i.e., similarity between meaning) between texts.

Semantic similarity approaches were developed in order to “overcome the vocabulary mismatch problem between searchers and document creators” (S. T. Dumais, 2007, p. 303). A large number of studies have investigated different methods to measure the semantic similarity between two texts, such as similarity between words (J. J. Jiang & Conrath, 1997; Resnik, 1995; Taieb et al., 2014), concepts (Pedersen et al., 2004), named entities (Do et al., 2009), sentences (Mihalcea et al., 2006) and documents (S. T. Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988; Gabrilovich & Markovitch, 2007; Wan, 2007). Semantic similarity approaches have been investigated for the purpose of finding similar news articles (Pouliquen et al., 2004), identifying paraphrases (Fernando & Stevenson, 2008) and classifying documents (Wu et al., 2017). Compared to lexical similarity approaches, these approaches often require some knowledge-based resources to identify the relations between semantically similar words.

These approaches, further referred to as *knowledge-based measures*, often require the use of an ontology or a thesaurus, such as WordNet (Miller, 1995). Similar words (or concepts) are identified by counting the number of edges between the two concepts in the thesaurus; concepts with fewer number of edges are considered to be more similar than those with more edges (Resnik, 1995; Taieb et al., 2014). These approaches, however, have various limitations due to the limited availability and the non-dynamic state of these ontologies (Y. Jiang et al., 2015).

To overcome the limitations of knowledge-based measures, other studies have explored the use of *corpus-based measures*, by utilising a large corpus and incorporated corpus statistical measures for measuring word similarity (Agirre et al., 2009; S. T. Dumais et al., 1988; Gabrilovich & Markovitch, 2007; J. J. Jiang & Conrath, 1997; Koberstein & Ng, 2006; Mihalcea et al., 2006). Corpus-based measures are based on the assumption that terms or words that appear in similar contexts can be assumed to be similar to one another (S. T. Dumais et al., 1988). Term co-occurrence in the corpus, therefore, can be used to establish the relations between terms (Kandola et al., 2003). Koberstein and Ng (2006) further improved this method by utilising not only the proportion of documents that contain both words, but also the frequency of occurrences, and the distance between the words in the documents. Previous studies have utilised the use of Wikipedia (Koberstein & Ng, 2006) and a large 1.6 Terabyte Web corpus in measuring similarity of words (Agirre et al., 2009).

A well-known corpus-based measure is the *Latent Semantic Analysis (LSA)* (S. T. Dumais et al., 1988) (also referred to as Latent Semantic Indexing), which utilised a large corpus to identify similar words by analysing their occurrences in the corpus. First, this technique indexed a large collection of text and creates a term-document matrix; this matrix contains, for each term, its $tf\text{-}idf$ score of each document. A dimension reduction method was then performed using a reduced singular value decomposition. To calculate the similarity between two documents, cosine similarity between the two representation vectors were calculated in this reduced dimensional space. Using this approach, similar documents can be identified even though they do not contain exact-matching terms.

Another approach, *Latent Dirichlet Allocation (LDA)* is a probabilistic model that created an underlying set of topics based on a text corpus (Blei, Ng, & Jordan, 2003). It created a representation for each document by measuring the probabilities over the underlying set of topics, allowing it to be applied for various tasks such as document modelling and text classification. Different to LSA, each topic in LDA is characterised with a set of words, making it easier for human to interpret the different topics.

Different to LSA and LDA, other works have also investigated the use of explicit concepts in representing a document vector. In this case, the vector representations of texts are represented as a pre-determined set of natural concepts. For example, similarity between academic papers has been measured using the co-occuring terms in the titles, journal title, abstract, full-text content and assigned index terms (Trivison, 1987). Information such as the hierarchical structure of the concepts in the ACM's classification hierarchy has also been utilised in measuring similarity between academic papers (Lakkaraju et al., 2008).

Concepts from knowledge bases such as Wikipedia, Wikidata or DBPedia have also been extracted to assist in measuring similarity between documents (Benedetti, Beneventano, Bergamaschi, & Simonini, 2018; Gabrilovich & Markovitch, 2007; Y. Jiang et al., 2015; Wu et al., 2017). Studies that used Wikipedia to identify these concepts have further utilised information such as the Wikipedia summary, links and categories for the corresponding concepts to measure the similarity between the concepts (Gabrilovich &

Markovitch, 2007; Y. Jiang et al., 2015; Milne, 2007; Strube & Ponzetto, 2006). By representing a document as a set of its relevant concepts, much smaller vocabulary size is used yet the performance was found to be comparable to a word-based approach in a Vector Space Model (Milios et al., 2003). In a different study, however, Cooper et al. (2002) found that a term-based method even performed better in identifying similar Web documents compared to word-n-gram method.

The *Explicit Semantic Analysis (ESA)* approach (Gabrilovich & Markovitch, 2007) utilised Wikipedia to create term representations. A semantic interpreter containing a weighted inverted index was created; each word containing information of weighted score for each concept. Using this interpreter, one can use Machine Learning (ML) techniques to map fragments of text into a weighted sequence of concepts ordered by their relevance. To compute semantic relatedness between two documents, interpretation concept vectors between both documents are compared using cosine similarity metrics. The results show that ESA achieved much better correlation to human judgments in measuring similarity at the document level compared to the bag-of-words model and LSA. ESA is also able to compute relatedness between words (instead of documents) with much better results compared to other approaches using WordNet, Roget's Thesaurus, LSA (S. T. Dumais et al., 1988) and WikiRelate! (Strube & Ponzetto, 2006).

In recent years, another state-of-the-art feature for measuring similarity involves the use of word embedding approach, such as *word2vec* (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013b). This approach utilises deep learning methods in a large corpus to learn the representation of words using two different architectures. The first model, continuous bag-of-words (CBOW) architecture, uses a window of surrounding context words to predict the current word. The second model, continuous skip-gram architecture, models the current word by predicting the surrounding window of the context words. This approach performs well in measuring similarity or relatedness between words.

The applicability of word embedding method for measuring similarity between documents, however, has not been thoroughly investigated. A simple aggregation of the word embedding to represent all the words in the documents has been shown to achieve poor results in identifying similarity due to losing the information about the word order (similar to the bag-of-words issue) (Le & Mikolov, 2014). To solve this problem, an extension of this method was proposed to create a paragraph vector that can be used to represent a variable-length piece of text, such as phrases, sentences, or documents. This approach, *doc2vec* (Le & Mikolov, 2014), allows similarity between longer piece of of texts to be measured. Pre-trained monolingual word embedding and document embedding models are widely available. However, the availability of cross-lingual document embedding models are very scarce.

In measuring relatedness in Wikipedia articles, different approaches have been developed which utilise information only available in Wikipedia. Strube and Ponzetto (2006) mined the category information in Wikipedia and measured overlap of text to identify similar articles. In contrast, Milne (2007) used the Wikipedia link information to compute monolingual semantic relatedness between a pair of Wikipedia articles; this approach is referred to as Wikipedia Link Vector Model. Using this model, the link structures in Wikipedia were mined and used to index each Wikipedia article. Similarity was then measured by calculating the links overlap between articles. Links were weighted in a similar way to idf, i.e., links referred to by many documents (e.g. 'science') had less weighting than less common links (e.g. 'thermodynamics'). Upon dealing with ambiguous words, Milne (2007) used Wikipedia disambiguation pages which listed all possible concepts of a word and chose the most common concept. For example, the disambiguation page of the word "plane" lists the following concepts: "airplane" (which Wikipedia identified as the most common), "Planes (film)", "Plane (river)" which is a river in eastern Germany, and others. Therefore, "airplane" was chosen as the disambiguated concept. Since this approach relied on links overlap, article pairs with different length were punished as the larger article often contained many links not found in the other document.

Other studies investigated a different method to measure similarity at the document level. Instead of representing documents as relevant concepts, these approaches utilised semantic similarity approaches to identify similar words (Fernando & Stevenson, 2008; Mihalcea et al., 2006), named entities (Do et al., 2009), or sentences (Koberstein & Ng,

2006) in the documents, prior to aggregating this alignment information to represent the similarity at the document level. Mihalcea et al. (2006) utilised both knowledge-based measures and corpus-based measures to find the most semantically similar words in two sentences (with the same part-of-speech) and aggregating this information as the sentence similarity score. Sentences that score above the specified threshold are considered to be equivalent (i.e., paraphrases). They utilised both knowledge-based measures and corpus-based measures. Similar approach was investigated by Fernando and Stevenson (2008) who utilised word-to-word similarity measures to detect paraphrased sentences. Islam and Inkpen (2008) proposed a more advanced string matching method by analysing the longest common subsequence of similar words in measuring text similarity. When applied at the document level, these measures often require semantic distances between every possible pair of words to be calculated (Wan & Peng, 2005), prior to measuring document similarity using a many-to-many matching between words. This process is, therefore, not very efficient especially for a large number of documents.

Another study (Wan, 2007) proposed a different approach to measure similarity between a document pair, by utilising the structure similarity information. Given a document, it divides the document into different 'tiles', each tile represents a multi-paragraph content that describes a particular sub-topic. These sub-topics between the two documents are aligned, allowing a many-to-one alignment. Document similarity is then evaluated based on the similarity of the sub-topics.

This section has reviewed some of the most prominent work in semantic similarity approaches. Compared to the lexical overlap approaches, the semantic similarity approaches require a language-specific resource (such as an ontology or a large corpus) and are generally more expensive to compute. A combination of lexical similarity and semantic similarity has also been combined for measuring similarity between documents (L. Huang, Milne, Frank, & Witten, 2012).

2.4 Methods to measure cross-lingual similarity

The approaches described in Section 2.3 have been used to measure similarity between monolingual documents. When similarity is measured between documents in different languages (i.e., document d_1 and d_2 which are written in language l_1 and l_2, respectively, where $l_1 \neq l_2$), different approaches are required to identify the overlapping information between the two documents. These approaches are referred to as *cross-lingual similarity* approaches.

Measuring cross-lingual similarity is an essential task for various areas, such as Cross-Language Information Retrieval (CLIR) (Peters et al., 2012), the aim of which is to retrieve documents for a given topic in different languages. Identifying cross-lingual similarity is also used for the purpose of automatic classification of multilingual news articles, in which multilingual articles of the same topic are clustered together in order to aid bilingual/multilingual users (Gupta et al., 2012; Steinberger et al., 2002). In the past decade, cross-lingual similarity approaches have also been utilised to identify similar multilingual documents on the Web for the purpose of creating linguistic resources, such as building bilingual lexicons and dictionaries (Erdmann et al., 2008; Sadat, 2010) or comparable corpora (Otero & López, 2010; Tomás et al., 2008).

In some cross-lingual similarity approaches, translation resources, such as dictionaries or Machine Translation (MT) systems, are utilised to perform translation of document d_1 from language l_1 to l_2 prior to measuring monolingual similarity to d_2 written in l_2 (Munteanu & Marcu, 2005; Uszkoreit et al., 2010). These approaches are referred to as *language-dependent approaches* (described in Section 2.4.1) as they require available translation resources for each particular language.

Whilst these linguistic resources are abundant for some highly-resourced languages such as German or French, the availability of these resources is limited for languages that are under-resourced, such as Latvian or Greek. Moreover, the quality of these resources is poorer compared to those for highly-resourced languages (Skadiņa et al., 2012). As a result, these approaches are not applicable for use in under-resourced languages. To overcome this issue, some approaches have been developed to measure similarity without the need of language-specific translation resources or by utilising other resources that are widely available. These approaches are referred to as *language-independent approaches* (Section 2.4.2).

Some language-independent approaches may use resources, such as Wikipedia. One may argue that these approaches should be categorised as language-dependent approaches as Wikipedia is a language-dependent resource that is not available in all languages. On the other hand, Wikipedia is a widely available resource and therefore, the applicability of approaches that utilise Wikipedia extends to considerably more languages compared to language-dependent approaches that require SMT systems, parallel corpora or dictionaries for the particular language. In this chapter, approaches that utilise *widely available resources* in the Web are categorised as language-independent approaches.

2.4.1 Language-dependent methods

Language-dependent methods rely on the availability of language-specific resources to measure cross-lingual similarity. These methods require translation resources to translate bilingual contents into the same language prior to measuring monolingual similarity. In some studies, additional language-specific resources such as part-of-speech taggers, named entity recognisers, stemmers are also required to pre-process the content (Negri et al., 2012).

Most approaches in this section measure cross-lingual similarity by using Statistical Machine Translation (SMT) systems to translate multilingual contents into the same language, prior to measuring similarity using monolingual approaches (outlined in Section 2.3), such as using Jaccard coefficient of bag-of-words or cosine similarity (Agirre et al., 2009; Gottschalk & Demidova, 2017; Mehdad et al., 2010; Muhr, Kern, Zechner, & Granitzer, 2010; Yasuda & Sumita, 2008). The SMT systems used in these studies were built using parallel corpora, such as the Europarl Corpus (Koehn, 2005) which contains European parliament proceedings in 21 European languages, JRC-Acquis Corpora containing European Union legal documents in 21 languages (Steinberger et al., 2006), and

DGT-TM, a corpus of aligned sentences of European legislation documents in 24 EU languages (Steinberger, Eisele, Klocek, Pilos, & Schlüter, 2012). Parallel corpora, however, are often only available in well-resourced languages and limited domains (e.g., legal). In cases where SMT or parallel corpora were not used, other multilingual resources, such as EuroVoc thesaurus,[3] have also been used to assist with the translation process (Pouliquen et al., 2004). Others opted on utilising commercial MT systems such as Google Translate, Microsoft Bing Translator (Aker, Kanoulas, & Gaizauskas, 2012; Negri et al., 2012; Wäschle & Fendrich, 2012; Yasuda & Sumita, 2008); these systems, however, are not free to use and may limit the size of content to be translated.

For languages with different alphabets, if translation resources are not available, a step of *transliteration* can be performed instead, in which characters are mapped from language l_1 to l_2. For example, performing transliteration of the Greek word "ιστορια" (meaning "history") to Latin alphabets will result in the word "istoria". Using linguistic features, such as character-n-grams, it is then possible to align the original Greek word "ιστορια" to the English word "history" without using any translation resources, as "istoria" and "history" share many overlapping characters ("istor"). This approach has been used to identify similarity between news articles (Argaw & Asker, 2005).

In this section, the author further reviewed language-dependent approaches that have been used to calculate similarity at the sub-document level (e.g., sentences or phrases) and the document level.

Similarity at the sub-document level

Cross-lingual similarity approaches have been used to measure cross-lingual similarity between words, by training an SMT system to translate the multilingual content into English before utilising WordNet to measure the semantic similarity between the translated words (Agirre et al., 2009).

A different approach was explored by Bollegala, Kontonatsios, and Ananiadou (2015) to detect translated terms in the biomedical domains. Instead of using an SMT system,

[3] http://eurovoc.europa.eu/

they instead extracted a number of features, such as char-n-gram overlap, and unigrams and bi-grams of the context surrounding the terms. They reduced the dimensionality using “prototype vector projection (PVP)” then, using a training data of 5,000 pairs of translated biomedical terms in French, Spanish, Greek and Japanese (all paired to English), learnt a mapping between the terms in the source and target language using “partial least squares regression”. These features were used in a binary classification task to identify translated biomedical terms.

The use of a classifier has also been explored in identifying translated sentences. Instead of using SMT, Munteanu and Marcu (2005) have incorporated the use of probabilistic dictionaries to translate the sentences prior to measuring their similarity. A number of features, such as the proportion of overlapping words, length of contiguous connected words, and length of contiguous unconnected words were then measured (Munteanu & Marcu, 2005). Language-independent features, such as sentence length differences and sentence length ratios, were also computed in this study. These features were combined into a Maximum Entropy classifier trained on a parallel corpus of 5,000 sentences. Using a dictionary learnt from 1M out-of-domain parallel corpora, the classifier managed to identify parallel sentences from a different parallel corpus with 97% precision and 22% recall; the recall increased to 46% with no loss in precision when using a bigger dictionary (trained from 50M out-of-domain parallel corpora).

As briefly described at the beginning of this section, parallel corpora are only available in well-resourced languages and very few domains. Due to this reason, Yasuda and Sumita (2008) reported an alternative approach to exploit available commercial MT system as an initial approach and employ a bootstrapping method to build a sentence-aligned corpus from the Japanese and English Wikipedias. First, they used a commercial MT system to translate the Japanese Wikipedia corpus to English, and the English Wikipedia corpus to Japanese. Similarity was calculated in both directions by first, comparing English and translated Japanese sentences, followed by Japanese and translated English sentences. Sentences were identified to be translations of each other if they contained more than 60% word overlap. This approach achieved over 80% precision although had very low recall (10%). Using this approach, they managed to create a Japanese-English sentence-aligned corpus which was then used to train their own SMT system. Unlike the previous approaches, this approach does not require the availability of parallel corpora. However, it required a high-quality translation system for the language pair, which is also often unavailable for under-resourced languages.

In identifying textual entailment between sentences, MT systems have been used to translate the sentences prior to measuring similarity between the content (Mehdad et al., 2010; Vilarino et al., 2012; Wäschle & Fendrich, 2012). Jaccard similarity of the overlapping words and a number of rules were used to identify the entailment (Vilarino et al., 2012). Wäschle and Fendrich (2012) also utilised additional features, such as token ratio, bag-of-words (Jaccard coefficient and overlap coefficient on unigram, bigrams and trigrams) and a simple distance measure based on string edit distance. Finally, they also utilised an external sentence similarity tool[4] which carried out a word alignment between the sentences prior to calculating the number of unaligned words, percentage of aligned words, and length of the longest unaligned subsequence. All these features were then combined into an SVMlight classifier. Other approaches in this area also used other bilingual resources, such as Europarl, bilingual dictionaries, language-specific resources such as part-of-speech taggers, named entity recognisers, noun phrase identifier, stemmers and stopwords lists (Negri et al., 2012).

Similar approach was also reported by Gottschalk and Demidova (2017) who utilised a commercial MT system (Bing Translation API) for translating texts prior to measuring similarity between similar text passages in Wikipedia articles. In this study, they found that the best results were achieved by combining the cosine similarity of the tf-idf (after translations), the similarity between named entities, and the similarity between time expressions included within the text passages.

[4] Meteor tool (Denkowski & Lavie, 2011) supports English, Spanish, French and German.

Similarity at the document level

Cross-lingual similarity has been measured at the document level for the purpose of finding parallel documents from the Web (Uszkoreit et al., 2010) and identifying similar news (Pouliquen et al., 2004). Different approaches and features were used due to the different nature of the texts.

An approach using SMT was made by Uszkoreit et al. (2010) to perform a cross-language near-duplication detection task to identify parallel documents from the Web. Six languages were used in this study: English, Arabic, Chinese, French, Russian and Spanish. First, the authors trained an SMT system using the Europarl Corpus (Koehn, 2005), DGT-TM (Steinberger et al., 2012) and the United Nations ODS Corpus (United Nations, 2006). The SMT system was then used to translate all non-English documents into English. Afterwards, an index was created to map each unique word-n-gram found in the corpus into the documents in which it occurred. Candidate parallel documents were identified by matching word-n-grams between the translated non-English and English documents. These candidates were then further filtered out using more rigorous matching with lower order n-grams. The evaluation corpus contained over 2 billion Web pages with 7,286 document pairs previously identified as parallel by the website: `http://america.gov`. Using this information as the gold-standard data, Uszkoreit et al. (2010) reported that this method managed to identify parallel documents with 93% precision and 65% recall.

A slightly different approach was investigated by Ture, Elsayed, and Lin (2011) to find similar articles in German-English Wikipedia. Firstly, they built document vectors using tf-idf (bag-of-words). The foreign language document vector was then projected into the target language document vector using translation resources from either a SMT system or parallel corpora, prior to measuring cosine similarity.

Instead of using SMT, Munteanu and Marcu (2005) used parallel data to build a probabilistic dictionary as translation resources. They used this approach to identify documents that contained similar sentences from Arabic-English news corpora. Initially, an index was created for the English corpus. Similarity was first identified at the document level by translating every word in each Arabic document into English using the probabilistic dictionary. The top five English translations were extracted for each Arabic word. These translations were then concatenated and used as a query to retrieve similar English documents. The top 20 English documents were retrieved and documents which were published outside a five day window (i.e., more than two days before and after) of the publication date of the Arabic document were filtered out. The remaining documents were regarded to be similar and were paired to the Arabic document for further use as resources for a sentence extraction task.

In identifying similar news articles, Aker et al. (2012) proposed the use of Google Translate to translate the non-English titles into English, prior to measuring cosine similarity of the titles. The time and date of publications were also used as additional metrics to identify similar articles. A combination of these three features were shown to perform comparably to measuring the similarity between the content of the articles (Aker et al., 2012). Geographical information of the news articles and named entities have also been shown to be useful in identifying similar news articles (Pouliquen et al., 2004).

Linguistic resources such as SMT systems, parallel corpora, and thesauri, are very important in measuring cross-lingual similarity, as shown in the previous approaches. However, they are not widely available for under-resourced languages, such as Amharic (the official language of Ethiopia). To overcome this limitation, Argaw and Asker (2005) explored a transliteration-based approach to pair Amharic-English comparable news articles. First, they transliterated the titles of Amharic news articles and calculated the *edit distance* between each Amharic title and each English title, i.e. the number of operations (e.g., removal, insertion or substitution of character) required to transform one title into the other. They also used the metadata information,[5] such as the date and location of the news event being reported, to identify comparable news articles. With an estimation of a 100% recall of comparable articles, this approach achieved 56% precision. Increasing the word matching thresholds improved the precision to 74% although the recall dropped

[5] Metadata information is often available for news documents and is considered very useful in identifying news articles about the same events; however, they are not always available for other Web sources, such as Wikipedia or blogs.

significantly to 45%.

The previous approaches have worked by translating or transliterating the multilingual content before measuring similarity. In other studies, a different approach was used instead by mapping the contents into a set of keywords. Steinberger et al. (2002) utilised EuroVoc, a multilingual thesaurus for EU languages, to assign relevant terms as descriptors to documents prior to measuring similarity between the documents. They used manually built training data to train the classifier. Evaluated on its ability to find the correct document pair in a parallel corpus, the classifier reached 88% precision (P@1). This technique, however, was not evaluated on measuring comparable documents.

A different method that has been proposed involves the use of multilingual topic models for measuring cross-lingual similarity. Mimno, Wallach, Naradowsky, Smith, and McCallum (2009) built a multilingual topic model using an extension of LDA (Blei et al., 2003) by utilising parallel corpora (Europarl) and comparable corpora (Wikipedia) to create 400 topics in 12 languages. For the latter, only the first 1,000 characters were used to build the topic model. The results indicated that the resulting topics can be used to identify topically similar documents, and might also be useful to create bilingual lexicon as they contained some translated words. Saad et al. (2014) investigated the performance of bilingual topic models by merging Arabic-English documents for training cross-lingual LSI (S. T. Dumais, Letsche, Littman, & Landauer, 1997). Given a pair of documents, similarity is measured by mapping each of them into vectors in the LSI space, prior to measuring cosine similarity between the vectors. No translation was used in this method. This approach was shown to achieve a comparable result to training monolingual LSI model using the Arabic articles only, and utilising Google Translate to translate the English documents into Arabic, prior to measuring similarity of the LSI vectors between the Arabic document and the translated English document. Evaluated on its accuracy in aligning similar documents, the results show that bilingual models trained on parallel corpora achieved much higher accuracy (R@1=0.97) compared to models trained on comparable corpora, i.e., Wikipedia (R@1=0.42).

In the recent years, more work have also explored the use of prediction-based methods such as word embedding (Mikolov, Le, & Sutskever, 2013a) in measuring cross-lingual similarity. A detailed survey of these cross-lingual word embedding models was reported in Ruder, Vulić, and Søgaard (2017). Although these approaches are not strictly language-dependent, as they can be trained using any resources, most of their applications in previous studies have relied on the use of bilingual resources, such as bilingual corpora or dictionaries. In earlier studies, word embeddings required the use of bilingual corpora (Mikolov et al., 2013a). However, due to the limited availability of bilingual corpora, recent studies have trained monolingual word embeddings for each language separately using a monolingual corpus (which are more widely available), although a parallel corpus or a bilingual dictionary is still required to project both monolingual embeddings to the same space (Mikolov et al., 2013b; Mogadala & Rettinger, 2016). Recently, Artetxe, Labaka, and Agirre (2017) developed an approach to project the monolingual embeddings into the same space by requiring only a very small dictionary (25 terms) and initial alignment containing only numbers. Large monolingual corpora, however, are required in this study.

A different approach was proposed by Vulić and Moens (2015) who built bilingual word representations using Wikipedia as a comparable corpus. Instead of building word representation for each language then projecting them into the same space, they built a 'joint' document by merging and shuffling the words from the source and target documents. Document embedding was created by concatenating the vectors for each word (either using basic addition, or a method similar to tf-idf). The representations were trained using Wikipedia comparable corpora and Europarl parallel corpora. However, it did not investigate the performance when trained using comparable corpus only. This approach was tested in a Cross-Lingual Information Retrieval Task (CLIR), where the document representations were used to measure the similarity between a given query and document.

Finally, bilingual word embedding approaches have also been utilised to measure similarity between bilingual words (Søgaard et al., 2015) and to extract bilingual word lexicon (Artetxe et al., 2017; Vulić & Moens, 2016). However, its use to measure cross-lingual similarity at the document level has not been investigated.

2.4.2 Language-independent approaches

The approaches described in Section 2.4.1 require high-quality lexical resources for particular language(s), such as a dictionary, an SMT system, or parallel corpora. However, these language-dependent resources are not widely available for under-resourced languages and significantly limit the language coverage of these approaches. In this section, previous studies which have measured cross-lingual similarity without requiring language-dependent linguistic resources are discussed. Techniques that utilised widely available resources[6] (such as Wikipedia) in measuring cross-lingual similarity are also described in this section.

Similarity at the sub-document level

A number of language-independent features have been used to identify similarity between sentences. Gale and Church (1993) identified translated sentences within a parallel corpus using a statistical approach based on the *character length difference of sentences*. They identified that long sentences would translate to long sentences, and similarly, short sentences would translate to short sentences. The difference between character lengths was used to assign a probability score for a given sentence pair, to identify the likelihood that both sentences were of a translation relation, prior to using a dynamic programming to identify the maximum likelihood alignment. This approach was tested in aligning English, French and German parallel corpora and achieved a high accuracy (4.1% error rate). Furthermore, when only 80% of the highest scoring sentence pairs were used, the error rate decreased further to 0.7%. Similarly, the *length of words* has been used as a feature to identify parallel sentences (Munteanu & Marcu, 2005; Patry & Langlais, 2011) and parallel documents (Resnik & Smith, 2003).

Character-n-gram overlap has also been utilised as a language-independent feature

[6] Since these resources (e.g., Wikipedia) are widely available in a large number of languages, they do not provide the same limitations on the language coverage as other approaches described in Section 2.4.1.

to measure similarity without requiring any translation resources (McNamee & Mayfield, 2004). This feature has been used in the past and was shown to be valuable in identifying cross-lingual words, especially in similar languages.

In some cases, some translation resources are required to identify *overlapping information between the multilingual content.* To avoid the need of resources such as those described in Section 2.4.1, some studies have investigated the use of Wikipedia as a linguistic source to aid with translation instead. This is performed by utilising the Wikipedia *interlanguage-links information* between languages (Adafre & de Rijke, 2006; Potthast et al., 2008; Tomás et al., 2008).

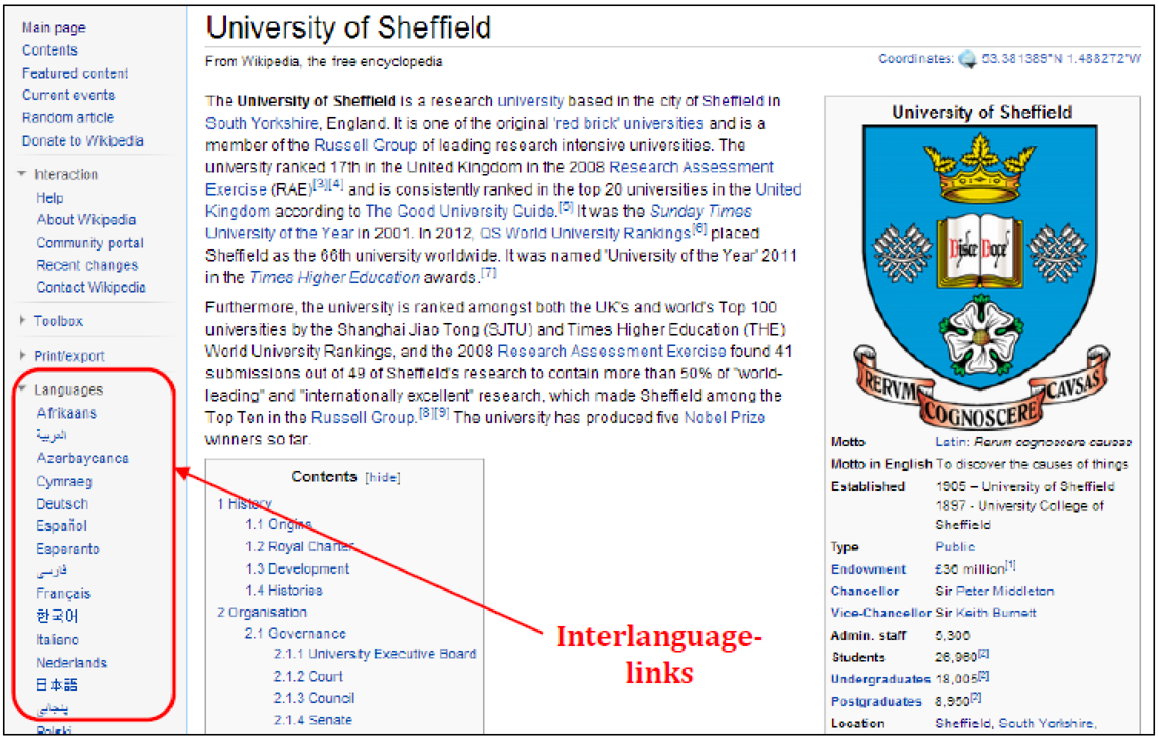
Tomás et al. (2008) utilised Wikipedia interlanguage-linked articles to collect parallel documents for building an SMT system to assist in measuring lexical similarity between the multilingual sentences. Their approach is very similar to one used by Munteanu and Marcu (2005) (previously described in Section 2.4.1). The difference is that the former approach does not require the use of parallel corpora for building SMT. Instead, it utilised Wikipedia (in Catalan-English) to extract parallel documents using features such as file sizes, numbers of HTML tags and number of paragraphs. The same features were shown to be valuable in identifying parallel document in previous study (Resnik, 1999). A statistical dictionary was built using GIZA++ (Och & Ney, 2003), an SMT toolkit that provides a word-based translation system, on the corpus. The IBM model 1 (Brown, Pietra, Pietra, & Mercer, 1993) was employed as a word alignment model: this model identifies word translations by analysing occurrence of words in each sentence pair, i.e., a pair of words which repeatedly appear in bilingual sentence pairs have higher probabilities to be translations of each other. Using this information as a translation resource, features such as proportion of overlapping words, length of contiguous connected and unconnected words and sentence length were extracted; the same features were also used in Munteanu and Marcu (2005), Smith et al. (2010) and Bharadwaj and Varma (2011b). In addition, Tomás et al. (2008) also analysed the alignment information of previous sentences. I.e., if sentences that appear before the sentence pair are aligned, it is more likely that the current sentences are also a translation of each other. This approach managed to identify parallel sentences within a set of Wikipedia parallel documents with 87% precision and 73% recall.

Bharadwaj and Varma (2011b) exploited cross-lingual links in Wikipedia to identify parallel sentences in English-Hindi Wikipedia articles. First, they indexed the document content by treating each sentence as a bag-of-words, and creating separate indexes for each language. To identify whether or not a sentence pair was parallel, they performed a retrieval for each sentence from the appropriate index; English sentences were queried on the English index and Hindi sentences were queried on the Hindi index, resulting in a set of documents for each language. Different features were then extracted, such as (1) the number of English articles for which corresponding Hindi articles (according to Wikipedia interlanguage links) were retrieved (and vice versa), (2) number of English articles whose corresponding Hindi articles were not retrieved (and vice versa), (3) total number of articles retrieved in both languages, and (4) difference of sentence lengths. A binary sentence classifier was then developed and trained on an English-Hindi word-aligned parallel corpus. They reported that the binary sentence classifier achieved 78% accuracy in identifying parallel sentences. This result is significantly higher than the accuracy achieved by the baseline approach, which was performed by translating sentences using an available dictionary and computing similarity using Jaccard Coefficient. Whilst all these features can be extracted without the need of linguistic resources, this approach did require the availability of parallel corpus for training the classifier.

Wikipedia interlanguage links information has further been utilised to measure the cross-lingual semantic relatedness task between two words (Hassan & Mihalcea, 2009). In this task, they used ESA to create vectors of concepts of a particular word (using a max of 10,000 concepts), then used interlanguage links to map to create vector in the other language. Semantic similarity is calculated using the similarity (using Lesk-like metric) of the concept vectors. They tested the approach in English, Spanish, Arabic and Romanian and found that their performance correlates with the Wikipedia size for the corresponding language, i.e., languages with bigger corpora size perform better than those with smaller ones.

Titles from Wikipedia interlanguage-linked articles have also been utilised as the basis of translation resources for finding similar sentences in Wikipedia articles (Adafre & de Rijke, 2006). First, they created a Dutch-English lexicon by extracting the titles of all interlanguage-linked Wikipedia articles in Dutch and English. An example of interlanguage-links in Wikipedia is shown in Figure 2.1. This figure shows the English version of "University of Sheffield" article in Wikipedia, with the list of interlanguage links shown on the bottom left of the figure. For example, by clicking "Nederlands", one is then directed to an article about the same topic (titled "Universiteit van Sheffield") written in the chosen language (Dutch) as shown in Figure 2.2. Therefore in this case, the English title "University of Sheffield" and the Dutch title "Universiteit van Sheffield" are extracted for building the bilingual lexicon. The Wikipedia IDs are representations of the article title with the spaces replaced into "_", e.g., the Wikipedia ID for "University of Sheffield" article is "`University_Of_Sheffield`". Synonym information were also gathered from the Wikipedia redirection pages, for example, since the Wikipedia page "Sheffield University" is redirected to "University of Sheffield", the former can be considered as a synonym of the latter. An example of how this method works is illustrated in Table 2.5 and described below.

Firstly, given a pair of interlanguage-linked articles, Adafre and de Rijke (2006) split the Wikipedia articles into sentences. Given an original sentence (see Stage 1: "Original sentence" in Table 2.5), where the original links included in Wikipedia articles are shown in blue, Adafre and de Rijke (2006) identified more links by generating word-n-grams within the sentence and finding whether these phrases existed in the bilingual lexicon. If they existed, these new links were added into the sentence (Stage 2). In this example, "Universiteit van Sheffield" and "University of Sheffield" were identified as phrases that exist in the bilingual lexicon and therefore were added as new links. In the next process, all links (both original and newly added) in the sentence were extracted (Stage 3) and mapped into their unique Wikipedia IDs (Stage 4). They then replaced the Dutch Wikipedia IDs to their corresponding English IDs using the previously created bilingual lexicon (Stage 5). This version was then used as a representation of the original sen-



University of Sheffield

From Wikipedia, the free encyclopedia

The **University of Sheffield** is a research university based in the city of Sheffield in South Yorkshire, England. It is one of the original 'red brick' universities and is a member of the Russell Group of leading research intensive universities. The university ranked 17th in the United Kingdom in the 2008 Research Assessment Exercise (RAE)[3][4] and is consistently ranked in the top 20 universities in the United Kingdom according to The Good University Guide.[5] It was the *Sunday Times* University of the Year in 2001. In 2012, QS World University Rankings[6] placed Sheffield as the 66th university worldwide. It was named 'University of the Year' 2011 in the *Times Higher Education* awards.[7]

Furthermore, the university is ranked amongst both the UK's and world's Top 100 universities by the Shanghai Jiao Tong (SJTU) and Times Higher Education (THE) World University Rankings, and the 2008 Research Assessment Exercise found 41 submissions out of 49 of Sheffield's research to contain more than 50% of "world-leading" and "internationally excellent" research, which made Sheffield among the Top Ten in the Russell Group.[8][9] The university has produced five Nobel Prize winners so far.

Contents [hide]

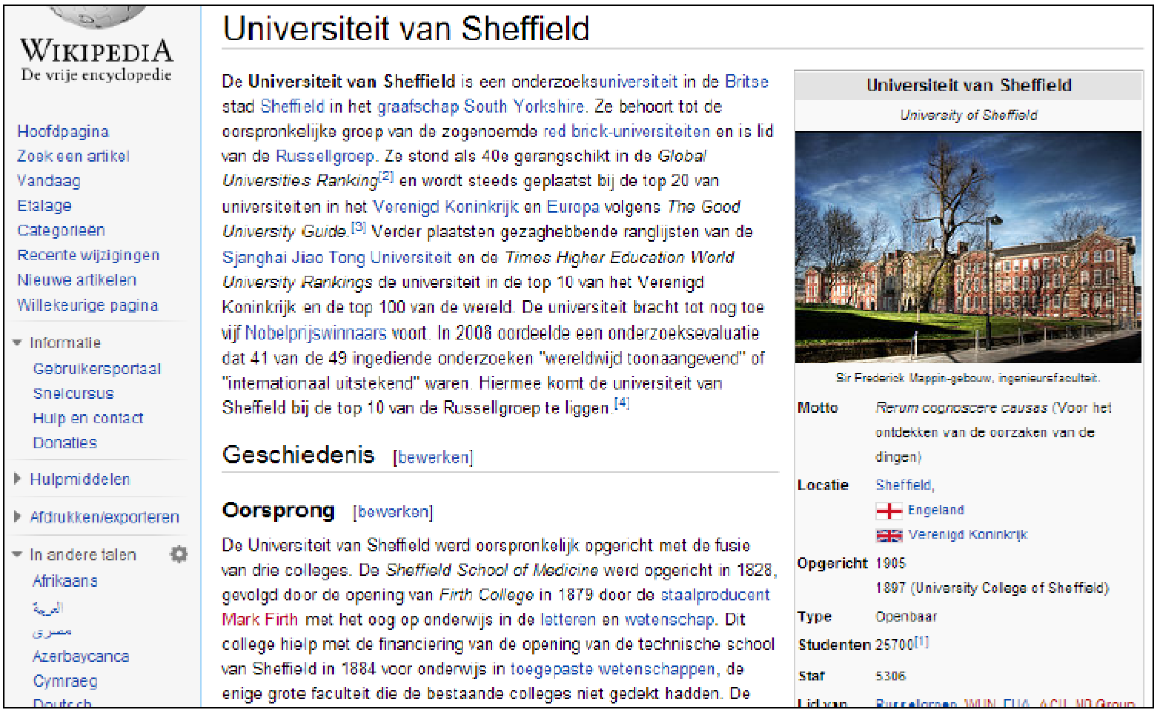
1 History
1.1 Origins
1.2 Royal Charter
1.3 Development
1.4 Histories
2 Organisation
2.1 Governance
2.1.1 University Executive Board
2.1.2 Court
2.1.3 Council
2.1.4 Senate

University of Sheffield

Motto	Latin: *Rerum cognoscere causas*
Motto in English	To discover the causes of things
Established	1905 – University of Sheffield 1897 - University College of Sheffield
Type	Public
Endowment	£30 million[1]
Chancellor	Sir Peter Middleton
Vice-Chancellor	Sir Keith Burnett
Admin. staff	5,300
Students	26,980[2]
Undergraduates	18,005[2]
Postgraduates	8,950[2]
Location	Sheffield, South Yorkshire,

Interlanguage-links

Fig. 2.1 English version of "University of Sheffield" article



Universiteit van Sheffield

De **Universiteit van Sheffield** is een onderzoeksuniversiteit in de Britse stad Sheffield in het graafschap South Yorkshire. Ze behoort tot de oorspronkelijke groep van de zogenoemde red brick-universiteiten en is lid van de Russellgroep. Ze stond als 40e gerangschikt in de *Global Universities Ranking*[2] en wordt steeds geplaatst bij de top 20 van universiteiten in het Verenigd Koninkrijk en Europa volgens *The Good University Guide*.[3] Verder plaatsten gezaghebbende ranglijsten van de Sjanghai Jiao Tong Universiteit en de *Times Higher Education World University Rankings* de universiteit in de top 10 van het Verenigd Koninkrijk en de top 100 van de wereld. De universiteit bracht tot nog toe vijf Nobelprijswinnaars voort. In 2008 oordeelde een onderzoeksevaluatie dat 41 van de 49 ingediende onderzoeken "wereldwijd toonaangevend" of "internationaal uitstekend" waren. Hiermee komt de universiteit van Sheffield bij de top 10 van de Russellgroep te liggen.[4]

Geschiedenis [bewerken]

Oorsprong [bewerken]

De Universiteit van Sheffield werd oorspronkelijk opgericht met de fusie van drie colleges. De *Sheffield School of Medicine* werd opgericht in 1828, gevolgd door de opening van *Firth College* in 1879 door de staalproducent Mark Firth met het oog op onderwijs in de letteren en wetenschap. Dit college hielp met de financiering van de opening van de technische school van Sheffield in 1884 voor onderwijs in toegepaste wetenschappen, de enige grote faculteit die de bestaande colleges niet gedekt hadden. De

Universiteit van Sheffield

University of Sheffield

Sir Frederick Mappin-gebouw, ingenieursfaculteit.

Motto	*Rerum cognoscere causas* (Voor het ontdekken van de oorzaken van de dingen)
Locatie	Sheffield, Engeland Verenigd Koninkrijk
Opgericht	1905 1897 (University College of Sheffield)
Type	Openbaar
Studenten	25700[1]
Staf	5306
Lid van	Russelgroep, WUN, EUA, ACU, N8 Group

Fig. 2.2 Dutch version of "University of Sheffield" article

Table 2.5 Link-based bilingual lexicon approach

Stage	Dutch	English
1. Original sentence *(original links are shown in blue)*	De Universiteit van Sheffield is een onderzoeksuniversiteit in de Britse stad Sheffield in het graafschap South Yorkshire.	The University of Sheffield (informally Sheffield University) is a research university based in the city of Sheffield in South Yorkshire, England.
2. Sentence after links expansion *(newly identified links are shown in red)*	De Universiteit van Sheffield is een onderzoeksuniversiteit in de Britse stad Sheffield in het graafschap South Yorkshire.	The University of Sheffield (informally Sheffield University) is a research university based in the city of Sheffield in South Yorkshire, England.
3. Extracted links	Universiteit van Sheffield universiteit Britse Sheffield graafschap South Yorkshire	University of Sheffield university Sheffield South Yorkshire England

4. Unique Wikipedia ID (before translation)	`Universiteit_van_Sheffield`
	`Universiteit`
	`Verenigd_Koninkrijk`
	`Sheffield`
	`Graafschappen_van_Engeland`
	`South_Yorkshire`
		`University_of_Sheffield`
		`University`
		`Sheffield`
		`South_Yorkshire`
		`England`

5. Unique Wikipedia ID (after translation)	`University_of_Sheffield`
	`University`
	`United_Kingdom`
	`Sheffield`
	`Counties_of_England`
	`South_Yorkshire`
		`University_of_Sheffield`
		`University`
		`Sheffield`
		`South_Yorkshire`
		`England`

tence. Dutch Wikipedia IDs that were not included in the bilingual lexicon were left un-translated. Given a document pair, all possible sentence pairs were generated, and Jaccard similarity was calculated on the sentence representations (i.e., English Wikipedia IDs). Finally, they used these scores to create a one-to-one sentence alignment between the two documents, representing sentences that were identified to be similar. They compared this approach to one using an MT system to translate all Dutch Wikipedia articles into English and calculating monolingual similarity between each sentence pair in order to determine similar sentences. They reported that the MT-based approach generated more sentence pairs but link-based bilingual lexicon approach achieved a significantly higher accuracy than MT (26% and 45%, respectively).

Other information from Wikipedia, such as categories, disambiguation pages, hyperlinks, and redirect pages have also been exploited for measuring semantic similarity between concepts (Y. Jiang et al., 2015).

Vulić and Moens (2013) proposes an approach to identify semantically similar words across languages based on their similarity of their semantic responses for the purpose of a bilingual lexicon extraction. Their semantic responses were generated using the LDA model with the number of topics set to be around 2000. They also compared whether better semantic responses could be learned from higher quality data (by comparing Wikipedia corpus only against Wikipedia+Europarl). They used a very small set of documents (between 7,612-18,898 pairs) and relied on a POS tagger to consider only nouns (that occured 5 times or more in the corpus). Their findings show that the quality of the vectors are dependent on the quality of the training data. When Wikipedia is supplemented with Europarl, the overall performance scores drastically increase.

Their later work reported the use of bilingual word embeddings without the use of any parallel corpora (Vulić & Moens, 2016). Instead it utilised a document-aligned comparable corpus (ES-EN, IT-EN and NL-EN Wikipedia) to create a pseudo-bilingual documents by intermingling cross-lingual words within the document contents whilst still keeping the order of the monolingual words. These pseudo-bilingual documents were then used for training the bilingual word embedding. This approach was evaluated on bilingual lexicon extraction only and for suggesting word translations in the content. However, its use for measuring cross-lingual similarity at the document level has not been investigated. Furthermore, its performance was also not compared to one using parallel corpora.

Similarity at the document level

The use of language-independent approaches in identifying bilingual parallel (translated) texts in the Web has been investigated in many studies. Resnik (1999) and Zhang, Wu, Gao, and Vines (2006) measured the similarity of HTML structures and URL paths of documents to identify those that are parallel. These approaches do not require language resources and work well in identifying translated articles, such as different language versions of a multi-language website. Tested on French-English articles in the Web, these approaches achieved a very high accuracy (over 90% precision).

Patry and Langlais (2011) explored a different approach by using an Information Retrieval (IR) system to narrow down the candidate document pairs by analysing the similarity of the document contents. The system they developed, PARADOCS, contained three components. First, it indexed all documents in the corpus using *hapax words*[7] and numerical entities. To search for candidate pairs, a sequence of hapax words or numerical entities was used as the query. Once candidate pairs were retrieved, a supervised classifier was used to classify whether or not the pair was parallel. This classifier used three language-independent features: normalised edit distance between each representation (i.e., numerical entities, hapax words and punctuation marks) of the two documents, total number of entities and a binary feature which determined whether or not the pair had the smallest edit distance among all document pairs. Tested on French-English Wikipedia, PARADOCS was able to identify parallel (and noisy parallel) articles with 80% precision.

Tomás et al. (2008) proposed a different method to identify parallel articles in Catalan-English Wikipedia using several features, such as differences in the HTML structures, the HTML tags, the number of paragraphs and the file sizes. Article pairs whose differences were above the previously set thresholds were filtered out. The contents of the remaining documents were further analysed using sentence similarity features, such as the sentence length difference and the number of overlapping cognates[8] (Gale & Church, 1993). Finally, Tomás et al. (2008) discarded article pairs which did not contain sufficient aligned sentences. The remaining article pairs were identified as parallel documents. Tested on 200 Catalan-English bilingual Wikipedia articles, this approach achieved 100% precision and 78% recall.

Previous studies have identified that several features, such as HTML structures, the number of paragraphs, file sizes and similarity of URLs, are extremely important in iden-

[7] Hapax words are any words containing more than 4 characters and occuring only once in the document.

[8] Cognates are words which derive from the same original word, e.g. "father" in English and "Vater" in German. A pair of words were considered as cognates if the edit distance was below a specified threshold.

tifying parallel documents (Resnik, 1999; Tomás et al., 2008; Zhang et al., 2006). These features, however, are not suitable for identifying comparable documents due to the following reasons. First, comparable documents are often developed independently and published in different websites (such as news articles which are published in different news site); therefore, these documents often do not share the same or similar URLs. Secondly, the contents of comparable documents, whilst describing the same concept, are often written by different authors who represent the contents differently. As a result, these documents may have different file sizes and HTML structures. Features for identifying parallel documents, therefore, are not suitable for identifying comparable documents.

Wikipedia interlanguage links have also been exploited to create a multilingual retrieval model for cross-language similarity analysis, referred to as *Cross-Language Explicit Semantic Analysis (CL-ESA)* (Potthast et al., 2008). Similar to ESA (Gabrilovich & Markovitch, 2007) (described in Section 2.3.1), CL-ESA also uses a set of concept documents in measuring similarity. The difference is that CL-ESA uses multilingual concept documents instead of monolingual documents. Given two languages, l_1 and l_2, firstly, a set of Wikipedia documents in l_1 was chosen as concept documents. Concept documents in l_2 were then selected by retrieving the corresponding Wikipedia interlanguage-linked documents from each concept document in l_1. These interlanguage-linked documents were considered to be equivalent to each other as they described the same topic. The similarity was then calculated monolingually between documents in l_1 to each of the concept documents in l_1 in order to create a similarity vector; this process was repeated on documents in l_2 to create the corresponding similarity vectors. As the concepts in both languages were treated as equivalent, these similarity vectors could then be compared in a language-independent manner. Tested on a corpus containing 1,000 parallel articles (from JRC-Acquis), 1,000 comparable articles (Wikipedia interlanguage-linked articles) and 1,000 unaligned articles, CL-ESA managed to find the correct pair with over 91% accuracy.

This section has identified a number of language-independent approaches that have been used for measuring cross-lingual similarity. It also highlighted the importance of Wikipedia as a resource for measuring similarity across languages.

2.5 Wikipedia as a linguistic resource

Previous studies (Adafre & de Rijke, 2006; Gabrilovich & Markovitch, 2007; Y. Jiang et al., 2015; Potthast et al., 2008; Sadat, 2010; Wu et al., 2017) have utilised Wikipedia as a resource to assist in various linguistic tasks due to a number of reasons. Firstly, Wikipedia contains articles from a large number of languages and domains. By March 2018, Wikipedia contained more than 47.6 million articles written in 298 languages and covering a wide range of domains.[9] The coverage of languages and domains have also significantly increased throughout the years (Clark, Ruthven, & Holt, 2009; Voss, 2005). Secondly, Wikipedia articles are freely available to download and there are available open-source Wikipedia extraction tools, such as JWPL (Zesch, Gurevych, & Mühlhäuser, 2007) that can be used to extract Wikipedia content easily. Thirdly, multilingual articles of the same topics in Wikipedia are aligned to each other using interlanguage links.[10] This document alignment is very valuable for creating and extracting bilingual resources from Wikipedia.

In this section, the author reviewed previous studies that have mined Wikipedia for a variety of purposes, such as extracting similar text for linguistic resources (Section 2.5.1), and assisting with computational tasks, such as classification and assessing the semantic relatedness of documents (Section 2.5.2).

2.5.1 Extraction of similar texts

A number of studies have extracted valuable bilingual resource from Wikipedia by utilising different information available in the corpus. The titles of interlanguage-linked articles, for example, have been found to correspond in a translation relation and have been

[9] Data collected in March 2018 from `http://meta.wikimedia.org/wiki/List_of_Wikipedias`.

[10] Some examples of interlanguage-linked articles have been discussed in Figure 2.1 and Figure 2.2 in Section 2.4.2.

extracted and used to build a bilingual dictionary (Adafre & de Rijke, 2006; Erdmann et al., 2009). Extracted for German-English language pair, these bilingual terms were shown to achieve high precision (92.3%) but very low recall (19.1%) compared to a dictionary as they were mostly available for proper names and domain-specific terms, but not general terms. Similar accuracy (92%) was reported when extracting Portugese-Spanish terms from Wikipedia (Gamallo & Garcia, 2012) and they managed to extract more than 27,000 new pairs of terms compared to an existing dictionary. Erdmann et al. (2009) investigated the use of a Support Vector Machine classifier to extract bilingual terms from Wikipedia using a number features, such as redirection pages and anchor texts (both treated as synonyms as the page title they refer to). This approach managed to increase the recall to 35% but reduced the precision to 77.2%.

Some studies have assumed that the contents of interlanguage linked articles are similar because they discuss the same topic (Mohammadi & GhasemAghaee, 2010; Sadat, 2010), although some have found evidence to suggest otherwise (Filatova, 2009; Patry & Langlais, 2011) (discussed further in Section 2.6). The former studies, therefore, have extracted these interlanguage-linked articles for the purpose of building comparable corpora (Mohammadi & GhasemAghaee, 2010; Sadat, 2010). Sadat (2010) further proposed a method to use Wikipedia to build comparable corpora for specialised domains by utilising the links between the monolingual articles. Given a query of *n* words that are relevant to the required domain, Wikipedia search engine was used to retrieve relevant articles. The links found in these documents were then explored recursively in order to expand the corpus. Lastly, information from interlanguage links was exploited to identify comparable articles in the target language.

Whilst previous studies assumed that interlanguage-linked article pairs were comparable to each other, other researchers found that many article pairs did not contain exactly the same contents, yet still shared large amount of similar fragments (Yu & Tsujii, 2009). Therefore, information extraction in Wikipedia has also been performed at the sub-document level, e.g., sentences, terms or 'infoboxes'. W.-P. Lin, Snover, and Ji

(2011) extracted information found in Wikipedia infoboxes[11] to gather *named entities* (e.g., names of people, places or organisations, dates or events) and other equivalent information in different languages. Other studies gathered parallel sentences in Wikipedia by extracting the captions of similar images in interlanguage-linked articles (Smith et al., 2010) or identifying bilingual sentences that shared high overlap of links (Adafre & de Rijke, 2006; Tomás et al., 2008).

A number of studies have also extracted similar fragments (e.g., phrases or words) from the content of the articles for the purpose of creating bilingual dictionaries (Bharadwaj & Varma, 2011a; Tyers & Pienaar, 2008) and extracting bilingual terms (Erdmann et al., 2008, 2009; Sadat, 2010; Yu & Tsujii, 2009). The accuracies of these approaches, however, have been shown to be quite poor as they retrieved many incorrect translations (Erdmann et al., 2009; Sadat, 2010), possibly caused by the low similarity of the content of the documents. These incorrect terms, however, were further shown to be related to each other (e.g., "evaporation" and "vaporized") or contain a hypernym/hyponym relation ('e.g., "car" and "vehicle") and therefore were still useful for enriching monolingual or bilingual language resources.

These studies show that Wikipedia is a valuable source for extracting bilingual content from the Web.

2.5.2 Assistance for other tasks

Wikipedia has also been used to assist with various tasks, such as document clustering and classification (Ni, Sun, Hu, & Chen, 2009), Cross-Language Information Retrieval (CLIR) (Schönhofen, Benczúr, Bíró, & Csalogány, 2008) and the assessment of semantic relatedness (Gabrilovich & Markovitch, 2007; Potthast et al., 2008; Søgaard et al., 2015). Different to the tasks described in Section 2.5.1, in these tasks, Wikipedia is used as a corpus without the need for extracting similar fragments within it. This section describes

[11] An infobox is a table summarising some facts and statistics which are common for the article's type. For example, articles about a person may contain information such as 'name', 'nationality' and 'profession' in the infoboxes, while those describing a country may contain facts such as 'capital city', 'official languages' and 'currency'.

some of the major works in this area.

Wikipedia links have also been utilised to extract the relations between monolingual concepts, allowing Wikipedia to be used as a knowledge base to enrich the representation of a given document (Benedetti et al., 2018; Y. Jiang et al., 2015; Wu et al., 2017). This information has been shown to be valuable in assisting with tasks, such as document classification (Wu et al., 2017). The relations between concepts have also been exploited or the purpose of disambiguating word senses (Turdakov & Velikhov, 2008) and named entities (Cucerzan, 2007). These approaches made use of the similarity between the contextual information surrounding the ambiguous words/named both in the Wikipedia articles and in the documents (e.g., news articles). Wikipedia category information of the named entities was also utilised in Cucerzan (2007)'s approach.

The links between articles were also shown to be useful in building topic models, both monolingually and cross-lingually (Ni et al., 2009; Saad et al., 2014). Ni et al. (2009) performed a Cross Lingual Text Classification (CLTC) task by using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) (previously described in Section 2.3.1) to model Wikipedia topics in English and Chinese. This process produced bilingual representations for a large number of topics where each representation contained different words (or terms) in one language. Given a document in a source language, they used the bilingual representations to classify that document to the corresponding classification in the target language, allowing this task to be performed without the need of any translation. The results showed that CLTC using multilingual LDA outperformed a classification performed by translating the source document to the target language and using a Support Vector Machine to classify documents.

Wikipedia has also been exploited for tasks such as semantic relatedness assessment between a pair of monolingual documents (Gabrilovich & Markovitch, 2007; Milne & Witten, 2008; Nakayama, Hara, & Nishio, 2007; Ponzetto & Strube, 2007; Zesch & Gurevych, 2010) and bilingual documents (Potthast et al., 2008), and to train bilingual word embedding (Vulić & Moens, 2015, 2016). The descriptions of these methods have been described in Section 2.3 and Section 2.4, respectively.

2.6 Identifying similarity (or dissimilarity) in Wikipedia

A number of studies have been carried out to gain a better understanding of Wikipedia, including its growth (Voss, 2005), how the articles evolved over time (Clark et al., 2009), the quality of the articles (Hu, Lim, Sun, Lauw, & Vuong, 2007) and how its categories compare to other encyclopaedias (Holloway, Bozicevic, & Börner, 2007). Their findings show that the categories provided by Wikipedia are comparable to those provided by other encylopaedias (Britannica and Encarta), containing popular categories such as 'Science', 'Society', 'History', 'Geography', and 'Maths' (Holloway et al., 2007). Furthermore, Wikipedia had additional top categories which were not available in the other two, such as 'People' and 'Topic Lists'.[12] Holloway et al. (2007), however, also reported that the categories were not fully hierarchical. I.e., although the categories contained a set of structure, some circularities existed between some categories and not all categories were linked to the top-level categories.

Despite its rich coverage of domains, the quality of the articles has been found to differ sigificantly (Hu et al., 2007). Some articles in Wikipedia were manually labelled by the Wikipedia editorial team to represent its quality (i.e., the completeness of information, the writing style and the layout) using 6 different labels (shown in Table 2.6).[13] Hu et al. (2007) analysed the quality labels of 242 English articles listed in the 'List of Countries' page[14] and found that the qualities of these articles varied widely; 6% of these articles were labelled as 'Featured Articles', 8% were labelled as 'A-class' articles, 4% were 'Good Articles', 64% were considered to be 'B-class' articles, 12% were 'Start' articles, and 5% were unlabelled. These findings indicate that only a small number of articles were assessed to be of a professional quality and contain complete facts of the concepts being described.

Although these studies have shown some insights of Wikipedia, they did not analyse how the quality and the content of articles differ across languages. In fact, very few

[12] This category represents Wikipedia articles which provide lists of topics, such as List of Countries, List of People by Occupation, etc.

[13] http://en.wikipedia.org/?title=Wikipedia:Version_1.0_Editorial_Team/Assessment

[14] http://en.wikipedia.org/wiki/List_of_countries

Table 2.6 Diferent labels used to represent the quality of Wikipedia articles

Quality label	Definition
Featured articles	Well-written, well-researched and comprehensive articles that do not require further content addition. This is the best article quality in Wikipedia.
A-class articles	These articles contain complete descriptions of the subjects but still require an expert knowledge to improve the contents and styles of the articles.
Good articles	Well-written articles useful to most readers but not yet reaching the quality of a professional encyclopaedia.
B-class articles	These articles are mostly complete but still require further work to reach good article standards.
Start articles	Incomplete articles which were still developing.
Stub articles	Articles with very basic description and generally represent bad-quality articles. This is the lowest article quality in Wikipedia.

studies aimed to analyse the similarity between multilingual Wikipedia articles, although many studies (as outlined in Section 2.5) have utilised Wikipedia for supporting many cross-lingual tasks. In Section 2.6.1, the author reviewed past studies that aimed to analyse Wikipedia articles, specifically those that identified the similarity (or dissimilarity) in Wikipedia interlanguage-linked articles. Research aimed at improving similarity in Wikipedia are then discussed in Section 2.6.2.

2.6.1 Evaluating similarity in interlanguage-linked articles

For tasks such as assisting semantic relatedness between multilingual documents, building comparable corpora or building bilingual dictionaries, multilingual Wikipedia articles of the same topic (i.e., interlanguage-linked articles) have often been assumed to be ‘comparable’, ‘equivalent’, or ‘similar’ because these articles discussed the same topic (Mohammadi & GhasemAghaee, 2010; Otero & López, 2010; Sadat, 2010). Another study indicated that "Wikipedia ... contains sets of documents that are not translations of one another, but are very likely to be about similar concepts." (Mimno et al., 2009, p. 880). However, few studies that evaluated the contents of some of these documents found that the similarity of these articles varied widely (Filatova, 2009; Otero & López, 2010; Patry &

Langlais, 2011; Tomás et al., 2008; Yu & Tsujii, 2009).

Tomás et al. (2008) manually analysed a set of 72 Spanish-English document pairs from the 'Pharmacology' category and found that only 4% of these document pairs were parallel. Patry and Langlais (2011) also carried out a similar analysis on 200 randomly-sampled pairs of French-English Wikipedia in 2009. Using the comparability levels proposed by Fung and Cheung (2004) (previously described in Section 2.2), they found that 14% were parallel, 11% noisy parallel, 29% topic-related and almost half of them (46%) were non-similar.

Filatova (2009) also manually investigated a set of Wikipedia articles that described 48 different people in a large number of languages. Her findings show that not only the description of the entries varied considerably across different languages, in some cases, the contents were even contradictory. Additionally, these interlanguage-linked articles also differed in length, and shorter documents were not necessarily summaries of larger ones and may contain different information. Lastly, Filatova (2009) also found a substantial difference between the number of languages in which an article was available; five people were described in one language only, while another person was described in 86 languages (in average, a person was described in 25 different languages).

The proportion of comparability at the document level, however, was shown to differ to the comparability at the sub-document level. Yu and Tsujii (2009) found that document pairs of lower comparability might still contain valuable parallel fragments (such as terms). Meanwhile, Tomás et al. (2008) reported that although only 4% of the Spanish-English document pairs were parallel, 15% of the sentences in the Spanish articles were exact translations of the English sentences.

The findings reported in these studies were based on a very small dataset and might not be representative of the overall similarity of Wikipedia. Nevertheless, they provided an evidence that the degree of similarity in Wikipedia interlanguage-linked articles varied considerably and these articles should not be assumed to be similar.

2.6.2 Improving similarity in Wikipedia

Whilst some studies have focused on analysing similarity between Wikipedia articles, others have focused their work on improving the similarity in Wikipedia by synchronising contents across languages. One of the earlier works was carried out by Adar, Skinner, and Weld (2009) who developed an automatic system, “Ziggurat”, to leverage Wikipedia infobox information in four different languages: English, French, German and Spanish. It used a self-supervised learning to align corresponding attributes in the infoboxes in the different languages. A number of features were used in the alignment, such as overlapping words, overlapping character-3-grams, and translated contents using dictionary and Wikipedia interlanguage links as translation resources. They used the trained classifier to identify missing attributes in the infobox and add the corresponding attributes and values where required.

More recently, a couple of studies (Bronner, Negri, Mehdad, Fahrni, & Monz, 2012; Cosma, 2015) have also proposed approaches to improve the similarity of contents across articles. Cosma (2015) proposed a semi-automatic method to complete Wikipedia articles by utilising a Machine Translation (MT) system and a Cross-Language Information Retrieval (CLIR) method. This method works by identifying parallel paragraphs from Wikipedia, prior to identifying the missing text fragments. The missing text paragraphs are then translated and used to synchronise the content between articles. Meanwhile, Bronner et al. (2012) used an MT system to enhance articles in other language upon addition of new content in one language version. This approach identified the content overlap between documents (in order to avoid inclusion of redundant information), identified insertion point for translated contents, analysed edits between factual content changes and any corrections of the MT output and use the knowledge to improve the MT system.

These studies have identified that there is a need to improve similarity in Wikipedia articles across languages. However, very few studies have investigated the work in increasing similarity between different Wikipedia language versions.

2.7 Summary

In this chapter, the author has reviewed the literature around the definition of similarity, approaches for measuring monolingual and cross-lingual similarity, and a number of studies analysing the use of Wikipedia and its similarity across languages. Despite the continuing interest in Wikipedia, however, there appears to be little work on comparing similarity at the article level, specifically what features or characteristics make two Wikipedia articles similar, such as similarity in structures, length of articles, etc.

Many of the previous studies were also limited to well-resourced languages in Wikipedia. Given the variance in size of Wikipedia in different languages, the performance of similarity measures is likely to vary (particularly for languages where translation resources are limited). Further work is therefore needed to understand the notion of similarity between interlanguage-linked Wikipedia articles. In addition, to the best of the author's knowledge, there has been little or no research on comparing automatically-derived similarity scores and human judgments. This work aims to address this and provide empirical evidence demonstrating the success of measuring cross-language similarity between different language pairs.

Chapter 3

Methodology

The goal of this thesis is to develop language-independent techniques to measure the similarity of Wikipedia articles across different languages. Firstly, the author summarises the methodology taken in Section 3.1. The remaining sections described the research stages carried out in this study. First, an initial study to identify relevant features that contribute to similarity in Wikipedia was performed (Section 3.2). Secondly, the author gathered human judgments on similarity to gain more understanding on similarity characteristics in Wikipedia and to build an evaluation corpus (Section 3.3). These features (together with findings from previous literature) were then incorporated into a number of approaches for measuring similarity in Wikipedia (Section 3.4). These approaches were evaluated and analysed using the methodology described in Section 3.5. Finally, the Wikipedia corpus used in this study is reported in Section 3.6.

3.1 Overview of methodology

The aim of this study is to develop language-independent approaches to measure similarity in Wikipedia. As discussed in Chapter 1, three research questions were identified in this study:

RQ1. What are the characteristics of similar interlanguage-linked articles in Wikipedia?

RQ2. Can we create an evaluation benchmark for Wikipedia? I.e., do human assessors agree on Wikipedia similarity?

RQ3. Can language-independent approaches be used to identify cross-lingual similarity in Wikipedia?

To answer these research questions, the author revisits the structure of work described in Chapter 1 (shown in Figure 3.1), as it shows the two main stages of work that were carried out in this study. The first stage aims to understand similarity in Wikipedia in order to identify similarity characteristics in Wikipedia (to answer Research Question 1) and to

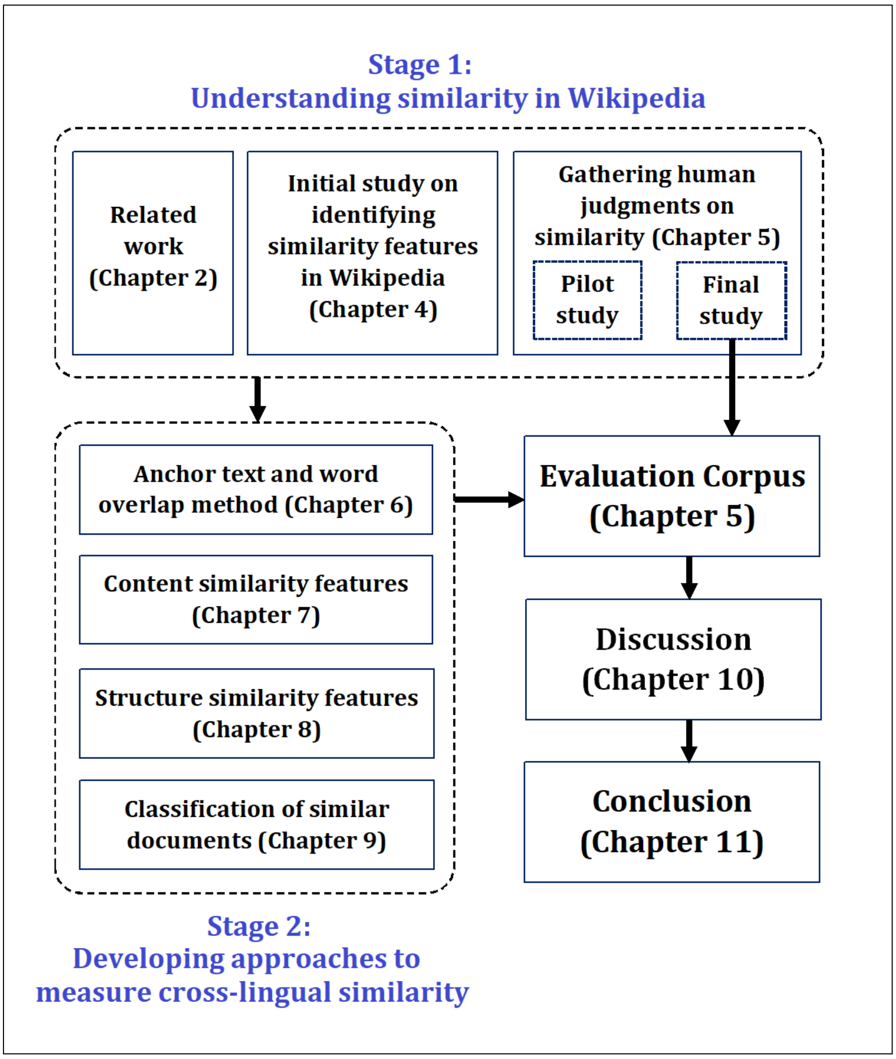

Stage 1:
Understanding similarity in Wikipedia
Related work (Chapter 2)
Initial study on identifying similarity features in Wikipedia (Chapter 4)
Gathering human judgments on similarity (Chapter 5)
Pilot study
Final study
Anchor text and word overlap method (Chapter 6)
Content similarity features (Chapter 7)
Structure similarity features (Chapter 8)
Classification of similar documents (Chapter 9)
Evaluation Corpus (Chapter 5)
Discussion (Chapter 10)
Conclusion (Chapter 11)
Stage 2:
Developing approaches to measure cross-lingual similarity

Fig. 3.1 Structure of work

gather human judgments on similarity that are then used as an evaluation benchmark (to answer Research Question 2). The similarity characteristics are used as insights to develop approaches to measure cross-lingual similarity (to answer Research Question 3), which is the second stage of this work. These approaches are then evaluated using the evaluation benchmark (i.e., the evaluation corpus) created in the first stage.

Pragmatism is the underlying philosophy for this study whereby methods are chosen for their practical applications (Creswell, 2017; Johnson & Onwuegbuzie, 2004). A mixed methods research design is considered to be most suitable approach to carry out the two stages in this study. Mixed method research combines the use of qualitative and quantitative research methods in conducting research and, therefore, was thought to provide a better understanding than insights gathered using individual methods (Creswell, 2017). In mixed method research, one method may be of emphasis compared to the other, and the phases may be carried out sequentially or concurrently (Johnson & Onwuegbuzie, 2004). Furthermore, the use of mixed-method approach has been shown to be suitable for development purposes, in which the results from one method is used to develop or inform another method (Greene, Caracelli, & Graham, 1989).

The first stage, understanding characteristics that contribute to similarity in Wikipedia, requires exploratory aspects that are best aligned with qualitative methods. However, quantitative methods were also required in this stage to gather human judgments on similarity scores of Wikipedia documents for the purpose of building the evaluation corpus. The second stage, the development of approaches, requires these approaches to be evaluated and analysed quantitatively against the evaluation corpus created in the first stage. The two stages in this work were performed sequentially; with the insights gathered in the first stage being used to develop and evaluate the approaches in the second stage. Therefore, a multi-phase mixed method research design was adopted in this study.

3.2 Initial study on investigating similarity in Wikipedia

As described in Section 2.6.1, few studies (Filatova, 2009; Patry & Langlais, 2011) have manually measured the cross-lingual similarity of interlanguage-linked articles. These findings provided evidences that the degree of similarity between interlanguage-linked articles in Wikipedia varied widely. However, these studies did not explore the similarity characteristics between these documents nor identify relevant features affecting similarity in Wikipedia. Identifying these similarity features are required in order to develop suitable approaches for measuring similarity in Wikipedia.

The author carried out an initial study to explore similarity in Wikipedia using a small set of Wikipedia interlanguage-linked articles. Three interlanguage-linked article pairs (six articles) were selected for this analysis. These articles are written in Indonesian (ID) and English (EN). These languages were selected because the author is a native speaker of Indonesian and a fluent speaker in English. The selection of these articles were performed manually to include articles containing varying lengths, varying qualities (as assessed by Wikipedia editors), and different similarity characteristics.

For each article pair, the author read the entire content of the articles, then identified and aligned similar contents in the articles. Similarity and dissimilarity analysis was performed at different levels, e.g., sentences, paragraphs, and article structures. A similar method was utilised by Gottschalk and Demidova (2017) in aligning similar paragraphs in Wikipedia. Furthermore, the author analysed the relations between Wikipedia similarity and other Wikipedia features, such as the qualities and the lengths of articles.

Considering that a small number of articles and language pairs were used in this analysis, this analysis was carried out to discover useful features in Wikipedia for identiying similarity, rather than to summarise all types of similarity and dissimilarity in Wikipedia. These initial findings were then used to complement findings from the related work (Chapter 2) and the evaluation corpus (Chapter 5) to identify relevant features to measure similarity in Wikipedia.

Although based on a very small set of documents, this task resulted in better understanding of different types of similarity and dissimilarity in Wikipedia, which was not available in previous studies. A number of features that could indicate the degree of similarity in Wikipedia articles were also identified in this task. The main findings of this analysis are described in Chapter 4.

3.3 Gathering human judgments on similarity

The aim of this task is to gather human judgments on degrees of similarity and to investigate the similarity characteristics found in a larger number of Wikipedia articles. The initial study task described in the previous section allowed similarity within a document pair to be analysed in a detailed manner. The analysis of overlapping contents within the documents, structures, and sentences, however, also proved to be a a time-consuming task which limited the evaluation to be applied for a larger number of documents. Therefore, a different approach was carried out to identify similarity features for a larger number of documents in a larger set of language pairs.

These data were gathered for two main purposes. Firstly, these data allowed for further investigation of which characteristics affect the similarity between Wikipedia articles. These findings were beneficial for selecting relevant features for developing the cross-lingual similarity approaches. Secondly, these data also provided gold-standard data that could be used for evaluating the approaches. At the time of writing, there were no suitable evaluation corpus that can be used to evaluate approaches for measuring similarity in Wikipedia articles. The work in creating this evaluation corpus (both for the pilot and final study) is described in Chapter 5. This section describes the methodology in selecting the languages, selecting the evaluation documents and creating the evaluation tasks. The background of the assessors is also described.

3.3.1 Selection of languages

In order to assess the effectiveness of techniques for computing cross-lingual similarity across different languages, seven under-resourced languages have been selected in this study: Greek (EL), Estonian (ET), Croatian (HR), Lithuanian (LT), Latvian (LV), Romanian

(RO), and Slovenian (SL). These languages had limited translation resources available and therefore would likely benefit from language-independent methods for identifying cross-language similarity. One well-resourced language, German (DE), was also chosen in order to compare performance against as a language which is well-resourced and for which high-quality translation resources are available. Each language was paired to English resulting in eight language pairs. The selected languages cover different language groups: Hellenic (EL), Baltic (LV and LT), Slavic (HR and SL), Romance (RO) and Germanic (DE and EN). Therefore, this allows the proposed approaches to be evaluated in a wide range of languages and language groups.

3.3.2 Selecting evaluation documents

To allow similarity characteristics to be investigated in detail, it was important to gather these information for document pairs with varying degrees of similarity. One approach to do this is to measure the similarity scores between all interlanguage-linked pairs in Wikipedia and to carry out a stratified sampling to purposively include document pairs for different range of similarity scores. Since most of the language pairs were under-resourced, it was not possible to use methods that rely on translation resources as they were mostly unavailable for the 7 under-resourced language pairs. Automatic approaches, such as Google Translate, was available for these language pairs during the study. However, it had a strict limitation on the amount of free translation per day[1] and therefore was infeasible to translate the entire Wikipedia.

Therefore, the author considered the use of language-independent approaches in selecting the evaluation documents. As described in the related work (Chapter 2), the *link-based bilingual lexicon method* (Adafre & de Rijke, 2006) was shown to perform with high accuracy in identifying translated sentences in Wikipedia documents, although it achieved a very low recall. Since this was the only method that had been evaluated and shown to work on Wikipedia documents, this method was applied in selecting the evalu-

[1] During this work (carried out in 2012), Google Translate had a limit of 2M characters to be translated each day (`http://developers.google.com/translate/v2/pricing/`). By the end of this study (February 2019), Google Translate was a paid service and did not provide any free translation service.

ation documents for the corpus.

Some adaptations were made into the approach, further referred to as the *anchor text and word overlap method* ($anchor + word$ method); a detailed description of this method is described in Chapter 6. Firstly, the link-based bilingual lexicon method was adapted to consider both links *and* word overlap in order to increase the recall score. Although this approach was likely to affect the accuracy (precision) of the method in finding translated sentences across documents, this adaptation was intended to allow the method to perform better in identifying similar (yet non translated) sentences. Furthermore, since the document selection process required similarity to be measured at the document level, the link-based bilingual lexicon method (originally created to identify translated sentences) was adapted to aggregate the sentence similarity scores to represent similarity at the document level. The $anchor + word$ method then was used to measure similarity across all Wikipedia articles, prior to carrying out a stratified sampling to select 100 document pairs for each language pair with varying similarity scores.

The use of this method in selecting the evaluation documents might introduce a potential bias. Firstly, the distribution of the overlap of links or word overlap in the selected documents might differ considerably to the distribution of these features in general Wikipedia articles. The purpose of this evaluation corpus, however, was not to create a corpus that represent the nature of Wikipedia. Instead, its purpose was to include document pairs with a wide range of similarity that allowed different approaches to be evaluated against human judgments, and to investigate the similarity characteristics between Wikipedia documents with different similarity scores. This purpose was further shown to be achieved using this evaluation corpus.

Another bias that might have been introduced with this approach is that the $anchor + word$ method was the only method used to pool the evaluation documents, and therefore, this approach might have advantages in the evaluation corpus. The use of more methods in the pooling of evaluation documents would have been preferred. However, at the time of carrying out the document selection task, there was no other language-independent method that have been investigated and shown to work in Wikipedia articles that could have been applied along side the $anchor + word$ method. A random sampling of Wikipedia documents was considered; however, due to the large number of non-similar document pairs in Wikipedia (Patry & Langlais, 2011; Tomás et al., 2008), this approach was likely to select a large number of document pairs with low similarity which would not have been useful for use as the evaluation corpus. Furthermore, a maximum of 100 document pairs per language pair was able to be evaluated due to the limited number of annotators. Using multiple methods to gather a larger number of documents, therefore, could not be pursued in this study.

Given this possible bias, however, the use of this approach was later shown to be able to achieve the two purposes of this evaluation corpus. Firstly, it was able to include document pairs with varying degree of similarity that allowed similarity characteristics to be investigated in more detail. Furthermore, this corpus also allow different approaches to be evaluated against human judgments to identify more approaches that can be used to identify similarity in WIkipedia. These approaches should be investigated as a future work to improve the selection process to increase the size of the evaluation corpus.

3.3.3 Assessors

A total of 23 assessors participated in the task of gathering human judgments on similarity. Seven assessors, who were either native speakers or fluent speakers in English, participated in the pilot evaluation task to assess five document pairs. In the final evaluation tasks, assessment was gathered for 800 document pairs for 8 language pairs. For each language pair, two assessors who were native speakers of the non-English language and fluent speakers in English, participated in the task; this resulted in 16 assessors participating in the final task.

All the participating assessors were partners and contributors of the ACCURAT project,[2] an EU-funded project aiming to exploit comparable corpora from the Web for the purpose of improving machine translation for languages and domains that are under-

[2] ACCURAT (Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation) project, `http://www.accurat-project.eu`.

resourced (Skadiņa et al., 2012), in which the author worked as a project partner during the beginning of the study. Their tasks in the ACCURAT project included identifying comparable documents in the Web, extracting parallel fragments within them, and using them to improve machine translation performance for under-resourced languages. All the assessors were professionals working in the area of comparable corpora and, therefore, had a great knowledge and understanding of cross-lingual similarity and comparability.

The evaluation tasks were carried out as a part of the work tasks during the ACCURAT project. The assessment was carried out anonymously. Annotators were allowed to take as many breaks as they needed during the evaluation tasks.

3.3.4 Evaluation tasks

The evaluation scheme used in the evaluation tasks required assessors to assign a similarity score to each document pair using a 5-point Likert scale (1='very different' and 5='very similar'). This scheme was tested in a pilot evaluation task, before was carried out in a final task. In the pilot task, seven assessors were given five document pairs to assess. They were asked to assign a similarity score for each document pair using a 5-point Likert Scale and to specify their reasons. The findings gathered in the pilot task were then used to improve the final evaluation task to gather similarity assessment for 800 document pairs for 8 language pairs (see Section 3.3.1). In this stage, assessors were asked to annotate more aspects of the document pair, including its similarity score, the proportion of shared content across the documents, the similarity of sentences in the shared content, and its comparability score. Similar to the pilot task, assessors were also asked to specify the document characteristics that contribute to their assigned similarity score. Sixteen assessors participated in the final task.

No training was provided for the assessors prior to their participations in the task due to two reasons. Firstly, all the assessors had an extensive work experience in retrieving similar multilingual documents for the purpose of building comparable corpora (further described in Section 3.3.3). As a result, they had an extensive understanding on cross-lingual similarity and comparability. Secondly, as described in Chapter 2 (specifically Section 2.2), there is a gap in the literature that have identified the definition of similarity in Wikipedia. As an example, since all interlanguage-linked articles in Wikipedia all describe the same topic, they would have been assigned to be "comparable" using Fung and Cheung (2004)'s definition. This example indicates that the available similarity schemes may not be able to capture the different degrees of similarity found in Wikipedia articles. The lack of suitable similarity scheme for Wikipedia articles made it difficult to provide more detail for the task or to provide specific training for the assessors as there was not enough understanding on how to define similarity specifically for Wikipedia. For the same reasons, the author did not specifically ask annotators to consider specific characteristics of the document pairs when considering the degree of similarity of the document pairs, such as to consider the similarity between the meaning of the documents (although given their background in the area, it was likely that they would have considered this aspect when judging the document pairs) or to consider the similarity between the structures of the documents. Instead, the author decided to approach the task in a qualitative manner and to rely on assessors' experiences and backgrounds to use their own understandings to assess the degree of similarity of the documents.

It was considered that the lack of specific guidelines or training for these tasks might introduce disagreements between assessors; however, the results (described in Chapter 5) show that a moderate agreement between the assessors was achieved in the evaluation task. This shows that, despite the lack of training, the assessors were still able to produce reliable judgments to be used as a gold-standard evaluation dataset. Furthermore, the feedback from assessors also allowed the different similarity characteristics that contributed to the different level of similarity, specifically in Wikipedia, to be studied further. This knowledge is important to understand prior to defining more specific annotation tasks in the future.

3.4 Development of approaches to measure similarity in Wikipedia

In this thesis, the author aims to investigate a number of approaches for identifying cross-lingual similarity. Such approaches should be applicable for languages with limited translation resources (*under-resourced languages*). Using the findings from the literature review (Chapter 2) and the previous stages (Section 3.2 and Section 3.3), the author developed four different approaches utilising different sets of features to measure similarity in Wikipedia:

1. *Anchor text and word overlap method*: The first approach is a technique that uses Wikipedia links (anchor texts) and word overlap to measure similarity on sentence level and aggregates them to represent the document score. This method is adapted from the link-based bilingual lexicon approach proposed by Adafre and de Rijke (2006). This experiment is described in Chapter 6.

2. *Content similarity features*: The second approach explores the use of features extracted from the entire document content to measure similarity at the document level. This is described in Chapter 7.

3. *Structure similarity features*: The third approach explores the use of features extracted from only the document structure. This experiment is described in Chapter 8.

4. *Classification approach*: In the final approach, these features are combined into a classifier which, given a document pair, will predict the similarity level of the document pair. This approach is described in Chapter 9.

3.4.1 Resources

Some of the features investigated in this study can be extracted without any language-independent resources, such as the ratio of lengths between the articles, word overlap, and char-n-gram overlap. Some features, such as the link overlap, utilised interlanguage links information available in Wikipedia in order to identify overlapping information across the different languages. Wiktionary, a free multilingual dictionary, was also utilised in this study to measure the structural similarity features as it contained more lexical knowledge compared to information available in Wikipedia interlanguage links (Müller & Gurevych, 2009). These two resources were utilised due to their availabilities in a large number of languages.[3]

This study focuses on the development of lightweight language-independent approaches that can be easily computed and applied for a large number of languages. As a result, they are limited to approaches that measure the lexical similarity between article content. Previous studies have shown that extending these approaches to consider semantic similarity, i.e., similarity between meanings, of the articles is likely to improve the results. This can be investigated by, for example, using WordNet to expand the content using synonyms, or using word embedding approaches to represent the content prior to measuring similarity. However, these resources are likely to increase the complexity of the methods and limit the number of languages that these approaches can be applied to. Due to these reasons, they were not investigated in the current study. Utilising these approaches, however, is a promising avenue and should be investigated for future work.

3.4.2 Language-dependent baseline

The performances of the language-independent approaches outlined in the previous section are compared against a language-dependent approach. In this study, the author used an approach based on Google Translate as the language-dependent baseline. Google Translate was chosen as it was available in all the language pairs investigated in this study and was one of the state-of-the-art translation tools that were freely available to use at the time of the study. Although Google Translate had a strict limitation on the

[3] By February 2019, Wikipedia is available in 303 languages (https://meta.wikimedia.org/wiki/List_of_Wikipedias) and Wiktionary is available in 174 languages (https://meta.wikimedia.org/wiki/Wiktionary/Table).

amount of free translation a day, it was possible to translate the entire evaluation corpus as it contain a small number of documents for each language pair.

As identified in previous literature, translation qualities of under-resourced languages may differ to highly-resourced languages. Although not specifically investigated in detail, this might have been the case for Google Translate (further explored in Chapter 6). The varying translation quality between the highly-resourced language pair and under-resourced language pairs might directly affect the performance of the baseline, with the latter having poorer performances.

Other approaches to achieve higher quality translation resources, such as hiring professional translators, were considered as these would have resulted in much higher quality translated documents that were more consistent between the language pairs. However, there were other disadvantages that were introduced by hiring human translators as the translation task would be significantly more expensive and more time-consuming. Moreover, multiple translators were required in the translation task in order to accommodate the different human interpretation which might have been introduced by the individual translators. This approach was, therefore, infeasible to be carried out in this study.

3.5 Evaluation

Numerous methods have been used to evaluate the similarity of information in Wikipedia at the sub-document level, e.g., evaluating the accuracy of extracted parallel phrases (Yu & Tsujii, 2009) or accuracy of extracted parallel sentences (Adafre & de Rijke, 2006; Smith et al., 2010). Evaluation of similarity at the document level, on the other hand, was mostly done extrinsically. For example, when similar Wikipedia documents were extracted as resources for specific applications (such as building SMT or bilingual lexicons), evaluation was performed by assessing the performance of the resulting applications, such as the SMT's performance (Munteanu & Marcu, 2005) or the quality of translations produced by the new bilingual lexicon (Hersh et al., 2004).

The above methods are useful for purposes such as investigating the usability of Wikipedia articles in performing a number of tasks. However, an intrinsic evaluation was required to specifically evaluate the accuracy of the approaches in identifying similar documents. Therefore, in this study, the approaches are evaluated using the evaluation corpus (Chapter 5), i.e., to identify how well the approaches correlated to human judgments.

Most approaches are evaluated by calculating the Spearman rank correlation scores between the automatic scores and the mean of human judgment similarity scores in the 5-point Likert scale. The evaluation of the classification approach (fourth approach) was also carried out using a set of metrics: Correctly Classified Instances, F_1-measure, Area under Receiver Operating Characteristic (AUROC) and Root Mean Squared Error (RMSE). Results from the regression approach was evaluated using RMSE and Pearson's Correlation Coefficient. Significance was calculated at p<0.05. Failure analysis was also carried out to analyse cases that the automatic approaches could not correctly identify.

3.6 Wikipedia corpus

This section describes the pre-processing tasks carried out in the Wikipedia corpus, and the statistics of the Wikipedia corpus used in this study.

3.6.1 Pre-processing of Wikipedia corpus

Wikipedia articles are freely available for download in the form of Wikipedia dumps[4] with newer versions of the database dumps published at least once a month. The content of all articles in a language version (such as English) are contained within a single XML file. This also contains additional metadata, such as document ID, interlanguage links,[5] redirection page, comment, and contributor.

[4] One can download all Wikipedia articles for a specified language using the following link: `http://dumps.wikimedia.org/[lang]wiki/[lang]wiki-[dateVersion]-pages-articles.xml.bz2`. For example, the link for an English Wikipedia dump is `http://dumps.wikimedia.org/enwiki/enwiki-20130503-pages-articles.xml.bz2`.

[5] Since 2013, interlanguage-links information is no longer included in the XML dump file but can still be accessed by processing the SQL dump of interlanguage-links which is available in the following link: `http://dumps.wikimedia.org/[lang]wiki/[date]/[lang]-wiki-[date]-langlinks.sql.gz`.

A number of open-source tools (e.g. JWPL (Zesch et al., 2007)) are available to process the XML file, including to clean the data and extract content in a plain-text format before loading the data into a database. Whilst these tools work well in extracting content in the entire documents, modifying them to suit the purpose of this research (e.g., extracting the section headings only, splitting documents into sentences) was found to be a complicated task. Therefore, the author instead developed a Wikipedia extractor tool which has been enhanced throughout this study in order to suit the research.

Document pre-processing

For the purpose of this study, only the textual contents of Wikipedia articles and information about the interlanguage links were required. Therefore, a set of pre-processing methods were implemented to filter out irrelevant metadata information and to extract the relevant content. These pre-processing methods are shown in Figure 3.2. The processes are described below with each output shown in italics:

1. **Document extraction**. First, an extraction tool[6] was developed to extract relevant information from the Wikipedia dump. In this process, document ID and titles of interlanguage-linked articles were extracted in order to build the *list of interlanguage-linked articles.* Pages which were not listed as main Wikipedia articles, such as articles describing Wikipedia users, categories, or discussions of a topic, were also deleted. Main article contents were then extracted and written into separate files, i.e., each file contained the content of one article. An example of English articles that has been pre-processed[7] is shown in Figure 3.3.

2. **Bilingual lexicon extraction**. Given the list of interlanguage-linked articles as an output from the previous process, an analysis was conducted to extract a *bilingual lexicon* by pairing titles of the interlanguage-linked articles from the source and the target language as shown in Figure 3.4, enabling a type of translation resources

[6] This tool is a part of the WikipediaRetrieval package which is available to download from `http://www.accurat-project.eu/index.php?p=accurat-toolkit`.

[7] `http://en.wikipedia.org/wiki/University_of_Sheffield`, accessed on 15 January 2013

to be built without any external linguistic resources. Most interlanguage-links pair articles of the same topic, however, a small number of irregularities such as incorrect links do appear in Wikipedia. Moreover, if the same topic does not exist in another language, sometimes an article is linked to the most similar topic instead (such as hyponym or hypernym). However, since the number of these cases is low (Adafre & de Rijke, 2006), this study assumes all interlanguage-linked documents to be correctly paired to each other. Title of redirection pages were not used in this study as they were shown to reduce the translation accuracy in previous study (Erdmann et al., 2009). On occasions where the titles for both language versions were exactly the same, such as articles about named entities (e.g. "Barack Obama" or "Nokia", which were spelled the same both in English and Slovenian) or dates (such as "1961"), these duplicate titles were eliminated from the lexicon.

3. **Document filtering and sentence splitting**. As shown in Figure 3.3, the content of a Wikipedia article contains various information, such as infoboxes, paragraphs, images, tables and lists of references. The main content (referred to as the main paragraphs) contains written text structured in sentences. Content included in infoboxes or tables is generally shorter in length (e.g., dates or named entities). In this study, the author only focused on similarity of the main content of the articles and filtered out information which was not included in the main paragraphs. The output from this process is referred to as *text and links documents*.[8] These documents are the main input for the language-independent similarity approach investigated in Section 6.2. An example of a document in this version is shown in Figure 3.5.

4. **Link removal**. In the final process, all links found in the documents were removed and replaced with the relevant plain text. An example of these *plain-text documents* is shown in Figure 3.6. These plain-text documents were also used as the basis for the manual cross-language similarity judgment (further described in Chapter 5).

[8] Links found in the documents were preserved as they would be required as a feature for identifying similarity. Links are shown as texts which are surrounded by the [[and]] brackets. Links which appears with a '|' character separating terms represents the referred article title and the document text as it appears to the user.

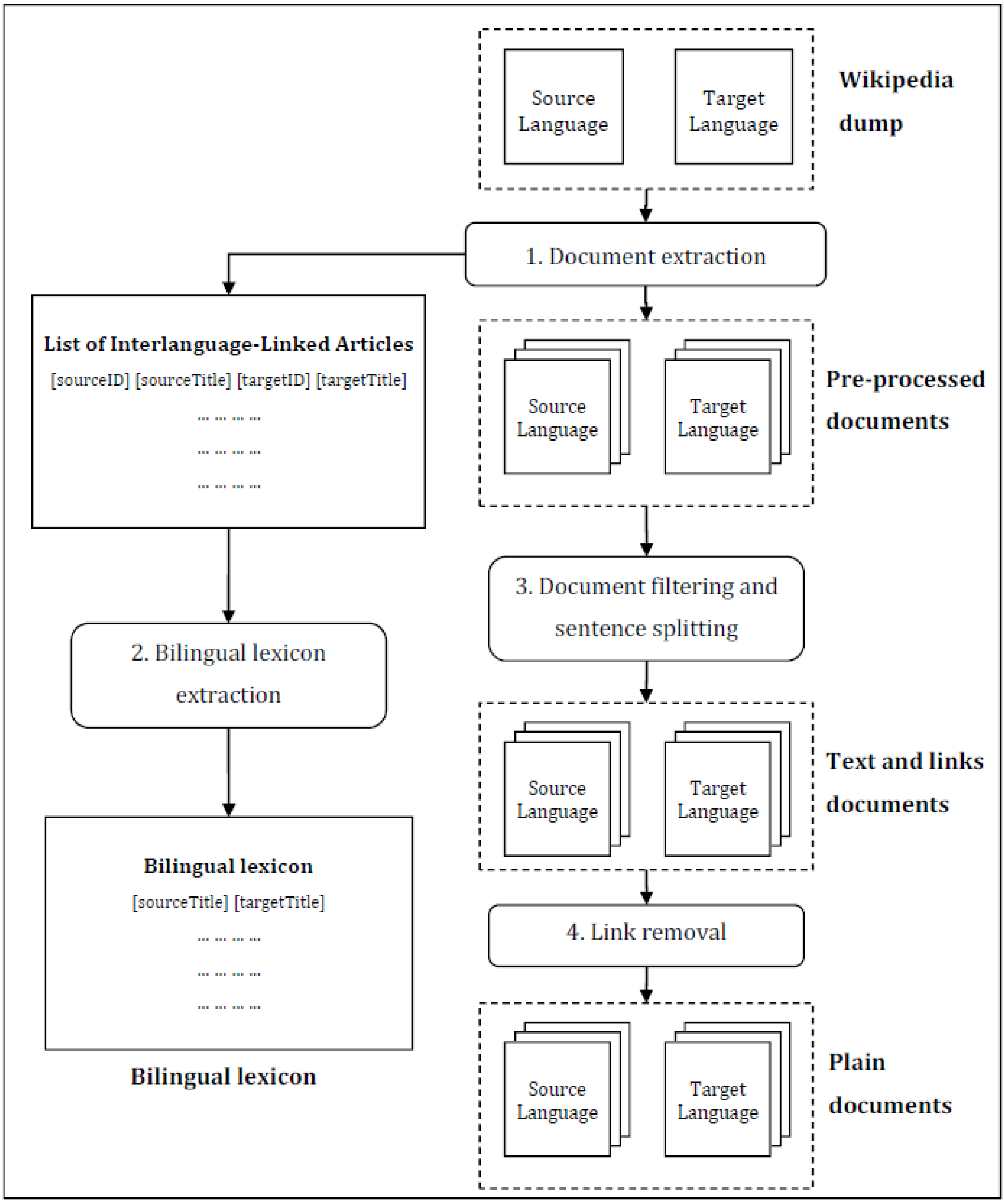
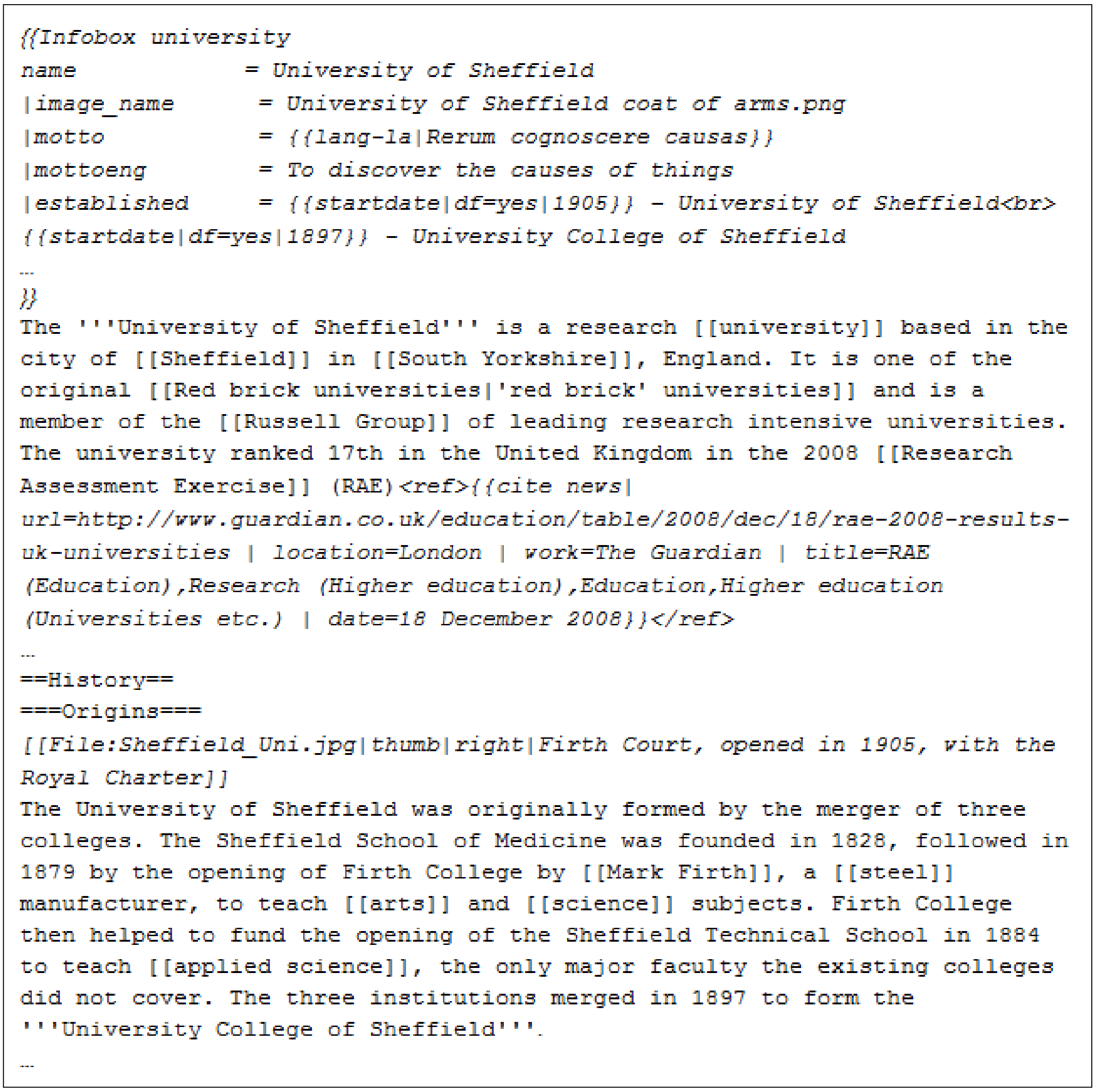
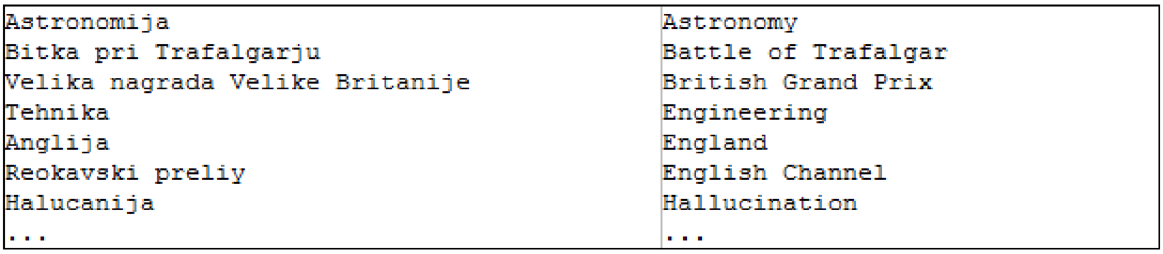


Fig. 3.2 Pre-processing methods



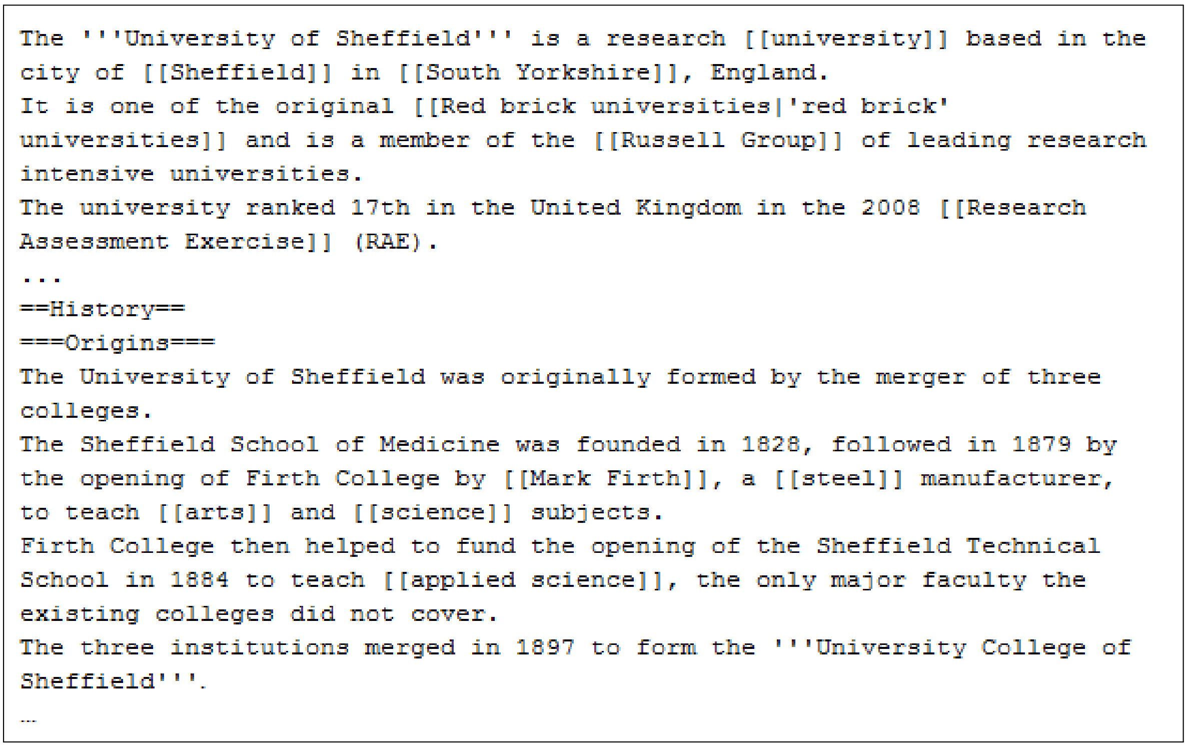
```
{{Infobox university
name             = University of Sheffield
|image_name       = University of Sheffield coat of arms.png
|motto            = {{lang-la|Rerum cognoscere causas}}
|mottoeng         = To discover the causes of things
|established      = {{startdate|df=yes|1905}} – University of Sheffield<br>
{{startdate|df=yes|1897}} – University College of Sheffield
...
}}
The '''University of Sheffield''' is a research [[university]] based in the
city of [[Sheffield]] in [[South Yorkshire]], England. It is one of the
original [[Red brick universities|'red brick' universities]] and is a
member of the [[Russell Group]] of leading research intensive universities.
The university ranked 17th in the United Kingdom in the 2008 [[Research
Assessment Exercise]] (RAE)<ref>{{cite news|
url=http://www.guardian.co.uk/education/table/2008/dec/18/rae-2008-results-
uk-universities | location=London | work=The Guardian | title=RAE
(Education),Research (Higher education),Education,Higher education
(Universities etc.) | date=18 December 2008}}</ref>
...
==History==
===Origins===
[[File:Sheffield_Uni.jpg|thumb|right|Firth Court, opened in 1905, with the
Royal Charter]]
The University of Sheffield was originally formed by the merger of three
colleges. The Sheffield School of Medicine was founded in 1828, followed in
1879 by the opening of Firth College by [[Mark Firth]], a [[steel]]
manufacturer, to teach [[arts]] and [[science]] subjects. Firth College
then helped to fund the opening of the Sheffield Technical School in 1884
to teach [[applied science]], the only major faculty the existing colleges
did not cover. The three institutions merged in 1897 to form the
'''University College of Sheffield'''.
...
```

Fig. 3.3 An example of *pre-processed version* of an English document



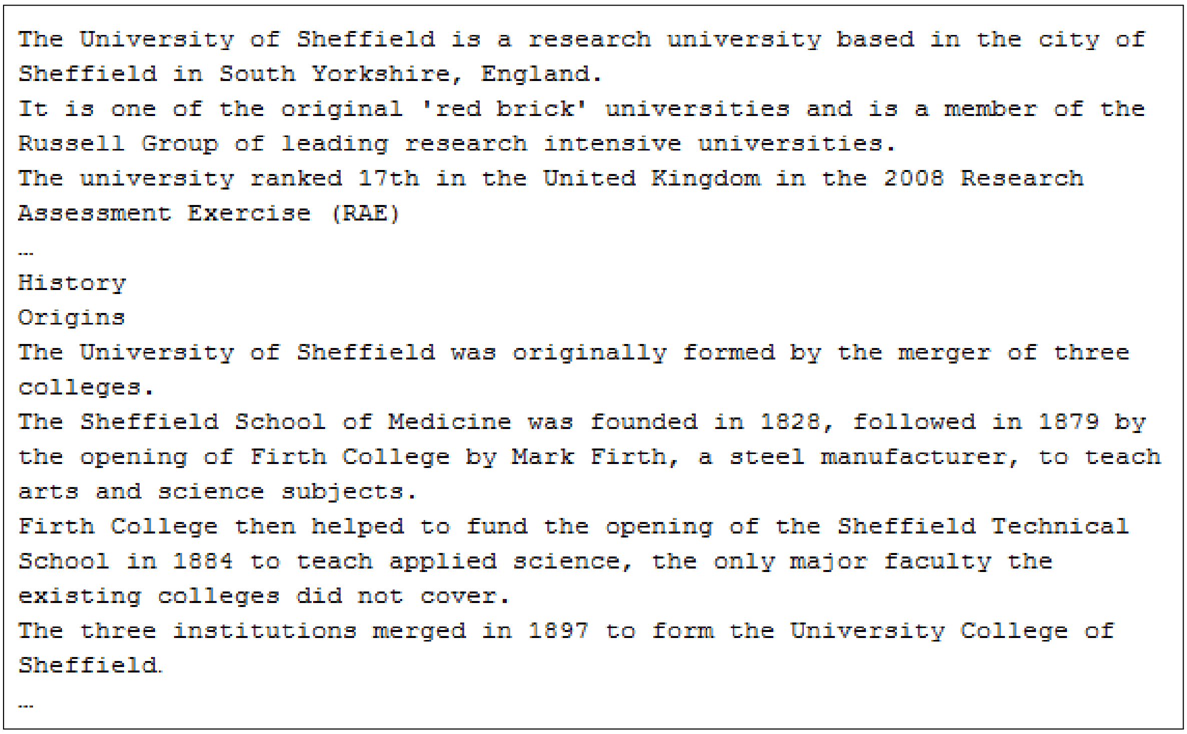
Astronomija	Astronomy
Bitka pri Trafalgarju	Battle of Trafalgar
Velika nagrada Velike Britanije	British Grand Prix
Tehnika	Engineering
Anglija	England
Reokavski preliy	English Channel
Halucanija	Hallucination
...	...

Fig. 3.4 An extract of Slovenian-English bilingual lexicon



```
The '''University of Sheffield''' is a research [[university]] based in the
city of [[Sheffield]] in [[South Yorkshire]], England.
It is one of the original [[Red brick universities|'red brick'
universities]] and is a member of the [[Russell Group]] of leading research
intensive universities.
The university ranked 17th in the United Kingdom in the 2008 [[Research
Assessment Exercise]] (RAE).
...
==History==
===Origins===
The University of Sheffield was originally formed by the merger of three
colleges.
The Sheffield School of Medicine was founded in 1828, followed in 1879 by
the opening of Firth College by [[Mark Firth]], a [[steel]] manufacturer,
to teach [[arts]] and [[science]] subjects.
Firth College then helped to fund the opening of the Sheffield Technical
School in 1884 to teach [[applied science]], the only major faculty the
existing colleges did not cover.
The three institutions merged in 1897 to form the '''University College of
Sheffield'''.
...
```

Fig. 3.5 An example of *text and links version* of an English document



```
The University of Sheffield is a research university based in the city of
Sheffield in South Yorkshire, England.
It is one of the original 'red brick' universities and is a member of the
Russell Group of leading research intensive universities.
The university ranked 17th in the United Kingdom in the 2008 Research
Assessment Exercise (RAE)
...
History
Origins
The University of Sheffield was originally formed by the merger of three
colleges.
The Sheffield School of Medicine was founded in 1828, followed in 1879 by
the opening of Firth College by Mark Firth, a steel manufacturer, to teach
arts and science subjects.
Firth College then helped to fund the opening of the Sheffield Technical
School in 1884 to teach applied science, the only major faculty the
existing colleges did not cover.
The three institutions merged in 1897 to form the University College of
Sheffield.
...
```

Fig. 3.6 An example of *plain text version* of an English document

3.6.2 Corpus statistics

For each language, one Wikipedia dump was downloaded and extracted at the beginning of this study (November 2009-March 2010). Articles containing fewer than 5 words were filtered out as these rarely contain useful sentences. The number of articles available for each language is shown in Table 3.1.

Although these dumps may be considered to be old, they still represent the nature of the current Wikipedia for the following reasons. Firstly, the Wikipedia dumps used in this corpus contain the same formats as the current Wikipedia dumps, which means that all features investigated in this study using this corpus are also available in the current Wikipedia versions. As a result, the approaches proposed in this study are also suitable for use in the newest version of Wikipedia. Secondly, previous studies have showed that the differences between these dumps and the current (or more up-to-date) Wikipedia dumps are limited to a few aspects, such as the number of articles, the development of articles, and the number of contributors, which are expected to have increased over time (Clark et al., 2009; Hu et al., 2007; Voss, 2005). However, the author believes that these changes do not significantly affect the aspects studied in this work, i.e., how the different similarity characteristics influence the degrees of similarity, and the effectiveness of approaches in measuring similarity in Wikipedia.

As shown in Table 3.1, the number of articles differ considerably between languages: the under-resourced languages have significantly lower numbers of documents com-

Table 3.1 Version of Wikipedias

No	Language Code	Language Name	Downloaded Version	Number of Articles
1	DE	German	6 February 2010	1,036,144
2	EL	Greek	15 March 2010	49,275
3	EN	English	3 November 2009	3,243,312
4	ET	Estonian	10 March 2010	72,231
5	HR	Croatian	11 March 2010	81,366
6	LT	Lithuanian	9 March 2010	102,407
7	LV	Latvian	13 March 2010	26,297
8	RO	Romanian	14 March 2010	141,284
9	SL	Slovenian	7 March 2010	85,709

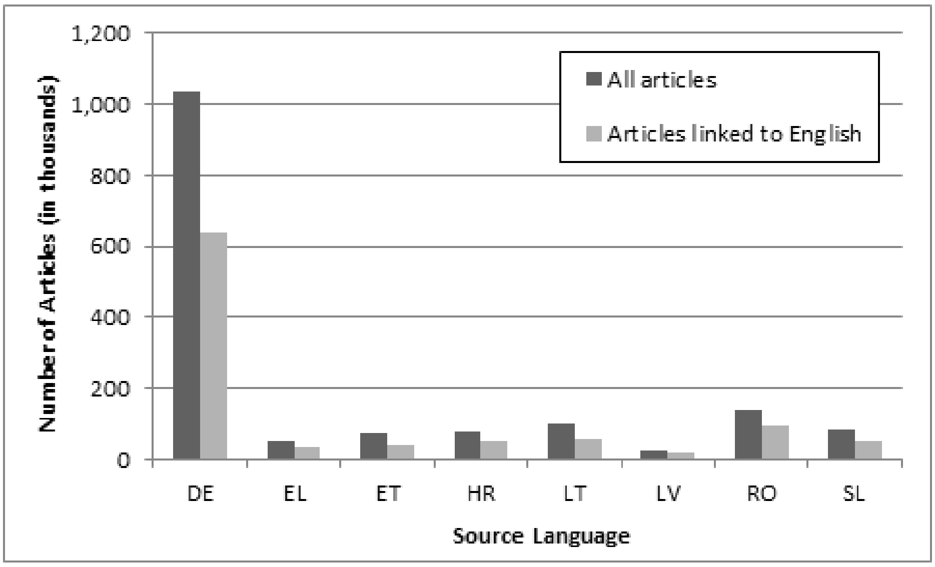


Fig. 3.7 Proportion of articles which contain interlanguage-links to English

pared to German and English. The Latvian Wikipedia is the smallest dataset with just over 26,000 articles (representing around 2.5% of the entire German Wikipedia). The significant differences between numbers of articles for the chosen languages are further shown in Figure 3.7, which also displays the number of interlanguage-linked articles to English.

Since the focus of this study is to investigate similarity between interlanguage-linked articles, those with no interlanguage-links to English were filtered out. As described in Section 3.6.1, the author created a bilingual lexicon by pairing titles of interlanguage-linked articles. If the article titles were the same for the language pair, they were discarded from the bilingual lexicon. For example, Table 3.3 shows article titles about "Barack Obama" in different languages. Apart from Greek (EL) and Latvian (LV), the titles for the remaining languages are spelled the same as the English article; therefore, they were deleted from the bilingual lexicon. The number of interlanguage-linked articles and the size of bilingual lexicon for each language pair are shown in Table 3.2. Since the creation of bilingual lexicon relies directly on the number of interlanguage-linked articles, the sizes of lexicon for the different language pairs also differ significantly, e.g., the size of German lexicon is more than 10 times larger than the Latvian one.

Table 3.2 Size of initial Wikipedia datasets

Language pair	Number of interlanguage-linked articles	Number of entries in bilingual lexicon
DE-EN	637,382	181,408
EL-EN	36,752	28,294
ET-EN	42,008	22,645
HR-EN	51,432	26,804
LT-EN	57,954	41,497
LV-EN	21,302	15,511
RO-EN	97,815	35,774
SL-EN	51,332	25,101

Table 3.3 Titles of “Barack Obama” articles in different language editions

Language	Article title
DE	Barack Obama
EL	Μπαράκ Ομπάμα
EN	**Barack Obama**
ET	Barack Obama
HR	Barack Obama
LT	Barack Obama
LV	**Baraks Obama**
RO	Barack Obama
SL	Barack Obama

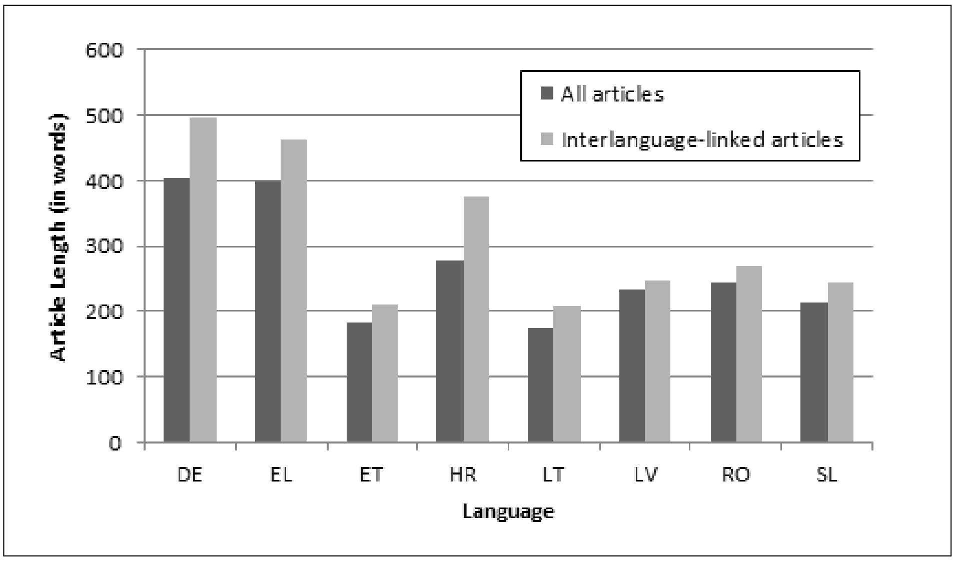


Fig. 3.8 Comparison of average size of Wikipedia articles

Analysing the average article lengths in different languages (Figure 3.8) shows that articles from under-resourced languages were found to be shorter in length than the highly-resourced language (DE). A comparison between the sizes of interlanguage-linked articles and all articles also shows that the length of interlanguage-linked articles is longer in average; this suggests that articles linked to English are more developed than others.

Chapter 4

Identifying Similarity Features in Wikipedia

The literature review described in Chapter 2 (especially Section 2.6) have identified that Wikipedia interlanguage-linked articles contain different degrees of similarity. The document characteristics that contribute to the similarity (or dissimilarity), however, have not been addressed in previous work. In this chapter, the author carried out an initial study to further understand the concept of similarity in Wikipedia, specifically, what features can be used to measure similarity between Wikipedia interlanguage-linked articles.

4.1 Background

Identifying features that contribute to similarity in Wikipedia is important in order to develop suitable methods for measuring similarity in Wikipedia. Previous works (e.g., Barker and Gaizauskas (2012); Braschler and Schäuble (1998); Fung and Cheung (2004); Skadiņa et al. (2012)) have addressed a detailed criteria that influences the similarity of news articles or Web articles (described in Section 2.2). Although few studies have aimed to identify similarity in Wikipedia in general (Filatova, 2009; Otero & López, 2010; Patry & Langlais, 2011; Tomás et al., 2008; Yu & Tsujii, 2009), none of these studies have analysed the contributing aspects behind the Wikipedia similarity (or dissimilarity).

Previous studies have also identified a number of features that can be used to measure document similarity in general, such as structure similarity (Wan, 2007) and word length (Resnik & Smith, 2003). Their applicability to identifying similarity in Wikipedia articles, however, has yet to be studied. Other work has also investigated Wikipedia-related feature, such as document quality (Hu et al., 2007), but did not assess if this aspect contributes to the similarity at the document level.

Various literature have utilised Wikipedia to extract bilingual resources at different granularity level, such as phrases (Erdmann et al., 2009), sentences (Adafre & de Rijke, 2006; Smith et al., 2010; Tomás et al., 2008), and paragraphs or text passages (Gottschalk & Demidova, 2017). The relations between these similarity aspects and the document similarity, however, have not yet been investigated.

Are similar documents more likely to have similar paragraphs or sentences? Are documents with the highest quality (and assessed to contain complete information about the concepts) more likely to be similar cross-lingually? These are some of the questions that the author aimed to investigate in this study. The findings in this chapter contribute to a further understanding of characteristics of similarity in Wikipedia articles that is addressed in the first research question:

RQ1. What are the characteristics of similar interlanguage-linked articles in Wikipedia?

Moreover, this initial study is also aimed to further understand the degree of similarity (and dissimilarity) between Wikipedia articles, as these were not available in previous studies. Furthermore, this study is also carried out to gain insights on the design of the annotation task for building the evaluation corpus (further described in Chapter 5).

Three pairs of interlanguage-linked articles were analysed in this study. These article pairs were manually selected because they contained varying lengths, qualities, and similarities. This initial study was limited to three pairs of articles due to the time-consuming analysis required to identify the similarity of these articles at different granularities (further described in Section 4.2).

Although the number of documents used in the initial study was very small, the different degrees of similarity between these three document pairs were able to provide some insights on the different types of similarity in Wikipedia and provided more information regarding the features that were not available in previous study. Due to the limited size, these analysis are to be used to complement the findings in the related work (described in Chapter 2) and the creation of evaluation corpus (described in Chapter 5), which involved a larger number of document pairs and language pairs.

4.2 Similarity analysis method

The author analysed the similarity within Wikipedia interlanguage-linked articles and how it relates to the aspects described in previous studies, such as similarity of sentences, phrases, structures, and article quality. The method carried out by the author can be summarised by the following:

- Step 1: Read both interlanguage-linked articles
- Step 2: Find and align similar contents in the articles
  - Step 2a: Analyse similarity between structures (Wan, 2007). The approach proposed by Wan (2007) required identifying the topic in each paragraph, prior to measuring the similarity between the topics. In Wikipedia articles, however, articles are often structured into different section headings. Therefore, the author utilised these section headings information in carrying out the structure similarity analysis.
  - Step 2b: Analyse similarity between paragraphs (Gottschalk & Demidova, 2017)
  - Step 2c: Analyse similarity between sentences (Adafre & de Rijke, 2006)
- Step 3: Analyse the relations between document quality (Hu et al., 2007) and word length (Resnik & Smith, 2003) and the similarity degree of the document pair

The first two steps were similar to the approach used by Gottschalk and Demidova (2017) to manually identify and extract parallel text passages from Wikipedia. The difference between the two approaches is that Gottschalk and Demidova (2017) focused on identifying similar text passages only. Their method, therefore, did not include identifying sentence similarity and structure similarity within the Wikipedia documents. They also did not carry out the last step (Step 3).

For Step 2 and Step 3, the author also identified examples of similar and dissimilar features found in the article. These examples are valuable to further understand the nature of Wikipedia articles. As far as the author's aware, no Wikipedia specific examples of similarity in different granularity levels have been described in previous studies.

4.3 Dataset

In this initial study, the author manually selected three interlanguage-linked article pairs (six articles) in Indonesian (ID) and English (EN). This language pair were selected because the author was a native speaker of Indonesian and a fluent speaker in English. The selection of these articles were performed manually to include articles containing varying lengths, qualities and degrees of similarity.

The varying length differences (in number of words) across the three article pairs as shown in Table 4.1. Article pairs 1 and 2 have similar lengths with the lengths of the Indonesian versions being 96% and 91% of the EN versions, respectively. In contrast, article pair 3 is shown to be extremely different lengthwise; the EN article of "Frédéric Chopin" is five times as long as the ID article, which are 10,527 words and 2,115 words, respectively. In all three cases, the ID versions of the articles are shorter than the EN versions.

As described in Section 2.6, Wikipedia editors have assessed the quality of some ar-

Table 4.1 Selected articles

Article Pair Id	English (EN) Article		Indonesian (ID) Article		Date Version
	Title	Length	Title	Length	
1	Indonesia**	5,564	Indonesia**	5,340	17 June 2014
2	England*	11,766	Inggris	10,743	13 May 2014
3	Frédéric Chopin**	10,527	Frédéric Chopin	2,115	13 May 2014

*Note: * represents good quality articles, ** represents featured articles.*

ticles in Wikipedia. By 19 January 2016, over 4,500 articles (less than 0.1% of all 5 million English Wikipedia articles), have been assessed to be *featured articles* (the highest article quality in Wikipedia).[1] Featured articles represent the best quality of articles in Wikipedia and were defined to have the following characteristics: well-written, comprehensive, well-researched, neutral and stable,[2] and that no further content addition (unless new information was made available) was necessary for these articles.[3] *Good quality articles*,[4] on the other hand, were assessed as having similar characteristics to featured articles, but some topics within the articles may have had weak or missing contents. Over 23 thousand English articles (0.5%) have been considered to be of good quality.

In this analysis, three of the selected articles were assessed as 'featured articles', one as a 'good article', while the remaining two articles had not yet been assessed by the time this study was carried out.

4.4 Initial findings

The author highlights the initial findings for each step in this section.

4.4.1 Similarity between structures

A Wikipedia article often contains a "Contents" section, listing all the headings and subheadings that appear in the article, similar to the 'table of contents' of a book or a report. This information displays how the various topics appearing in the article are structured. This structure information is further analysed to identify whether interlanguage-linked articles have any similarity at the structure/heading level.

As shown in Tables 4.2 and 4.3, both article pairs 1 and 2 show a large proportion of similar headings. In article pair 1 (Table 4.2), a large number of the sections can be

[1] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles accessed on 16 January 2016

[2] https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria, accessed on 16 January 2016

[3] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Wikipedia/Assessment, accessed on 16 January 2016

[4] https://en.wikipedia.org/wiki/Wikipedia:Good_articles, accessed on 16 January 2016

Table 4.2 Aligned section headings in article pair 1: "Indonesia"

EN Version	**ID Version (translations shown in brackets)**
1 Etymology	1 Etimologi *(Etymology)*
2 History	2 Sejarah *(History)*
	— 2.1 Sejarah awal *(Early history)*
	— 2.2 Kolonialisme *(Colonialism)*
3 Government and politics	— 2.3 Indonesia merdeka *(Indonesian's independence)*
4 Foreign relations and military	3 Politik dan pemerintahan *(Politics and government)*
5 Administrative divisions	4 Pembagian administratif *(Administrative divisions)*
6 Geography	5 Geografi *(Geography)*
	— 5.1 Sumber daya alam *(Natural resources)*
7 Biota and environment	6 Pendidikan *(Education)*
8 Economy	7 Ekonomi *(Economy)*
	8 Peringkat internasional *(International rank)*
9 Demographics	9 Demografi *(Demographics)*
10 Culture	10 Kebudayaan *(Culture)*
	— 10.1 Pertunjukan *(Performances)*
	— 10.2 Busana *(Fashion)*
	— 10.3 Arsitektur *(Architectures)*
	— 10.4 Olahraga *(Sports)*
	— 10.5 Seni musik *(Music)*
	— 10.6 Boga *(Culinary)*
	— 10.7 Perfilman *(Films)*
	— 10.8 Kesusastraan *(Literatures)*
	— 10.9 Bahasa *(Language)*
11 See also	11 Lingkungan hidup *(Environment)*
12 Notes	12 Lihat pula *(See also)*
13 References	13 Referensi *(References)*
14 External links	14 Pranala luar *(External links)*

aligned as they share exact title names (after translation). The order of the sections, however, are slightly different to each other. Headings in article pair 2 (Table 4.3), on the other hand, can be aligned to each other and follow the same order. In both cases, there are sections not available in other language in both articles. Furthermore, in article pair 1, the ID version contains a more detailed structure, with some sections consisting of a number of sub-sections; these were not available in the EN version.

On the contrary, article pair 3 is shown to be extremely different to each other (shown in Table 4.4). First of all, both articles do not share the same section heading names, although some sections discuss similar topics, e.g. "Life" (EN) vs "Biography" (ID), or "Music" (EN) and "Chopin's compositions" (ID). Furthermore, a large number of topics appearing in one language do not appear in the other language version. For example, the EN article describes Chopin's style of composition and TV documentaries that are not

Table 4.3 Aligned section headings in article pair 2: "England"

EN Version	**ID Version (translations shown in brackets)**
1 Toponymy	1 Etimologi *(Etymology)*
2 History	2 Sejarah *(History)*
3 Governance	3 Pemerintahan *(Government)*
4 Geography	4 Geografi *(Geography)*
5 Economy	5 Ekonomi *(Economy)*
	6 Ilmu pengetahuan dan teknologi *(Science and technology)*
	7 Transportasi *(Transportation)*
6 Healthcare	8 Kesehatan *(Health)*
7 Demography	9 Demografi *(Demography)*
8 Education	10 Pendidikan *(Education)*
9 Culture	11 Kebudayaan *(Culture)*
10 Sports	12 Olahraga *(Sports)*
11 National symbols	13 Simbol nasional *(National symbols)*
12 See also	14 Lihat pula *(See also)*
13 Notes	15 Catatan *(Notes)*
14 References	16 Referensi *(References)*
15 External links	17 Pranala luar *(External links)*

contained in the ID article. Meanwhile, the ID article describes a number of his compositions, which are not described in the EN version.

Upon analysing the structures of these article pairs, different types of dissimilarities were found. Given two documents D_1 and D_2 written in language l_1 and l_2, respectively, the author was able to identify different alignments:

- A heading in D_1 is aligned to a heading in D_2.
- A heading in D_1 is represented as a sub-heading in D_2.
- A heading or sub-heading in D_1 may not appear in D_2, but the contents exist in both articles. E.g., contents under the heading 'Demography' and sub-heading 'Population' in D_1 both appear under the heading 'Demography' in D_2.
- A section heading and its content appear in D_1 but not in D_2.

Table 4.4 Aligned section headings in article pair 3: “Frédéric Chopin”

English version	**Indonesian Version (translations shown in brackets)**
1 Life	1 Biografi (*Biography*)
— 1.1 Childhood	2 Awal karier (*Early career*)
— 1.2 Education	3 Pertemuan dengan George Sand (*Meet with George Sand*)
— 1.3 Youth	4 Komposisi-komposisi Chopin (*Chopin's compositions*)
— 1.4 Paris	— 4.1 Etudes
— 1.5 George Sand	— — 4.1.1 Etudes Op. 10 & 25
— 1.6 Final years	— 4.2 Mazurka
— 1.7 Death	— 4.3 Polonaise
— 1.8 Funeral	— 4.4 Impromptu
2 Memorials	— 4.5 Ballade
3 Music	— — 4.5.1 Ballade in F, Op. 38
— 3.1 Publishing	— — 4.5.2 Ballade in Ab Op. 47
— — 3.1.1 Opus numbering	— — 4.5.3 Ballade in F Minor, Op. 52
— 3.2 Influence	— 4.6 Sonata
— 3.3 Style	— 4.7 Sonata Op. 4 No. 1 in C minor
— 3.4 Rubato	— 4.8 Sonata in Bb minor, Op. 35 No. 2
— 3.5 Romanticism	— 4.9 Sonata in B Minor, Op. 58
— 3.6 Nationalism	— 4.10 Sonata in G Minor, Op. 65
4 TV documentaries	— 4.11 Concerto
5 Fiction	— 4.12 Scherzo
6 See also	— — 4.12.1 Scherzo in B Minor, Op. 20
7 Notes	— — 4.12.2 Scherzo in Bb Minor, Op. 31
8 Bibliography	— — 4.12.3 Scherzo in C# Minor, Op. 39
9 External links	— — 4.12.4 Scherzo in E, Op. 54
— 9.1 Biographies	— 4.13 Nocturne
— 9.2 Music scores	— 4.14 Prelude
— 9.3 Recordings	— 4.15 Waltz
— 9.4 Miscellaneous	— 4.16 Fantasia
	— 4.17 Lain-lain
	— — 4.17.1 Variations on La ci darem La mano, Op. 2
	— — 4.17.2 Krakowiak, Op. 14
	— — 4.17.3 Allegro de Concert, Op. 46
	— — 4.17.4 Berceuse14 in Db, Op. 57
	— — 4.17.5 Barcarolle15 in F#, Op. 60
	— — 4.17.6 Rondo (Dua Piano), Op. Posthumous
	— — 4.17.7 Variations sur un Air national allemande in E, Op. Posthumous
	— — 4.17.8 Chansons lithuanienne, Op. Posthumous
	— — 4.17.9 Fugue in A Minor (1842), Op. Posthumous
	— — 4.17.10 Largo in Eb, Op. Posthumous
	— — 4.17.11 Meine Freunde, Op. Posthumous
	— — 4.17.12 19 Polish Songs for Voice and Piano, Op. Posthumous
	5 Referensi (*References*)
	6 Pranara luar (*External links*)

4.4.2 Similarity between paragraphs

The previous section shows that both article pair 1 and 2 have similar structures as most of the sections can be aligned to each other based on the titles. In the same two document pairs, the author further found that most of the content can also be aligned at the paragraph level. However, the contents at a paragraph level often do not correspond at the translation level, likely to be the results of additions or deletions of information (e.g., phrases or sentences) in one language version. The type of similarities and dissimilarities found at the paragraph level are listed as follow:

- A paragraph in D_1 and D_2 represent the same content and correspond to a sentence by sentence translation.
- A paragraph in D_1 and D_2 represent the same content, but the sentences are split differently (see Table 4.5).
- A paragraph in D_1 contains additional information compared to its aligned paragraph in D_2 (see Table 4.6).
- A paragraph in D_1 contains the same content to more than one paragraphs in D_2. An example is shown in Table 4.7.

4.4.3 Similarity between sentences

Upon analysing the similarity at the sentence level, the author identified similar findings, i.e., that content in article pairs 1 and 2 can easily be aligned at the sentence level as they were found to be translations of each other. However, some dissimilarities were also found at this level, likely to be the results of additions or deletions of phrases in one language version. The similarities and differences found at the sentence level are listed as follows:

Table 4.5 Example of same content but sentences were split differently (paragraph level)

Lang	**Paragraph**
EN	"{From 1453 to 1487 civil war between two branches of the royal family occurred - the Yorkists and Lancastrians}^{t1} - {known as the Wars of the Roses.}^{t2} {Eventually it led to the Yorkists losing the throne entirely to a Welsh noble family the Tudors, a branch of the Lancastrians}^{t3} {headed by Henry Tudor who invaded with Welsh and Breton mercenaries,}^{t4} {gaining victory at the Battle of Bosworth Field where the Yorkist king Richard III was killed.}^{t5}"
ID (EN-translation)	"{From 1453 to 1487, civil war between two branches of the royal family occured - the Yorkists and Lancastrians.}^{t1} {This war is known as the Wars of the Roses,}^{t2} {which led to the Yorkists losing the throne entirely to a Welsh noble family the Tudors, a branch of the Lancastrians.}^{t3} {The Tudors, headed by Henry Tudor, invaded Britain with Welsh and Breton mercenaries.}^{t4} {They gained victory at the Battle of Bosworth Field where the Yorkist king Richard III was killed.}^{t5}"
ID	"{Dari tahun 1453-1487, perang saudara antara dua wangsa keluarga kerajaan terjadi (Wangsa York dan Wangsa Lancaster).}^{t1} {Perang ini dikenal dengan sebutan Perang Mawar,}^{t2} {yang berakhir dengan kekalahan York dan harus merelakan takhta jatuh ke tangan Wangsa Tudor dari Wales, yaitu penerus Lancaster.}^{t3} {Tudor, yang dipimpin oleh Henry Tudor, menginvasi Inggris bersama tentara-tentara Breton dan Wales.}^{t4} {Tentara ini berhasil memperoleh kemenangan dalam Pertempuran Bosworth dengan tewasnya raja York terakhir; Richard III.}^{t5} "
Note: Information in the English version was represented in two sentences, while the Indonesian version represents the same information in four sentences.	

Table 4.6 Example of some additional information (paragraph level)

Lang	**Paragraph**
EN	"Roman military withdrawals left Britain open to invasion by pagan, seafaring warriors from north-western continental Europe, chiefly the Angles, Saxons and Jutes who had long raided the coasts of the Roman province and began to settle, *initially in the eastern part of the country.* Their advance was contained for some decades after the Britons' victory at the Battle of Mount Badon, *but subsequently resumed, over running the fertile lowlands of Britain and reducing the area* under Brythonic control *to a series of separate enclaves in the more rugged country to the west* by the end of the 6th century. *Contemporary texts describing this period are extremely scarce, giving rise to its description as a Dark Age. The nature and progression of the Anglo-Saxon settlement of Britain is consequently subject to considerable disagreement.* Christianity had in general disappeared from the conquered territories, but was reintroduced by missionaries from Rome led by Augustine from 597 onwards and by Irish missionaries led by Aidan around the same time. *Disputes between the varying influences represented by these missions ended in victory for the Roman tradition.*"
ID (EN-translation)	"Roman military withdrawals left Britain open to invasion by pagan, seafaring warriors from north-western continental Europe, chiefly the Angles, Saxons and Jutes who had long raided the coasts of the Roman province and began to settle. Their advance was contained for some decades after the Britons' victory at the Battle of Mount Badon. Britain went under Brythonic control at the end of the 6th century. Christianity had in general disappeared from the conquered territories, but was reintroduced by missionaries from Rome led by Augustine from 597 onwards and by Irish missionaries led by Aidan around the same time. "
ID	"Penarikan tentara Romawi membuat Inggris terbuka atas serangan dari suku-suku pagan dan tentara pelaut yang berasal dari barat daya Eropa, terutama suku Angles, Saxon, dan Jute, yang sudah lama menduduki pesisir timur Britania dan selanjutnya mulai membangun pemukiman. Pengaruh mereka tetap bertahan selama berdekade-dekade lamanya hingga suku Briton berhasil memenangkan Pertempuran Gunung Badon. Setelah itu, Britania kembali jatuh ke tangan Briton pada akhir abad ke-6. Agama Kristen turut menghilang seiring jatuhnya Romawi, namun diperkenalkan kembali oleh para misionaris dari Romawi yang dipimpin oleh Agustinus sejak tahun 597 dan seterusnya, serta oleh misionaris Irlandia bernama Aidan pada periode yang sama."
Note: the additional information is shown in italic.	

Table 4.7 Example of a one-to-many paragraph alignments

Lang	Paragraph
EN	"{Many ancient standing stone monuments were erected during the prehistoric period, amongst the best known are Stonehenge, Devil's Arrows, Rudston Monolith and Castlerigg. With the introduction of Ancient Roman architecture there was a development of basilicas, baths, amphitheaters, triumphal arches, villas, Roman temples, Roman roads, Roman forts, stockades and aqueducts.}^{t1} {It was the Romans who founded the first cities and towns such as London, Bath, York, Chester and St Albans. Perhaps the best known example is Hadrian's Wall stretching right across northern England. Another well preserved example is the Roman Baths at Bath, Somerset.}^{t2}"
ID (EN-translation)	"{Many ancient standing stone monuments were erected during the prehistoric period, amongst the best known are Stonehenge, Devil's Arrows, Rudston Monolith and Castlerigg. With the introduction of Ancient Roman architecture there was a development of basilicas, baths, amphitheaters, triumphal arches, villas, Roman temples, Roman roads, Roman forts, stockades and aqueducts.}^{t1} {It was the Romans who founded the first cities and towns such as London, Bath, York, Chester and St Albans. Perhaps the best known example is Hadrian's Wall stretching right across northern England. Another well preserved example is the Roman Baths at Bath, Somerset.}^{t2}"
ID	"{Banyak monumen-monumen kuno yang dibangun pada masa prasejarah, yang paling terkenal adalah Stonehenge, Devil's Arrows, Rudston Monolith dan Castlerigg. Dengan diperkenalkannya arsitektur Romawi Kuno, bangunan-bangunan seperti basilika, pemandian, amfiteater, villa, kuil Romawi, benteng, dan saluran air model Romawi juga makin berkembang.}^{t1} {Romawi mendirikan kota-kota pertama seperti London, Bath, York, Chester dan St Albans. Contoh arsitektur terpentingnya adalah Tembok Hadrian, yang membentang di bagian utara Inggris. Peninggalan lainnya yang cukup terpelihara dengan baik adalah pemandian Romawi di Bath, Somerset.}^{t2}"
Note: the same contents correspond in sentence-by-sentence translation but have different representations in the paragraph level. I.e., the Indonesian article represents the content in two paragraphs, while the English article only has one.	

Table 4.8 Example of some additional information (sentence level)

Lang	Paragraph
EN	"Roman military withdrawals left Britain open to invasion by pagan, seafaring warriors from north-western continental Europe, chiefly the Angles, Saxons and Jutes who had long raided the coasts of the Roman province and began to settle, *initially in the eastern part of the country. ...* "
ID (EN-translation)	"Roman military withdrawals left Britain open to invasion by pagan, seafaring warriors from north-western continental Europe, chiefly the Angles, Saxons and Jutes who had long raided the coasts of the Roman province and began to settle. ..."
ID	"Penarikan tentara Romawi membuat Inggris terbuka atas serangan dari suku-suku pagan dan tentara pelaut yang berasal dari barat daya Eropa, terutama suku Angles, Saxon, dan Jute, yang sudah lama menduduki pesisir timur Britania dan selanjutnya mulai membangun pemukiman. ..."
Note: the additional information is shown in italic.	

- A sentence in D_1 contains the same content in D_2.
- A sentence in D_1 contains the same content in D_2 but the Wiki links are different.
- A sentence in D_1 may be represented in more than one sentences in D_2 (as previously shown in Table 4.5).
- A sentence in D_1 may contain additional information compared to its aligned sentence in D_2 (shown in Table 4.8).
- A sentence in D_1 may contain different information compared to its aligned sentence in D_2.

Most of the examples above showed dissimilarities between content which might have originated from sentence-by-sentence translations. On other cases, however, some contents were found to describe the same topic although they did not correspond in a translation manner. In an example shown in Table 4.9 from article pair 1, both contents describe the adoption of Islam religion in Indonesia. The EN article described the topic in more detail compared to the ID version. On the other hand, although the ID version is much shorter, it contains new information not available in the EN version (i.e., that the Muslim traders arrived through Gujarat, India).

Table 4.9 Example of content describing different aspects

Lang	Text
EN	"Although Muslim traders first traveled through Southeast Asia early in the Islamic era, the earliest evidence of Islamized populations in Indonesia dates to the 13th century in northern Sumatra. Other Indonesian areas gradually adopted Islam, and it was the dominant religion in Java and Sumatra by the end of the 16th century. For the most part, Islam overlaid and mixed with existing cultural and religious influences, which shaped the predominant form of Islam in Indonesia, particularly in Java."
ID (EN-translation)	"The arrival of Arabic and Persian traders through Gujarat, India, then brought Islam."
ID	"Kedatangan pedagang-pedagang Arab dan Persia melalui Gujarat, India, kemudian membawa agama Islam."

Table 4.10 Example of contradictory sentences

Lang	Text
EN	"Fossils ... show that the Indonesian archipelago was inhabited by Homo erectus, popularly known as 'Java Man', between *1.5 million years ago and as recently as 35,000 years ago*."
ID (EN-translation)	"Fossilized remains of Homo erectus, which by anthropologists also dubbed 'Java Man', raises suspicion that the Indonesian archipelago was inhabited *two million to 500,000 years ago*."
ID	"Peninggalan fosil-fosil Homo erectus, yang oleh antropolog juga dijuluki 'Manusia Jawa', menimbulkan dugaan bahwa kepulauan Indonesia telah mulai berpenghuni pada antara *dua juta sampai 500.000 tahun yang lalu*."
Note: the contradictory information is shown in italic.	

Furthermore, whilst both article pairs 1 and 2 describe similar topics and contain many translated sentences, some *contradictions* have also been found in the document contents. An example of these is shown in Table 4.10 (the contradictory information is shown in italics). Different content, although not necessarily contradictory, were also found when articles used different references or reported data that were collected at different times. For example, Table 4.11 shows that the EN sentence reports statistics gathered in 2012, whilst the data in the ID sentence were gathered in 2006. Contradictory or different information were also found as results of outdated information (i.e., information not updated in one language version following a change), or a mistake in one language version.

Although dissimilarities do occur in the contents of article pair 1 and 2, majority of the

Table 4.11 Example of different references

Lang	**Text**
EN	"However, *as of 2012, an estimated 11.7% of the population lived below the poverty line* and the official open *unemployment rate was 6.1%.*"
ID (EN-translation)	"However, the impact of that growth has not been large enough to affect *the unemployment rate, which amounted to 9.75%.* Estimates *in 2006, as many as 17.8% of the people live below the poverty line.*"
ID	"Namun demikian, dampak pertumbuhan itu belum cukup besar dalam memengaruhi *tingkat pengangguran, yaitu sebesar 9,75%.* Perkiraan *tahun 2006, sebanyak 17,8% masyarakat hidup di bawah garis kemiskinan.*"
Note: the content reporting different references is shown in italic.	

texts in these article pairs can be aligned to each other. On the other hand, when a similar analysis was performed on article pair 3, very little text was able to be aligned in this article pair. As shown by the structure similarity, both articles describe different aspects. For example, the EN article elaborates on Chopin's background and family whilst the ID article discusses these aspects very briefly. Furthermore, the ID article describes different compositions; these were not available at all in the EN version of the article. Overlapping information appearing in both texts instead occurred in content with higher granularities, such as dates, locations, or names, rather than translated sentences or paragraphs.

4.4.4 Relations between quality of articles, word length, and the similarity degree

The previous section discovered that article pair 1 and 2 are much similar than article pair 3. Both these document pairs had similar word lengths. On the opposite, article pair 3, which were shown to have very different lengths, also contain very little similarity. These findings indicate that article length may be a promising feature for identifying similarity.

Article quality, however, was not found to be a good feature in identifying similarity for a number of reasons. Firstly, this information is not available for all documents. For example, article pair 2, which is a pair of a good article and an unassessed article, was shown to be similar to each other, despite the varying quality between the articles. Article pair 3, on the other hand, which is a featured article and an unassessed article, contain very different contents to one another. Secondly, the 'Featured Articles' quality does not guarantee the completeness of the information available in both languages. For example, article pair 1 (both featured articles), although were shown to be very similar to each other, have been shown in this study to miss some information that is available in the other language version.

4.5 Conclusion

The analysis of the three different Wikipedia article pairs in this initial study enabled a number of features for measuring similarity in Wikipedia to be identified. These findings allowed us to further understand the characteristics of similar interlanguage-linked articles in Wikipedia, which informed further work in Chapter 5 and contributed to the answer of research question 1:

RQ1. What are the characteristics of similar interlanguage-linked articles in Wikipedia? Based on an analysis of the three pairs of articles, two characteristics, i.e., the *structure similarity* between articles and *similarity of word length*, were shown to be useful in identifying articles with similar content. Articles with similar (or alignable) section headings are more likely to contain content of similar aspects compared to those with different structures. The *article quality* information, however, was not shown to be a good feature for indicating similarity at the document level; furthermore, this feature is not available for all Wikipedia articles. The findings further show that highly similar articles very often contain *translated content*, although some may have dissimilarities in the way the sentences are split, or additional information that appear in one language version. However, cases were also found that show that these articles (although are highly similar) may still contain differences and contradictions at the sub-document level (e.g., paragraphs or sentences).

Further investigation using a larger set of document pairs is needed to summarise the similarity of Wikipedia content at the sub-document level. This thesis, however, aims to focus on measuring document similarity and therefore, will not further analyse similarity at the sub-document level in the remainder of this thesis.

Whilst the method utilised in this initial study was able to gain many insights on similarity in Wikipedia, analysing the content of these articles was a very time-consuming task, especially for long document pairs. It was infeasible to carry out the same evaluation task on a much larger set of document pairs. Therefore, a modification of this approach is required before gathering human annotations on a larger set of documents (further described in Chapter 5).

Chapter 5

Evaluation Corpus

As previously described in Section 3.5, an evaluation corpus is needed to directly evaluate the approaches to measure similarity in Wikipedia. More importantly, this corpus is also needed in order to further identify and understand similarity characteristics in Wikipedia articles in a larger set of document pairs and language pairs. The findings derived from the initial study on Wikipedia similarity (Chapter 4) has provided some insights on features that contributed to similarity in Wikipedia. However, it is infeasible to carry out this task on a large number of document pairs. Therefore, the author simplified this task to focus on the similarity of the documents at the document level. This chapter describes the work in creating the Wikipedia evaluation corpus, which as far as the author is aware, is the first corpus available for measuring similarity in Wikipedia articles.

5.1 Background

To extrinsically evaluate similarity measures, an evaluation corpus is needed to compare the automatic scores (produced by the similarity approaches) against human judgments. Most available evaluation corpus exists at the word or sentence level, such as the WordSimilarity-353 Test Collection (Finkelstein et al., 2002) and the Microsoft Paraphrase Corpus (Dolan & Brockett, 2005). However, very few evaluation corpora are available to measure similarity at the document level. A large number of studies that investigated similarity at the document level (Benedetti et al., 2018; Gabrilovich & Markovitch, 2007; L. Huang et al., 2012; Yeh, Ramage, Manning, Agirre, & Soroa, 2009) have evaluated their approaches using an evaluation corpus of 50 short news articles (between 51-126 words each) from the Australian Broadcasting Corporation (M. D. Lee et al., 2005). In this corpus, every combination of articles (1,225 pairs in total) were annotated by 83 students using a 5-point Likert Scale to represent the relatedness of the topics (1=highly unrelated; 5=highly related).

At the time of writing, there was no evaluation corpus that has been specifically created for evaluating cross-lingual similarity approaches, especially for Wikipedia articles. One of the main contributions of this thesis is to develop an evaluation corpus that can be used to: i) identify similarity characteristics specifically for Wikipedia articles, and ii) evaluate the language-independent approaches for measuring similarity in Wikipedia. The work in creating such evaluation corpus is described in this chapter.

This work aims to answer two research questions:

RQ1. What are the characteristics of similar interlanguage-linked articles in Wikipedia?

RQ2. Can we create an evaluation benchmark for Wikipedia? I.e., do human assessors agree on Wikipedia similarity?

The work in this area is reported in four sections. First, the creation of evaluation tasks used in the process of gathering human judgments is described in Section 5.2. Section 5.3 reports the process of selecting documents for the evaluation corpus, whilst Section 5.4 describes the assessors participating in the evaluation task. The annotation results of the evaluation corpus are reported in Section 5.5. Finally, Section 5.6 discusses the results and findings.

5.2 Creating evaluation tasks

To develop a suitable evaluation task, a pilot task (Section 5.2.1) was run to test an initial similarity assessment scheme. Findings from this task were then used to improve the evaluation scheme used in the main evaluation task (Section 5.2.2).

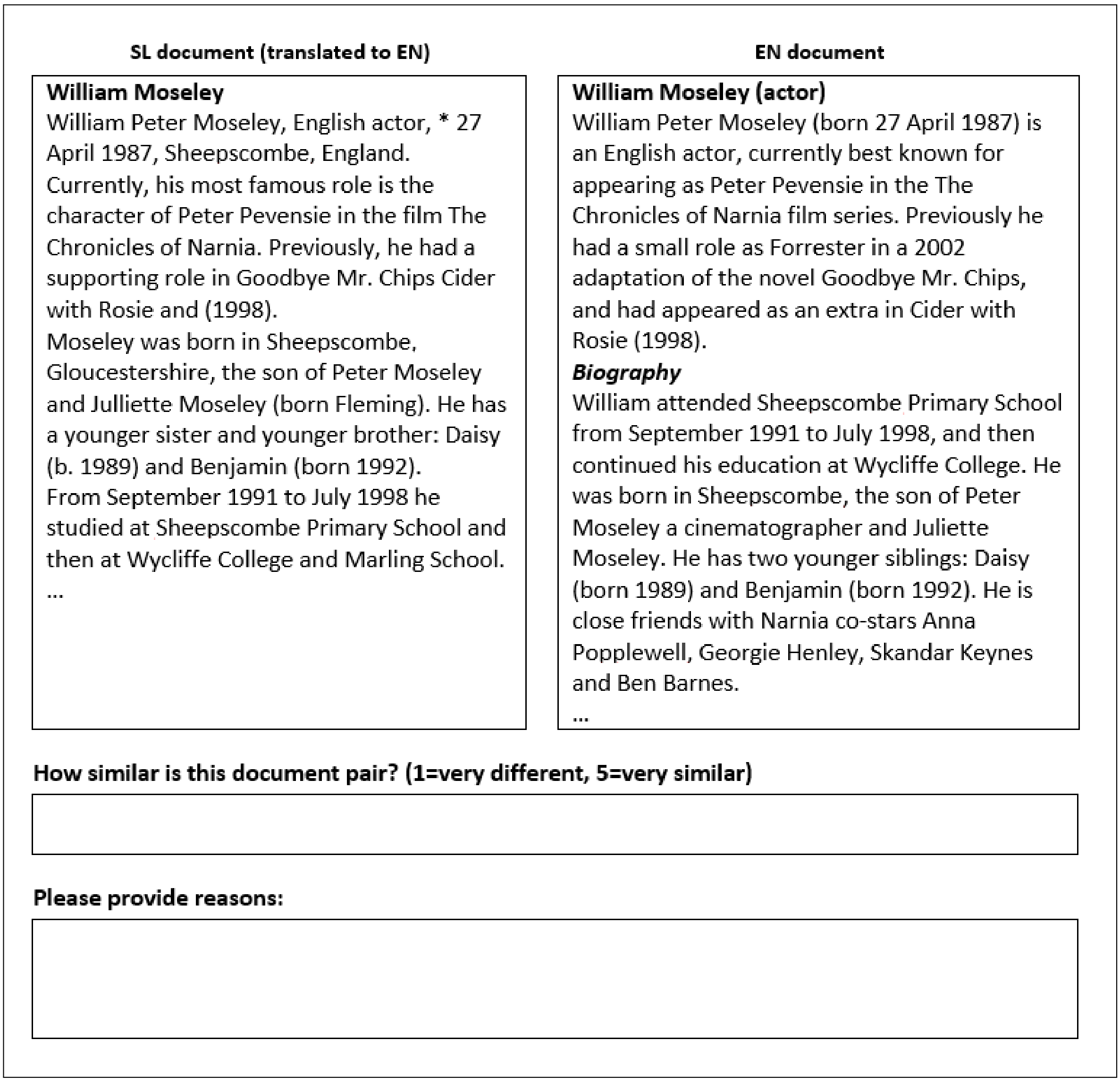
5.2.1 Pilot evaluation task

As previously explored in Section 2.2, there is no universally accepted definition of similarity; different degrees of similarity have also been described differently within the literature. Moreover, similarity characteristics that have been defined in general may also differ to those of Wikipedia articles. For the purpose of simplifying the task and understanding similarity characteristics that are specific to Wikipedia articles, the evaluation scheme used a 5-point Likert scale instead to specify the document similarity (1='very different' and 5='very similar'). Instead of providing definitions for each scale, assessors were asked to provide reasons they felt contributed to their judgments of similarity. The pilot evaluation task interface for one of the documents is shown in Figure 5.1.

Seven assessors participated in this task; all of which were either native English speakers or fluent speakers in English. For this experiment, the assessors were asked to judge the similarity of five document pairs. Document pairs used in the pilot evaluation task were manually selected from the Slovenian-English interlanguage-linked Wikipedia articles. Due to the language limitations of the assessors, the Slovenian documents were previously translated into English using Google Translate.[1] Assessors were asked to read the contents of each document pair, i.e., the English document and the translated Slovenian document, prior to identifying the similarity degree of the document pair and their reasons.

The author calculated the inter-rater reliability score using Krippendorff's α (Krippendorff, 1980), due to its applicability for ordinal data and multiple annotators (Artstein & Poesio, 2008). The pilot evaluation task suggested that assessors showed a moderate agreement in the 5-point Likert Scale (Krippendorrf's α=0.597). The distribution of scores between the seven assessors for each document pair is shown in Figure 5.2. In three out of the five document pairs, all assessors provided either the same scores or scores differing by one. In one case, assessors' scores differed by two (document pair 5), and one case occured where assessors' scores differed by up to three (document pair

[1] Whilst the translation quality from Slovenian to English was not perfect, it was considered to be sufficient for the assessors to assess the content similarity between the article pairs.



SL document (translated to EN)

William Moseley
William Peter Moseley, English actor, * 27 April 1987, Sheepscombe, England. Currently, his most famous role is the character of Peter Pevensie in the film The Chronicles of Narnia. Previously, he had a supporting role in Goodbye Mr. Chips Cider with Rosie and (1998).
Moseley was born in Sheepscombe, Gloucestershire, the son of Peter Moseley and Julliette Moseley (born Fleming). He has a younger sister and younger brother: Daisy (b. 1989) and Benjamin (born 1992).
From September 1991 to July 1998 he studied at Sheepscombe Primary School and then at Wycliffe College and Marling School.
...

EN document

William Moseley (actor)
William Peter Moseley (born 27 April 1987) is an English actor, currently best known for appearing as Peter Pevensie in the The Chronicles of Narnia film series. Previously he had a small role as Forrester in a 2002 adaptation of the novel Goodbye Mr. Chips, and had appeared as an extra in Cider with Rosie (1998).
Biography
William attended Sheepscombe Primary School from September 1991 to July 1998, and then continued his education at Wycliffe College. He was born in Sheepscombe, the son of Peter Moseley a cinematographer and Juliette Moseley. He has two younger siblings: Daisy (born 1989) and Benjamin (born 1992). He is close friends with Narnia co-stars Anna Popplewell, Georgie Henley, Skandar Keynes and Ben Barnes.
...

How similar is this document pair? (1=very different, 5=very similar)

Please provide reasons:

Fig. 5.1 Pilot evaluation task

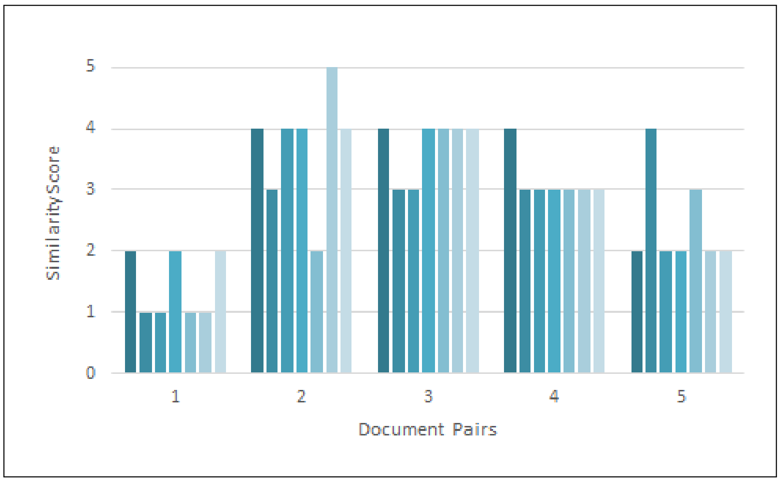
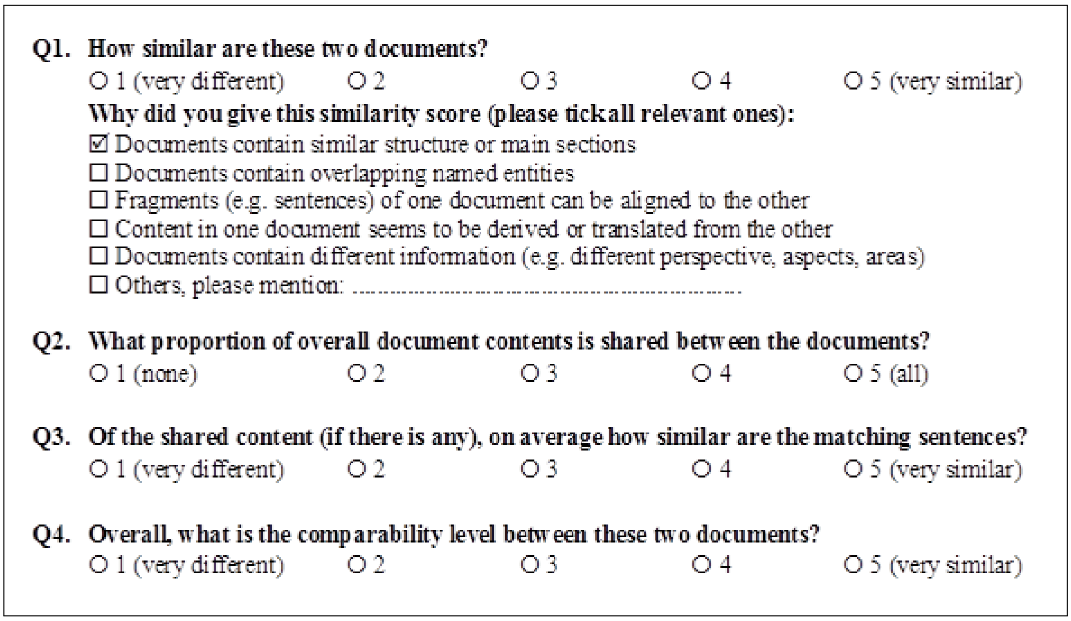


Fig. 5.2 Pilot evaluation task results (seven assessors for each document pair)

2). The latter document pair contains an article about Carlo Matteucci, an Italian physicist. Assessors who provided high scores identified that the article contents discussed the same topic (e.g., his life, works, awards, involvement in politics); on the other hand, another assessor identified that although the topics were the same, the contents of the documents had clearly been written by different authors (the SL content was not a translation of the EN content, and vice versa). This document pair was therefore punished by some assessors and given lower scores.

In the pilot task, assessors were also asked to provide reasons to justify their chosen similarity scores. Five reasons were identified to why a pair of documents were annotated to be similar (or dissimilar):

- Documents contain similar structure or main sections
- Documents contain overlapping named entities
- Fragments (e.g., sentences) of one document can be aligned to the other
- Content in one document seems to be derived or translated from the other
- Documents contain different information (e.g., different perspectives, aspects, areas)



Q1. How similar are these two documents?

○ 1 (very different) ○ 2 ○ 3 ○ 4 ○ 5 (very similar)

Why did you give this similarity score (please tick all relevant ones):

☑ Documents contain similar structure or main sections
☐ Documents contain overlapping named entities
☐ Fragments (e.g. sentences) of one document can be aligned to the other
☐ Content in one document seems to be derived or translated from the other
☐ Documents contain different information (e.g. different perspective, aspects, areas)
☐ Others, please mention: ....

Q2. What proportion of overall document contents is shared between the documents?

○ 1 (none) ○ 2 ○ 3 ○ 4 ○ 5 (all)

Q3. Of the shared content (if there is any), on average how similar are the matching sentences?

○ 1 (very different) ○ 2 ○ 3 ○ 4 ○ 5 (very similar)

Q4. Overall, what is the comparability level between these two documents?

○ 1 (very different) ○ 2 ○ 3 ○ 4 ○ 5 (very similar)

Fig. 5.3 Main evaluation task (evaluation questions)

5.2.2 Main evaluation task

Using the findings from the pilot evaluation task, the evaluation scheme for the main evaluation task were adapted. The resulting evaluation task, shown in Figure 5.3, contains four main questions. The first question (Q1), similar to the pilot task, contains two parts; in the first part, assessors were asked to judge the similarity based on the contents of the documents, and in the second part, assessors were asked to justify the chosen similarity scores by selecting all the relevant reasons that contributed to their answers. These reasons were previously derived from the pilot study. A new option was also included, ‘Others’, for assessors to add new reasons not previously included in the list.

Assessors then were asked to provide a score identifying the proportion of shared contents between the documents (Q2), and the similarity of sentences within these contents, if they exist (Q3). Lastly (Q4), assessors were asked to assign a score to reflect the degree of comparability between the two documents. (Assessors for this task were familiar with the notion of comparability, as further described in Section 5.4.) Lastly, to avoid any bias caused by translation quality, in the main evaluation task, all documents are shown in their original languages, i.e., non-English documents were not translated into English.

5.3 Selecting evaluation documents

In order to create a representative benchmark, a number of document pairs with varying degrees of similarity were needed for the human assessment. These documents were selected for the eight language pairs used in this study (see Section 3.3.1): German (DE), Greek (EL), Estonian (ET), Croatian (HR), Lithuanian (LT), Latvian (LV), Romanian (RO) and Slovenian (SL) – all paired to English (EN). Except for German and English, all these languages are under-resourced.

To select these document pairs, first, the anchor text and word overlap method was used to measure similarity within all interlanguage-linked articles in the Wikipedia corpus. The detail of this approach is described in Chapter 6 (specifically Section 6.2), and the corpus used in this study is described in Section 3.6.2. Firstly, the titles of interlanguage-linked articles were extracted and used as a bilingual lexicon, to identify overlapping links across languages. This method calculates similarity by finding similar sentences in the document pair; similar sentences are identified as those that contain a high proportion of overlapping links and words. The sentence alignment information is aggregated to represent the similarity at the document level, i.e., document pairs with many similar sentences are assigned high similarity scores.

For each language pair, the document pairs were then sorted based on the similarity scores assigned by the method. The score range, i.e., the difference between the minimum and maximum score, was then divided into 10 bins. From each bin, 10 document pairs were randomly selected,[2] resulting in a total of 100 document pairs per language pair. Whenever possible, a maximum word length of 1,000 tokens was set when selecting articles in order to ensure assessors were able to read and assess the articles in a reasonable time and to limit assessor fatigue. However, this was not always feasible for bins with limited number of articles. In total, 97% of articles included in the dataset contained fewer than 1,000 tokens.

[2] When this was not possible (i.e., fewer than 10 document pairs were found in a bin) the maximum number of document pairs in that bin were chosen for the evaluation set and a higher number of documents were chosen from the lower bins to achieve the total number of 100 document pairs.

Table 5.1 Summary of documents used for human similarity judgments

Number of languages	9 (DE, EL, EN, ET, HR, LT, LV, RO and SL)
Number of language pairs	8 (DE-EN, EL-EN, ET-EN, HR-EN, LT-EN, LV-EN, RO-EN & SL-EN)
Number of documents	1,589* (800 document pairs)
Number of documents per language pair	200 (100 document pairs)
Average number of words per document	450.59 (min: 107, max: 1,546)
Average number of sentences per document	51.31 (min: 22, max: 1,028)
*Note: * The same 11 English documents were selected for two different language pairs.*	

The document selection process above was performed for each language pair independently, i.e., the chosen document pairs in one language pair did not affect the selection for a different language pair.[3] Whilst the documents selection was performed independently, a small number of documents did overlap between several language pairs. Overall, 11 English documents were shared between two different language pairs.[4] This results in a total of 789 unique English documents.

A summary of the documents used for the evaluation dataset is shown in Table 5.1. The average document length for each language in the eight language pairs is shown in Figure 5.4. Overall, the selected non-English documents were almost 30% shorter (an average of 140 words fewer) than the English documents; the overall average length for non-English documents was 362 words, whilst the average length for English documents was 502 words. However, not all non-English documents were shorter than their paired English documents. Out of 800 non-English documents, 621 documents (77.6%) were shorter than their paired English documents, whilst the remaining 179 documents were longer. These numbers varied between language pairs as shown in Table 5.2; RO dataset had the fewest number of documents that were shorter than their EN documents (67%), meanwhile, up to 91% of LT documents were shorter than their EN documents.

[3] This was decided as the number of documents that appear in all 9 languages was significantly lower.

[4] Four English documents overlapped in both the RO and SL sets, and one English document overlapped in each of these language sets: EL and ET, LT and LV, LT and RO, HR and RO, ET and RO, EL and RO, DE and SL.

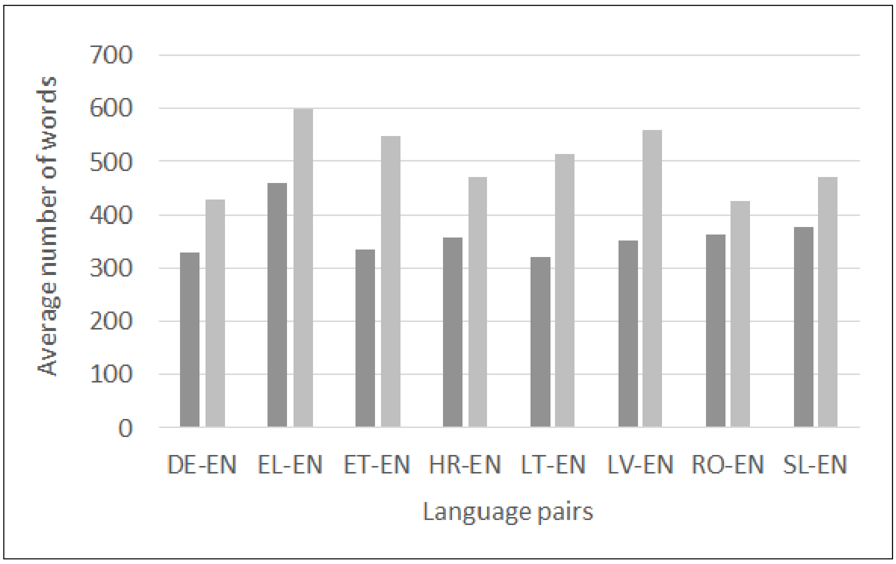


Fig. 5.4 Average number of words per document for each language pair

Table 5.2 Proportion of shorter documents in each language pair

Lang Pair	**Number of source documents**		**Total**
	$L_{source} < L_{EN}$	$L_{source} \geq L_{EN}$	
DE-EN	76	24	100
EL-EN	69	31	100
ET-EN	84	16	100
HR-EN	76	24	100
LT-EN	91	9	100
LV-EN	90	10	100
RO-EN	67	33	100
SL-EN	68	32	100
All pairs	621	179	800
Note: L_{source} represents the length of source (non-English) documents and L_{EN} represents length of English documents			

5.4 Assessors

Sixteen assessors participated in the main evaluation task (two assessors for each language pair). All assessors were native speakers of the non-English language and fluent speakers of English. All assessors were partners in the ACCURAT project,[5] an EU-funded project aiming to exploit comparable corpora from the Web for the purpose of improving machine translation; these corpora were shown to be valuable resources for languages and domains that are under-resourced (Skadiņa et al., 2012). The tasks in the project included identifying comparable documents in the Web, extracting parallel fragments within them, and using them to improve machine translation performance for under-resourced languages. The project partners, therefore, had a great knowledge and understanding of cross-lingual similarity and comparability.

5.5 Annotation results

Given a pair of Wikipedia articles in different languages, assessors were asked to read the articles and answer the four questions shown in Figure 5.3. Each of the 16 assessors assessed 100 document pairs, resulting on 1,600 annotations. The results for each of the four evaluation questions were analysed and are reported in this section.

Inter-annotator agreements between the assessors were measured using four different measures:

1. *Spearman's ρ.* Spearman rank-order correlation measures the correlation between ranks of documents based on the scores given by assessors in each of the four questions.
2. *Cohen's Kappa.* A weighted version of Cohen's Kappa (Cohen, 1968) was used to measure inter-annotator agreement over results on the 5-point scale. In this study, the weighted Cohen's Kappa is computed using a squared weight of the score difference (`scoreDiff`) between the two annotators, punishing annotations with bigger

[5] ACCURAT (Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation) project, www.accurat-project.eu.

score differences. I.e., cases scored 1 and 2 are considered to be a better agreement than cases scored 1 and 5.

3. *Krippendorff's Alpha* (Krippendorff, 1980). This inter-annotator agreement was calculated taking into account the 5-point Likert Scale as ordinal data.
4. *Percent agreement.* This measure reports the proportion of cases in which annotators provided the same scores.

5.5.1 Results on Q1: "How similar are the two documents?"

When asked to provide a score between 1-5 to identify the similarity between each document pair (1="very different" and 5="very similar"), the assessors achieved a moderate agreement with each other across the eight language pairs (Spearman's $\rho = 0.59$; weighted Cohen's Kappa=0.50, Krippendorrf's α=0.42). The results are shown in Table 5.3. Only an average of 40.5% cases were given the same similarity scores. This low agreement percentage may be caused by assessors' different understanding behind their chosen similarity scores, as there were no defined guidelines specifying the definition of the different similarity score. Further analysis, however, shows that up to 84.13% cases were scored either the same or differed by 1; this proportion further increased to 97.63% of the cases when considering scores that differed by two or less. When the score difference tolerance was increased to 1, i.e., including cases where the scores differed by 1 or less as agreements, the average weighted Cohen's Kappa increased to 0.70, showing a higher agreement between annotators in answering this question. The results for each language pair is shown in Table 5.4.

Figure 5.5 shows the distribution of the similarity scores assigned to document pairs where these scores were averaged across the two assessors. Overall, less than 1% (5 document pairs) were assessed to be different from each other (i.e., average score of 1). Just under 10% document pairs were scored an average of 2 or below, and 10% were scored an average of 2.5. The proportions of document pairs with average score between 3 and 4 were similar, around 15% each. Finally, around 35% of the document pairs were assessed to be very similar by both assessors (i.e., average score of 4.5 and 5).

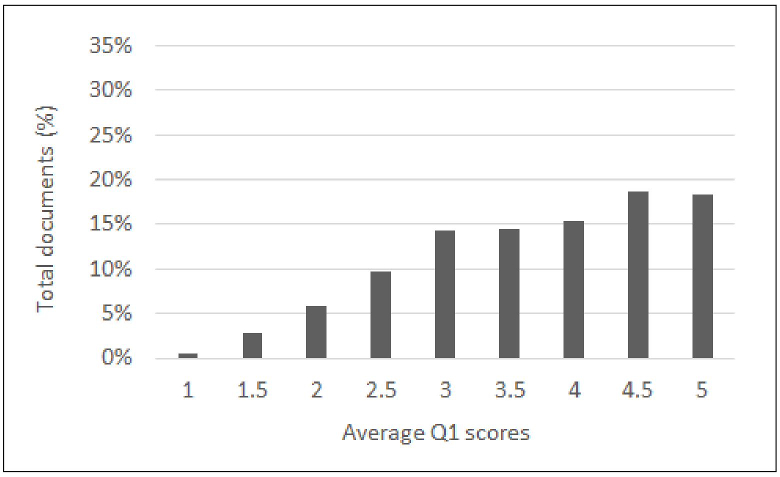


Fig. 5.5 Distribution of document-level similarity scores (Q1) averaged across both assessors (N=800)

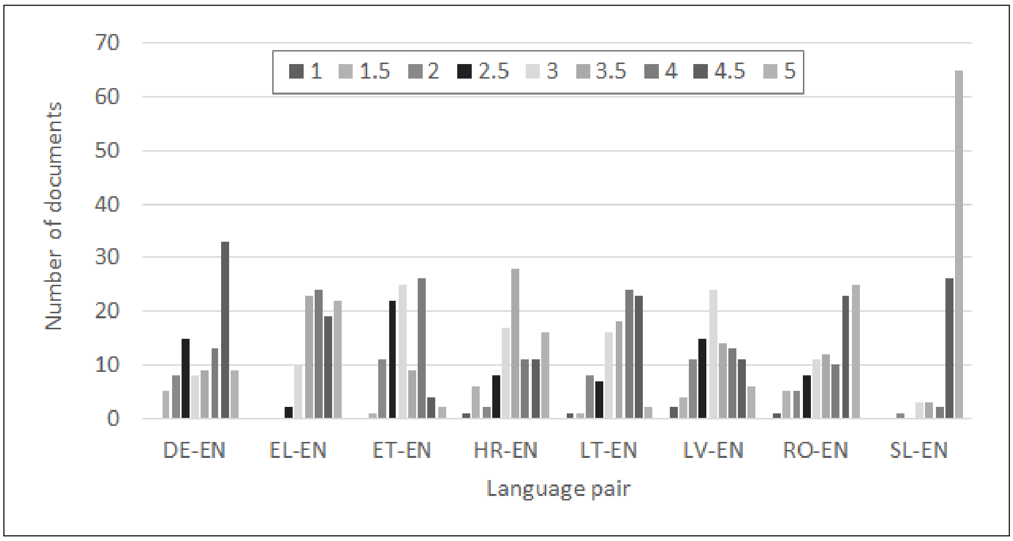


Fig. 5.6 Distribution of document-level similarity scores (Q1) per language pair (N=100), averaged across both assessors

Table 5.3 Inter-assessor agreement for Q1 (5 classes)

Lang pair	Spearman's ρ	Weighted Cohen's Kappa*	Krippendorff's α	Agreement
DE-EN	0.74	0.55	0.48	25%
EL-EN	0.46	0.32	0.31	43%
ET-EN	0.67	0.60	0.62	57%
HR-EN	0.76	0.54	0.45	28%
LT-EN	0.47	0.30	0.08	19%
LV-EN	0.60	0.55	0.57	45%
RO-EN	0.56	0.56	0.52	37%
SL-EN	0.43	0.56	0.33	70%
Mean	0.59	0.50	0.42	40.5%
* $weight = scoreDiff^2$				

Table 5.4 Inter-assessor agreement for Q1 (5 classes, score difference tolerance of 1)

Lang pair	Weighted Cohen's Kappa*	Agreement
DE-EN	0.73	84%
EL-EN	0.43	85%
ET-EN	0.79	93%
HR-EN	0.77	80%
LT-EN	0.48	63%
LV-EN	0.68	88%
RO-EN	0.73	81%
SL-EN	1.00	99%
Mean	0.70	84.13%
* If the scoreDiff is 1 or less, $weight = 0$; otherwise, $weight = (scoreDiff - 1)^2$		

The same results were analysed for each language pair, as shown in Figure 5.6. The results show that the similarity levels of document pairs vary widely between the different language pairs. Over 90% of the documents in SL-EN have an average score of 4 or above; whilst only 6% of ET-EN have the same score range.

To further analyse the results, the average scores were aggregated into two bins: 'similar' (average score equal 3.5 or above) and 'non-similar' (average score of 3 or lower). Using these categories, an overall of 67% of document pairs (536 document pairs) were assessed to be similar, whilst 33% (264 document pairs) were judged to be non-similar, as shown in Table 5.5. The proportion of similar and non-similar documents were roughly the same between most language pairs, except for ET-EN and LV-EN, in which the num-

Table 5.5 Similarity assessment for all language pairs

Language pair	Similar Document Pairs	Non-Similar Document Pairs	Total
DE-EN	64	36	100
EL-EN	88	12	100
ET-EN	41	59	100
HR-EN	66	34	100
LT-EN	67	33	100
LV-EN	44	56	100
RO-EN	70	30	100
SL-EN	96	4	100
All	536 (67%)	264 (33%)	800

bers of non-similar documents were higher than those of similar documents. Also, for two language pairs (EL-EN and SL-EN), the numbers of documents judged to be non-similar were found to be significantly lower, 12% and 4%, respectively.

5.5.2 Q1 reasons: "Why did you give this similarity score?"

When judging cross-language similarity, the assessors were asked to provide reasons that led them to make their decisions. A list of options were provided to help them identify the similarity characteristics between the document pairs; these included whether the structures of both articles were similar (‘*similar structure*’), whether documents contained overlapping named entities (‘*overlapping NEs*’), whether fragments of text from one document could be aligned to the other (‘*overlapping fragments*’), whether content in one article appeared with equivalent translations in the other (‘*contains translation*’), and whether different information were contained in the articles (‘*different information*’). Assessors were also allowed to provide other reasons not specified above.

Overall, assessors selected an average of 3.02 reasons (a minimum of 1 reason and a maximum of 5 reasons). Assessors in LV-EN selected an average of 2.3 reasons (lowest in the dataset), whilst the average for SL-EN is the highest (3.85 reasons). This may relate to the level of similarity in the evaluation documents, i.e., documents with higher similarity often have more reasons contributing to the similarity, compared to non-similar documents. As discussed in Section 5.5.1, SL-EN has the highest number of similar document pairs, whilst LV-EN has the second lowest number of similar document pairs.

To *understand the similarity characteristics that led assessors toward choosing different Q1 scores*, the reasons for each of the 1,600 annotations were grouped by their Q1 scores. The results, summarised in Figure 5.7, show that different characteristics were found when comparing annotations with different similarity scores.

The results show that when assessors annotated document pairs to be very different (i.e., Q1 score of 1), 24.14% were annotated to contain overlapping named entities and almost 85% were annotated to contain different information. Very few of these cases were annotated to contain similar structure, overlapping fragments and translated contents. In contrast, the majority of cases (85% or higher) with Q1 scores of 4 and 5 were annotated to contain similar structure, overlapping named entities and overlapping fragments. Furthermore, only a small percentage of these annotations were assessed to contain different information (5.9% and 0.2%, respectively).

The results further show that the proportion of cases annotated to contain similar structure, overlapping named entities, overlapping fragments and translated contents increased with the rising Q1 scores. On the other hand, the proportion of 'different information' substantially decreased when the Q1 scores increased.

The results also show that over 80% annotations that were scored 3 still contained overlapping named entities and overlapping fragments. This suggests that although the overall similarity of the document pair may be lower, it may still contain similarity at the sub-document level, e.g., same entities and similar sentences or phrases. However, this is not the case for the proportion of translated contents. A high proportion of translated contents (85.49%) was found in annotations of similar (i.e., Q1 scores of 5). However, this dropped to 44.69% for Q1 scores of 4, and 18.28% for Q1 scores of 3.

Furthermore, the similarity characteristics given by assessors were analysed with respect to the 800 document pairs in the evaluation dataset. This was carried out to *identify the overall characteristics of the similarity corpus*. Figure 5.8 reports the similarity characteristics chosen by, i) at least one assessor, and ii) both assessors overall the evaluation set. This figure summarises the characteristics of the evaluation dataset. The results

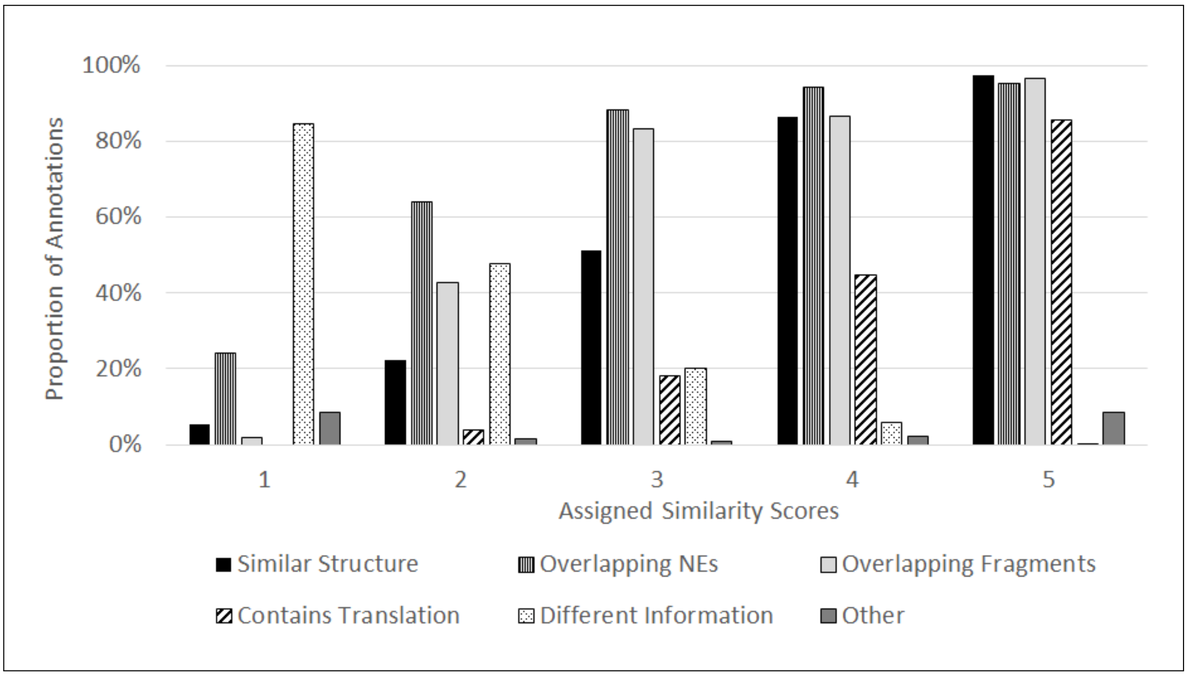


Fig. 5.7 Characteristics that capture various levels of similarity (N=1,600)

show that when considering similarity characteristics given by at least one assessor, majority of document pairs (98%) in the evaluation corpus were assessed to contain overlapping named entities, 93% to contain overlapping fragments, and 86% to share similar structure. The proportion of document pairs that were assessed to contain translation by at least one assessors is over 60%. When considering characteristics that both assessors agreed on, over three-quarters (76%) document pairs were assessed to contain named-entity overlap, 68% contain overlapping fragments, 55% had similar structure, and 28% contain translations. Only 5% documents were assessed to contain different information by both assessors; however, at least one assessor specified that document pairs contain different information in over around 27% of the evaluation dataset.

When document pairs contained characteristics not captured in the list of reasons in the evaluation scheme, assessors were able to select the option '*others*' and provide their own reasons. Overall, 7 assessors provided their own reasons in 64 annotations (4%); these annotations were distributed across 59 document pairs in four language pairs (DE-EN, HR-EN, RO-EN and SL-EN). The reasons given by assessors for these annotations were analysed and are reported below:

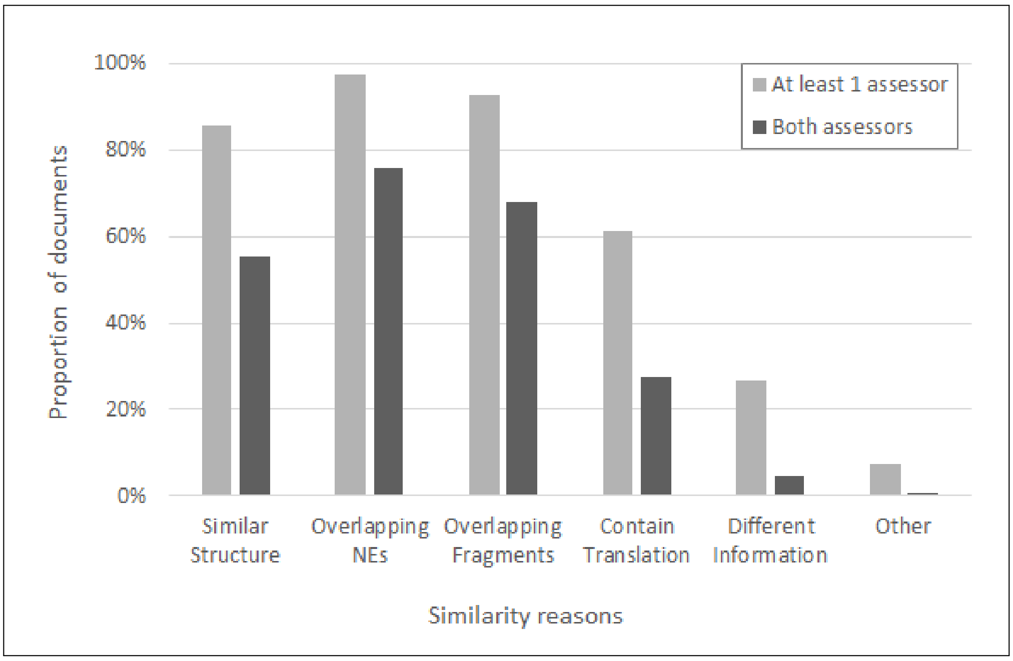


Fig. 5.8 Similarity characteristics of the evaluation dataset (N=800)

- *A large number of named entities* appearing in the document contents (35 annotations: 55%). Although assessors identified the overlap of named entitites to be one of the reasons contributing to a document similarity, document pair whose contents are largely named entities are often not deemed as useful as bilingual resources. An example of these document pairs is a RO-EN document pair listing the scientific names for cactus species, or a SL-EN document pair with the topic: "Members_of_the_European_Parliament_for_Sweden_1999–2004", in which both contents in the language pair list all the names of the parliament members. One assessor mentioned that "*not much relevant information can be extracted from alinged* [sic] *sentences .. they contain named entities ...*". The majority of these annotations appeared in the SL-EN corpus with 32 document pairs being assessed as containing "*a lot of (International) names*".

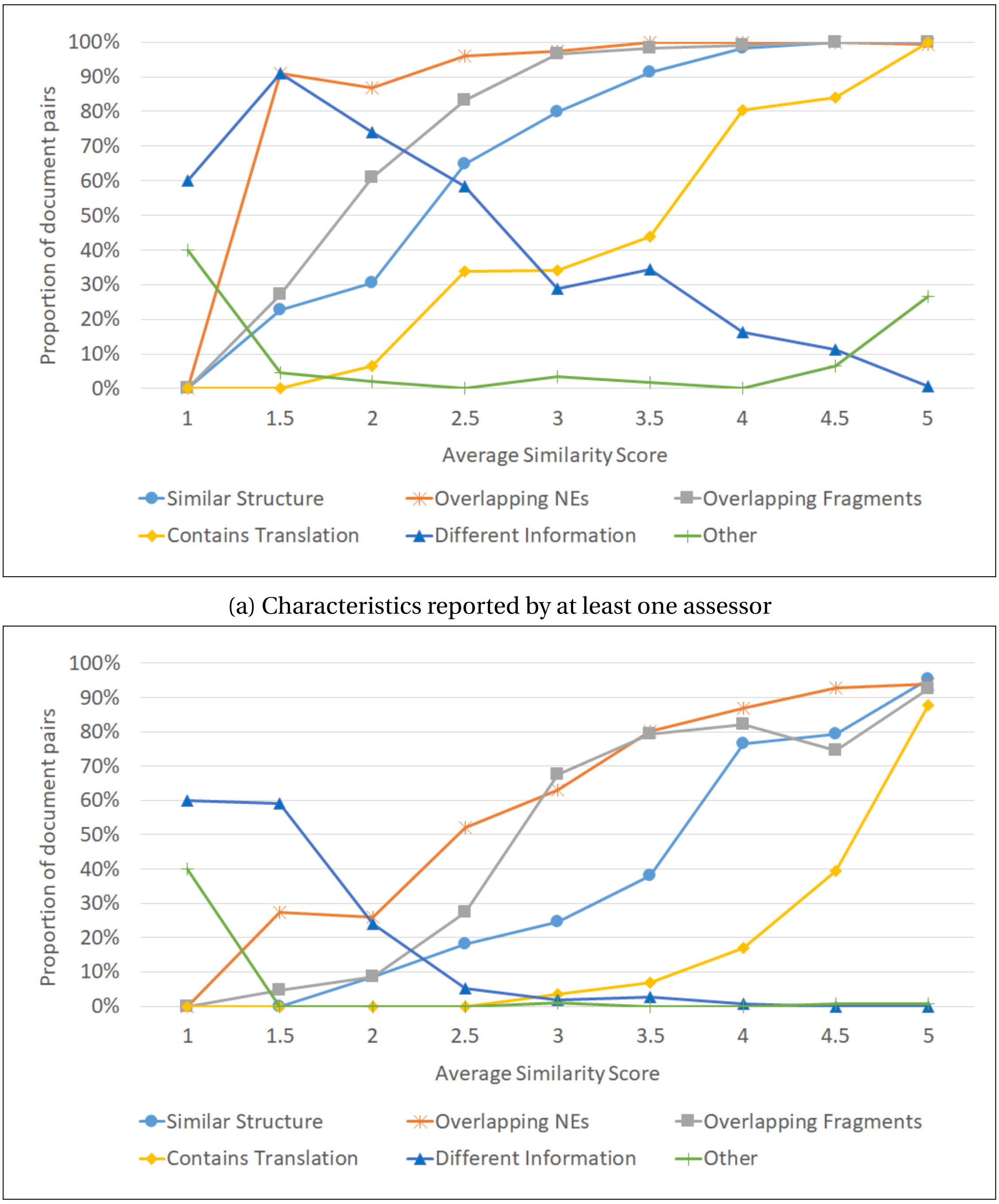
- *Language issues* (20 annotations: 31%). In several documents, assessors identified that some contents were not written in the correct language. In three-quarters of these annotations, English content was found in the non-English documents, and vice versa in the remaining annotations (i.e., non-English content were found in the English documents). In one case, one assessor mentioned that "*the romanian version of the text is translated from spanish and contains unmodified text*".

- *More information* about the similarity aspects (8 annotations: 12.5%). In these annotations, assessors simply used the 'others' option to provide further information about the document similarity, such as to indicate that more information appeared in the EN document, or to specify the specific paragraphs where similarity occurs. In one case, one assessor used this option to indicate that the document pair contained "*same topic, similar information (i.e. similar perspective, aspects ...), but described differently*".
- *Content error* (1 annotation: 1.6%). One assessor specified that some invalid characters and line breaks were found in the document content. This error, however, did not affect the similarity score of the document (this document was given a Q1 score of 5).

Finally, the author also combined the annotations between the two assessors to *analyse the characteristics of document pairs with different scores.* In this case, the Q1 scores provided by the two assesors were averaged. The author reported the characteristics of document pairs for each Q1 score in two figures. Figure 5.9a reports the characteristics annotated by at least one assesor and Figure 5.9b reports only the characteristics that were annotated by both assessors.

As expected, the characteristics for the 800 document pairs were similar to the characteristics for the 1,600 annotations (Figure 5.7). Analysing characteristics that were annotated by both assessors only, three document characteristics, i.e., similar structure, overlapping named entities and overlapping fragments, were found for most (over 74.67%) of the document pairs with Q1 scores of 4 and above. Both assessors also agreed that 87.76% document pairs with the highest similarity score (i.e., Q1 score of 5) contain translation.

Analysing document pairs with average similarity scores below 2, the author found very low number of document pairs with overlapping named entities, overlapping fragments, and translated contents. Instead, around 60% of these document pairs were an-



(a) Characteristics reported by at least one assessor

(b) Characteristics reported by both assessors

Fig. 5.9 Similarity characteristics of document pairs with different similarity scores (N=800)

notated to contain different information by both assessors.

These findings were very helpful to identify the characteristics of similar document pairs, and to develop suitable methods to automatically measure similarity in Wikipedia articles. Furthermore, these findings also provide more information on the usability of document pairs in Wikipedia. E.g., if one intends to extract translated contents, they should only consider Wikipedia articles with average similarity scores of 5. However, if one intends to extract overlapping fragments and named entities, they should also consider utilising document pairs with average similarity scores of 3 and above. Furthermore, document pairs with average score of 2 or below were shown to have very low proportion of overlapping content, which indicates their low value as a bilingual resource.

5.5.3 Results on Q2: "What proportion of overall document contents is shared between the documents?"

The second question in the evaluation task required assessors to identify the proportion of overall document contents that is shared between the two documents. The agreement between annotators in Q2 is shown in Table 5.6. When asked to identify the proportion of overall document contents that are shared between the documents, assessors showed a good agreement to each other (Spearman's ρ = 0.61). The highest Spearman's ρ was achieved in HR-EN (ρ = 0.74), whilst the lowest occured between LT-EN assessors (ρ = 0.51). Measuring the agreement in the 5-point scale, assessors agreed with each other in just under half of the cases (48%), weighted Cohen's Kappa = 0.47. However, the percentage agreement increases drastically when considering cases where ssessors either gave the same scores or scores differing by one (92% agreement in the 800 document pairs; the agreement scores for each language pair is shown in Table 5.7); this agreement further increased to 99% when considering cases where the scores differ by up to two.

The Q2 scores were averaged between the two assessors to identify the proportion of overall document contents for each document pair, as shown in Figure 5.10. The results show that less than 10% of the document pairs have very low content overlap (average score between 1 and 2). The proportion of documents with higher average Q2 scores

Table 5.6 Inter-assessor agreement for Q2 (5 classes, N=800 doc pairs)

Lang Pair	**Spearman's ρ**	**Weighted Cohen's Kappa**	**Krippendorff's α**	**Agreement**
DE-EN	0.71	0.65	0.63	46%
EL-EN	0.53	0.46	0.47	50%
ET-EN	0.66	0.59	0.61	58%
HR-EN	0.74	0.60	0.56	34%
LT-EN	0.51	0.43	0.43	43%
LV-EN	0.55	0.51	0.54	39%
RO-EN	0.55	0.50	0.49	38%
SL-EN	0.64	0.74	0.61	79%
Mean	0.61	0.56	0.54	48.38%

Table 5.7 Inter-assessor agreement for Q2 (5 classes, score difference tolerance of 1)

Lang Pair	**Weighted Cohen's Kappa***	**Agreement**
DE-EN	0.87	92%
EL-EN	0.66	90%
ET-EN	0.79	94%
HR-EN	0.90	95%
LT-EN	0.59	90%
LV-EN	0.67	92%
RO-EN	0.71	81%
SL-EN	1.00	100%
Mean	0.77	91.75%
* If the $scoreDiff$ is 1 or less, $weight = 0$; otherwise, $weight = (scoreDiff - 1)^2$		

increased and reached its peak on the average score of 4; this represents over one-fifth of the document pairs (22%). Lastly, around 25% documents pairs were judged to have very high proportion of content overlap (an average score of 4.5 or above). The distribution of these scores across the eight language pairs is shown in Figure 5.11.

When these scores are aggregated into two groups, shown in Table 5.8, similar proportion was found to Q1; 36% document pairs were found to have a low content overlap (average score between 1 and 3), and 64% have a high content overlap (average score of 3.5 and above). These proportions vary widely across the eight language pairs. The majority of SL-EN document pairs were assessed to have content overlap (95 document pairs), whilst the number of document pairs for ET-EN and LV-EN with high content overlap are fewer than half (44 document pairs each).

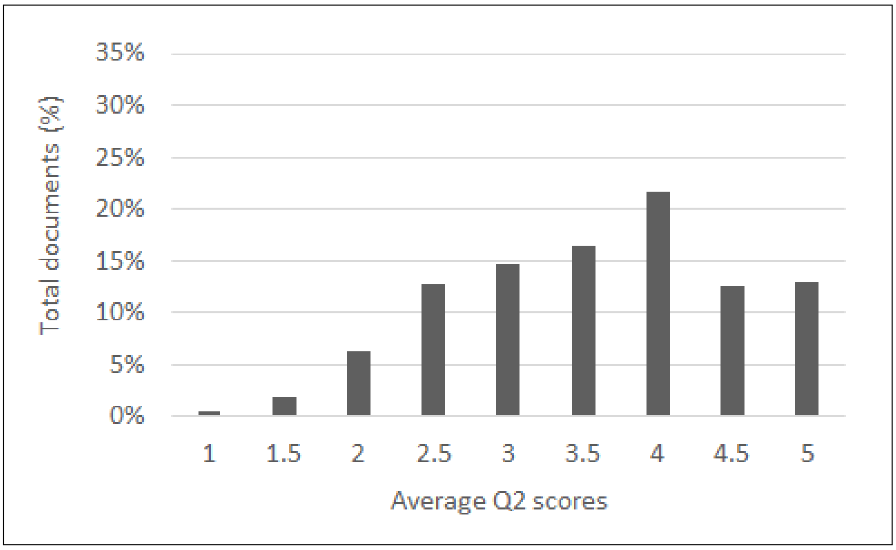


Fig. 5.10 Distribution of content-overlap scores (Q2) averaged across both assessors (N=800)

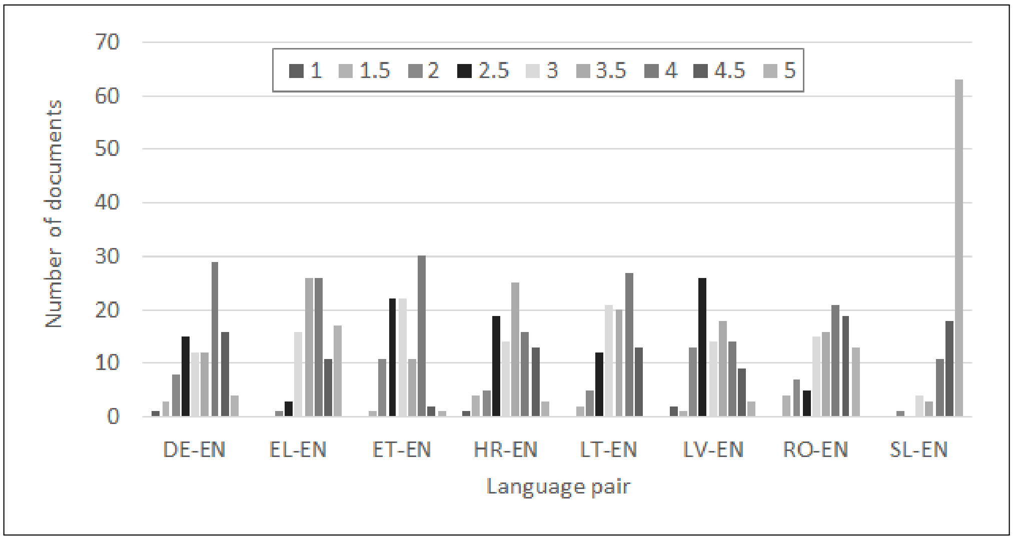


Fig. 5.11 Distribution of document-level similarity scores (Q2) per language pair (N=100), averaged across both assessors

Table 5.8 Proportion of content overlap for all language pairs

Language Pair	Document Pairs with Low Content Overlap (Avg Q2 < 3.5)	Document Pairs with High Content Overlap (Avg Q2 ≥ 3.5)	Total
DE-EN	39	61	100
EL-EN	20	80	100
ET-EN	56	44	100
HR-EN	43	57	100
LT-EN	40	60	100
LV-EN	56	44	100
RO-EN	31	69	100
SL-EN	5	95	100
All	290 (36%)	510 (64%)	800

5.5.4 Results on Q3: "Of the shared content (if there is any), on average how similar are the matching sentences?"

In the third question of the evaluation task, assessors were asked to identify the similarity between the matching sentences of the shared contents (if they exist). In a small number of cases (31 document pairs), at least one assessors identified that there was no content overlap between the document pairs (Q2 equals 1). Since the aim of Q3 is to identify the similarity between sentences in the overlapping contents (instead of the entire documents), cases where no content overlap was identified are not relevant in this analysis. Therefore, these 31 document pairs were taken out of the results. The remaining 769 document pairs that were annotated to contain overlapping contents by both assessors (i.e., Q2 equals 2 or higher) were analysed further in this section.

In this task, both assessors provided the same scores in 52.39% of the cases. The agreement scores are shown in Table 5.9. Upon increasing the tolerance to take into account cases with different scores, there were 90.2% agreement (694 cases) when taking into account cases that differ by one or less, and 97.9% agreement (753 document pairs) for those that differ by two or less. The agreement between assessors (with a score tolerance of 1) is shown in Table 5.10.

These scores were further averaged between assessors as shown in Figure 5.12. The resulting scores show that when document pairs were identified to have matching con-

Table 5.9 Inter-assessor agreement for Q3 (5 classes, N=769 doc pairs)

Lang Pair	Spearman's ρ	Weighted Cohen's Kappa	Krippendorff's α	Agreement
DE-EN	0.52	0.50	0.52	52.1%
EL-EN	0.45	0.42	0.45	56.0%
ET-EN	0.63	0.47	0.57	46.0%
HR-EN	0.53	0.52	0.48	52.6%
LT-EN	0.45	0.31	0.16	28.1%
LV-EN	0.61	0.58	0.55	53.8%
RO-EN	0.52	0.60	0.47	50.5%
SL-EN	0.39	0.24	0.14	72.0%
Mean	0.51	0.46	0.42	52.39%

Table 5.10 Inter-assessor agreement for Q3 (5 classes, score difference tolerance of 1, N=769 doc pairs)

Lang pair	Weighted Cohen's Kappa*	Agreement
DE-EN	0.62	85.4%
EL-EN	0.55	91.0%
ET-EN	0.64	85.7%
HR-EN	0.74	96.8%
LT-EN	0.59	88.5%
LV-EN	0.77	90.1%
RO-EN	0.74	89.2%
SL-EN	0.28	95.0%
Mean	0.62	90.2%

* If the $scoreDiff$ is 1 or less, $weight = 0$; otherwise, $weight = (scoreDiff - 1)^2$

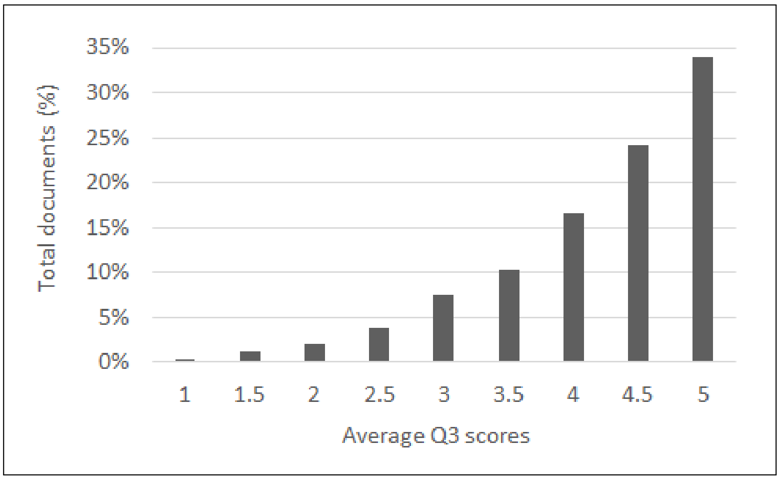


Fig. 5.12 Distribution of sentence-level similarity scores (Q3) averaged across both assessors (N=769)

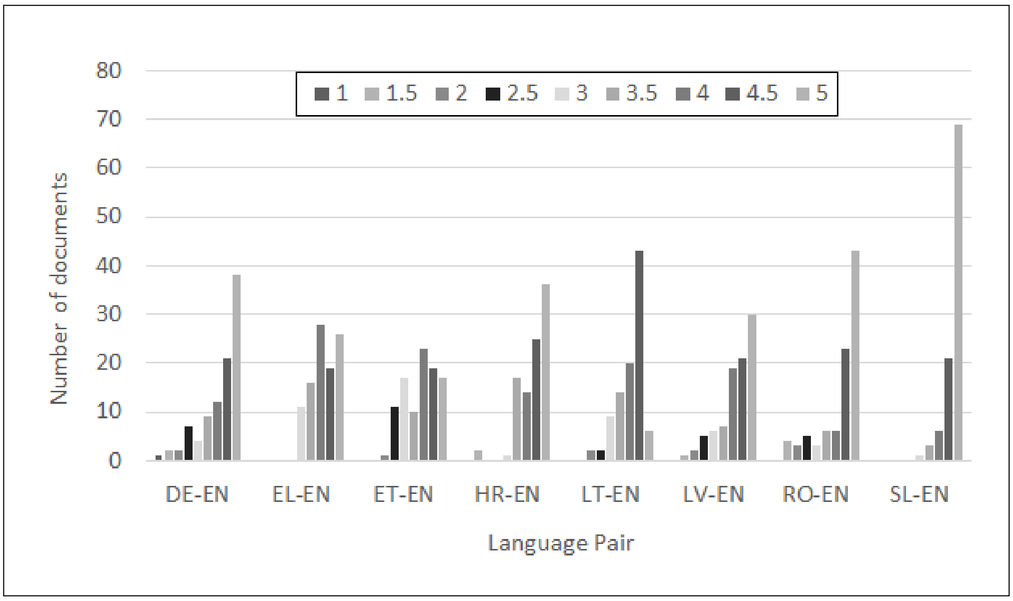


Fig. 5.13 Distribution of document-level similarity scores (Q3) per language pair (N=91-100), averaged across both assessors

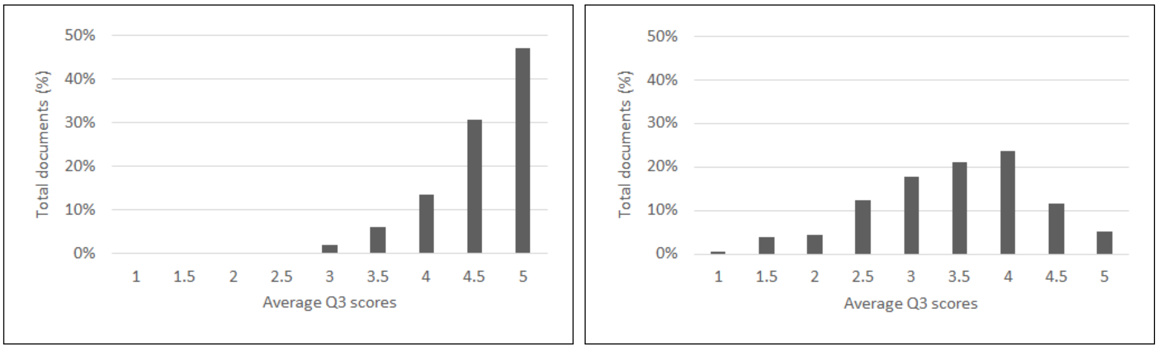
tents, very small percentage were annotated to be non-similar sentences or sentences with low similarity. In the evaluation set, the number of document pairs increased with higher degree of sentence similarity. Moreover, up to 34% of the documents had an average sentence similarity of 5, i.e., they were identified to have very high sentence similarity in the matching contents by both assessors. The results across the eight language pairs are shown in Figure 5.13.

Aggregating these scores into two groups results in the statistics shown in Table 5.11. Over all 769 document pairs, 86.74% document pairs were assessed to have high sentence similarity within the matching contents. Considering Q1 and Q2 results into account, i.e., that only two-thirds of the evaluation set are highly similar and contain high content overlap, these results suggest that low-similar documents may still contain similar sentences that can be extracted and used as valuable bilingual resources.

An analysis was further performed to identify whether the proportion of similar sentences depends on the degree of similarity of the documents, i.e., whether similar document pairs are more likely to have more similar sentences compared to non-similar document pairs. When considering only similar document pairs (average Q1 score ≥ 3.5)

Table 5.11 Proportion of similar sentences for all language pairs

Language Pair	**Document Pairs with Low Sentence Similarity (Avg Q3 < 3.5)**	**Document Pairs with High Sentence Similarity (Avg Q3 ≥ 3.5)**	**Total**
DE-EN	16	80	96
EL-EN	11	78	100
ET-EN	29	69	98
HR-EN	3	92	95
LT-EN	13	83	96
LV-EN	14	77	91
RO-EN	15	78	93
SL-EN	1	99	100
All	102 (13.26%)	667 (86.74%)	769



(a) Similar document pairs (N=536)

(b) Non-similar document pairs (N=233)

Fig. 5.14 Distribution of sentence-level similarity scores (Q3) in similar and non-similar document pairs

(Figure 5.14a), most of the overlapping contents (98%) contain highly similar sentences (average Q3 score ≥ 3.5). In contrast, over all non-similar document pairs (average Q1 score < 3.5), only 61% were annotated to contain highly similar sentences (average Q3 score ≥ 3.5). Furthermore, these non-similar document pairs also contain a larger proportion of sentences with low similarity (average Q3 score < 3.5), i.e., 39%, compared to those found in similar document pairs, i.e., 2%.

5.5.5 Results on Q4: “Overall, what is the comparability level between these two documents?”

In the last evaluation question, assessors were asked to assign a score to represent the comparability level of each document pair. As reported in Section 5.4, all assessors were familiar with the concept of comparability, as they were partners of an EU-funded project whose tasks were to identify comparable documents from the Web for the purpose of extracting bilingual resources for improving machine translation.

Correlation between assessors in answering Q4 is similar to Q1 (Spearman’s ρ = 0.59). Upon deciding a score on the document comparability on a 5-point scale (weighted Cohen’s Kappa = 0.47), assessors agreed with each other in 48% of the document pairs (shown in Table 5.12). This agreement increased to 93.25% when also considering cases where both assessors’ scores differed by one (shown in Table 5.13), and in 99.6% for scores differing by two or fewer.

Although comparability and similarity were expected to relate to each other (further explored in Section 5.5.6), the distribution of scores between them was slightly different. As shown in Figure 5.15, just under 15% document pairs were given an average comparability score of 4 (the highest in the dataset), and around 32% when considering those scoring 3.5 or above. The rest of the document pairs (68%) scored 3 or below and were assigned to be weakly-comparable. This is in contrast to the Q1 results, where 67% document pairs were judged to be similar. The distribution of Q4 scores per language pair is

Table 5.12 Inter-assessor agreement for Q4 (N=800)

Lang Pair	**Spearman’s ρ**	**Weighted Cohen’s Kappa**	**Krippendorff’s α**	**Agreement**
DE-EN	0.62	0.47	0.44	42%
EL-EN	0.58	0.48	0.49	59%
ET-EN	0.52	0.49	0.52	69%
HR-EN	0.70	0.45	0.35	24%
LT-EN	0.58	0.28	0.04	23%
LV-EN	0.56	0.45	0.38	43%
RO-EN	0.70	0.66	0.68	59%
SL-EN	0.45	0.44	0.29	65%
Mean	0.59	0.47	0.40	48%

Table 5.13 Inter-assessor agreement for Q4 (5 classes, score difference tolerance of 1, N=800 doc pairs)

Lang pair	Weighted Cohen's Kappa*	Agreement
DE-EN	0.69	91%
EL-EN	0.73	97%
ET-EN	0.82	99%
HR-EN	0.75	87%
LT-EN	0.62	92%
LV-EN	0.67	83%
RO-EN	0.90	97%
SL-EN	1.00	100%
Mean	0.77	93.25%

* If the $scoreDiff$ is 1 or less, $weight = 0$; otherwise, $weight = (scoreDiff - 1)^2$

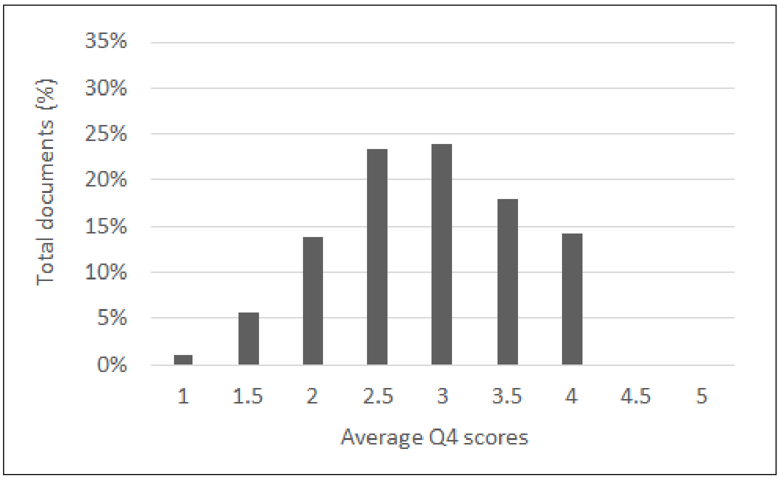


Fig. 5.15 Distribution of document-level comparability scores (Q4) averaged across both assessors (N=800)

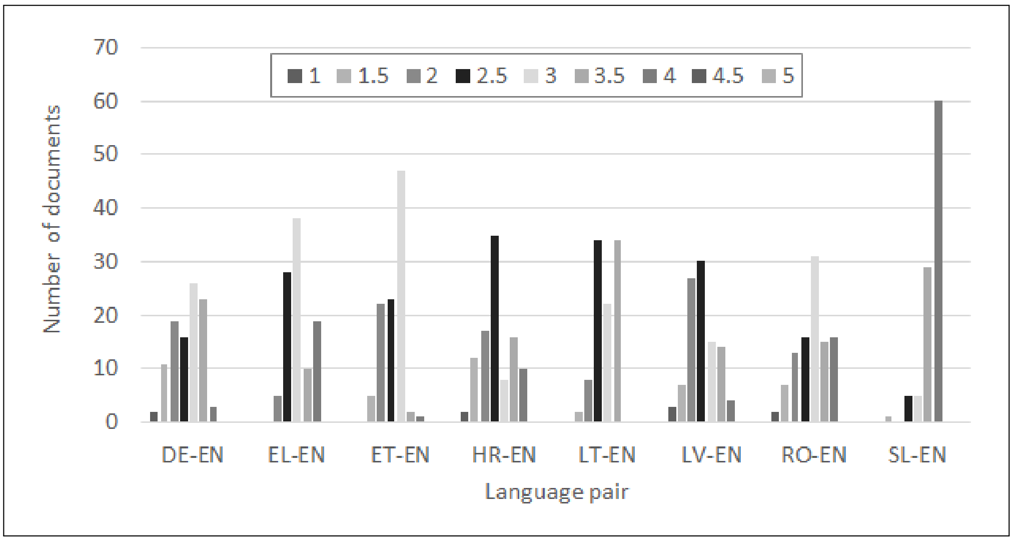


Fig. 5.16 Distribution of document-level comparability scores (Q4) per language pair (N=100), averaged across both assessors

shown in Figure 5.16. In all language pairs, except SL-EN, a higher proportion of documents were assessed to have an average scores of 2.5 and 3. SL-EN, however, have a high proportion of document pairs with a comparability score of 4.

The results show that whilst comparability is often deemed to be highly related to document similarity, the proportion of highly comparable documents is much lower than the proportion of similar documents in the dataset. Several reasons were found to explain this finding. First, when considering the comparability of document pairs, given the background of the assessors, they aimed to identify and gather documents that would be valuable for training machine translation systems. A number of document pairs that were included in the evaluation set contain, for example, contents that were both written in English (i.e. both in the English and non-English pair), many overlaps of named entities but not other contents (such as a document about a list of towns called ‘Georgia’); these documents, whilst similar, would not be as valuable as comparable documents and therefore were scored lower.

Table 5.14 Proportion of comparability levels for all language pairs

Language Pair	Weakly Comparable Document Pairs (Avg Q4 < 3.5)	Highly Comparable Document Pairs (Avg Q4 ≥ 3.5)	Total
DE-EN	74	26	100
EL-EN	71	29	100
ET-EN	97	3	100
HR-EN	74	26	100
LT-EN	66	34	100
LV-EN	82	18	100
RO-EN	69	31	100
SL-EN	11	89	100
All	544 (68%)	256 (32%)	800

5.5.6 Comparison between evaluation questions

Spearman's rank correlation coefficient (ρ) was used to analyse correlations between the different evaluation questions across the 5-point scale, shown in Table 5.15. In this section, the author reports the correlations between the average scores between both assessors for each pair of evaluation questions.

There is a significant correlation between the document similarity (Q1) and the overall proportion of shared contents (Q2) (ρ = 0.94; p<0.01), suggesting that the more overlap between information in article pairs, the greater the perceived degree of similarity (Figure 5.17a).

There is a strong correlation between the similarity level (Q1) and the comparability level (Q4) (ρ=0.92; p<0.01). However, document pairs that were previously identified to be similar were shown to have lower degrees of comparability (Figure 5.17b). Having strong backgrounds in comparability, annotators identified highly comparable doc-

Table 5.15 Spearman correlation between questions (mean between assessors, N=800)

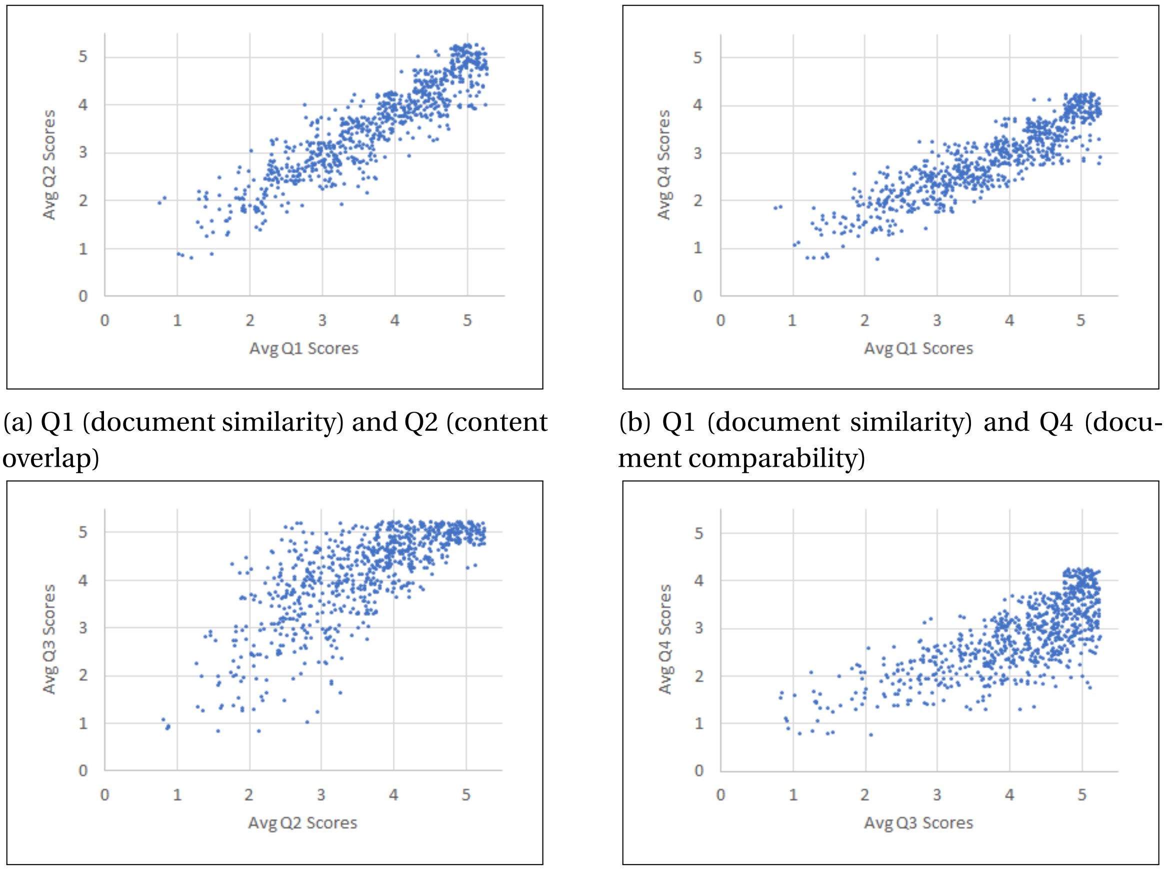
	Q1 (Similarity)	Q2 (SharedCont_Prop)	Q3 (SharedCont_SentSim)	Q4 (Comparability)
Q1	1	0.94	0.80	0.92
Q2	0.94	1	0.78	0.93
Q3	0.80	0.78	1	0.77
Q4	0.92	0.93	0.77	1

uments as those that contain valuable bilingual resources. This includes matching sentences written in multiple languages. Named entities, on the other hand, are not deemed to be as valuable; hence, documents containing many named entities were scored lower on their degrees of comparability.

A lower correlation was found when calculating correlation scores between document similarity (Q1) and the sentence similarity (Q3) (ρ=0.80; p<0.01). Similar results were identified when calculating correlations between the overall proportion of shared contents (Q2) and the sentence similarity in the matching contents (Q3) (ρ=0.78; p<0.01). Figure 5.17c shows that articles with high proportion of shared contents often have high sentence similarity too, and articles that do not have any shared contents have very low sentence similarity. Interestingly, article pairs that shared smaller proportion of contents (Q2 score between 2-4) can still have high sentence similarity. These documents need to be taken into account when extracting linguistic resources from Wikipedia.

The lowest correlation was found between the document comparability (Q4) and sentence similarity (Q3) (ρ=0.77; p<0.01). Figure 5.17d shows that the majority of article pairs perceived to be highly comparable (Q4 score of 4) also contained high sentence similarity in their overlapping contents. Article pairs scoring 2 or 3 may also contain high sentence similarity, although also contain a proportion of less similar sentences. Article pairs scoring 1, on the other hand, have more sentence pairs perceived with lower similarity; although they may also contain highly similar sentence pairs.

A comparison of the distributions of average scores for all the four evaluation questions are shown in Figure 5.18. The results show that the distribution of scores of Q1 and Q2 were similar, suggesting that the similarity between document pairs correspond to the amount of content overlap. Q3, on the other hand, captured a higher number of document pairs with high sentence-level similarity, an issue that was not captured in the other questions.



(a) Q1 (document similarity) and Q2 (content overlap)

(b) Q1 (document similarity) and Q4 (document comparability)

(c) Q2 (content overlap) and Q3 (sentence similarity)

(d) Q3 (sentence similarity) and Q4 (document comparability)

Fig. 5.17 Correlation of scores across different questions (N=800)

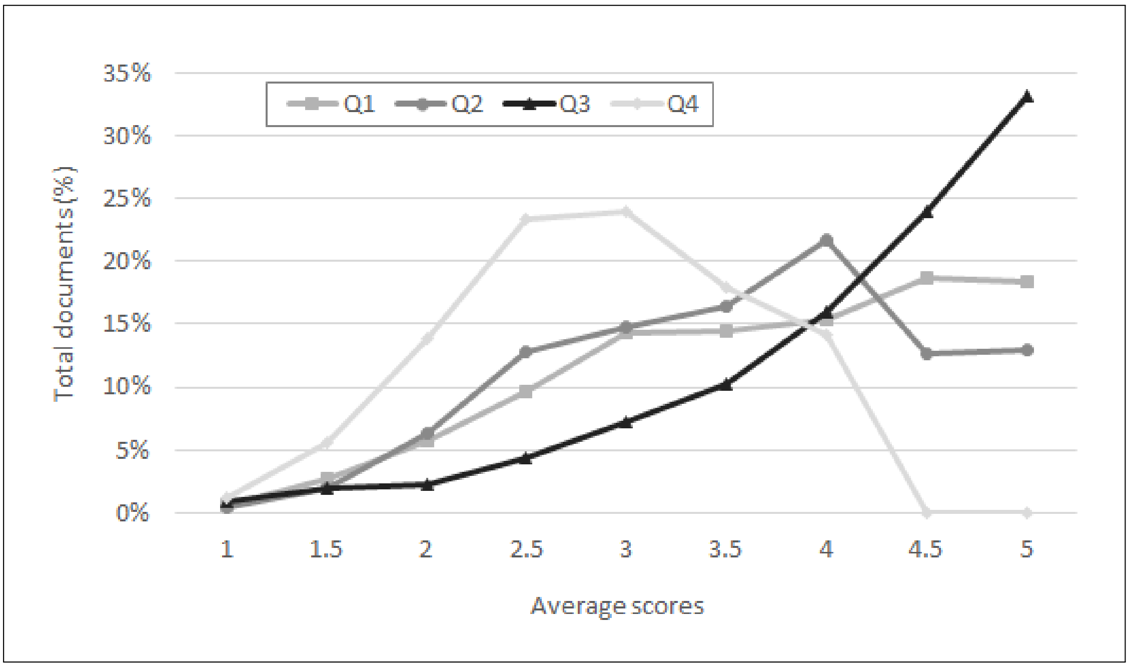


Fig. 5.18 Comparison of the distribution of average scores across all evaluation questions

5.6 Discussion

In this section, the author discussed the results further to identify the cases of major disagreements between assessors. The relations between 'similarity' and 'comparability' aspects of the document pairs are also discussed further. And finally, the author described the limitations of this work.

5.6.1 Analysis of disagreements

The author analysed cases where assessors had major disagreements on annotating the similarity of the document pair, i.e., the similarity scores differed by three points or more. The disagreement cases include document pairs that contained a large number of matching entities. Some assessors identified these contents to be similar and gave high scores to the documents; meanwhile, other assessors ignored these contents as they were not deemed to be valuable, and based their judgments on the remainder of the document pairs. A small number of disagreements occured due to human-error. Some examples were documents that looked similar at the document level, e.g., they shared similar structure, but the contents at the sub-document level were different. The latter might be missed by some assessors and therefore caused the scores to be different. Lastly, disagreements also often occured when the document lengths were very different. Some assessors punished these document pairs by lowering the scores although some contents were similar. As reported in the previous sections, however, these major disagreements only occured in a small number of cases: 2.37% (19 document pairs out of 800 pairs) for Q1, and 0.5% (4 document pairs), 3.38% (27 document pairs) and 0.37% (3 document pairs) for Q2, Q3 and Q4, respectively.

5.6.2 Relations between 'similarity' and 'comparability'

The evaluation corpus allowed us to investigate further the relations between 'similarity' and 'comparability'. The results indicated that these aspects correlated strongly (ρ=0.92; p<0.01), although document pairs generally achieved lower comparability scores than similarity scores. The scatter plot (Figure 5.17b) shows that although some document pairs were annotated as highly similar (score of 5), they only achieved a moderate to moderate-high comparability scores (comparability scores of 3 and 4). One reason for this us that a full comparability score is often defined to represent parallel documents (Fung & Cheung, 2004; Skadiņa et al., 2012), i.e., those that are translations of each other. Due to the open-editing nature of Wikipedia, even articles that were created as translations of another article would gradually differ in time based on various additions, removals, or modification of the contents. Wikipedia document pairs that resemble parallel documents are therefore scarce.

The scatter plot further shows that there are other cases found in the corpus where the similarity and comparability scores were considerably different. Firstly, some document pairs annotated to be 'similar' were annotated to be not 'comparable' due to the lack of translated contents. For example, there are Wikipedia articles that contains list of named entities, such as an article listing the names of athletes that competed in Olympic Games 2016. This article is available in a large number of languages, but they all contain the same names throughout the different language versions. Another example is one Wikipedia article that lists the names of schools in Estonia. The same article that is written in other language versions list the Estonian names of the schools, rather than the translations. Again, these types of documents are annotated to be similar. However, from comparability perspective, there is a lack of bilingual resources that can be extracted from these documents due to the amount of contents written in the same language; as a result, this document pair was given a low comparability score. These findings indicate that 'similar' articles are thought to be 'comparable' only if they contain contents written in different languages. E.g., document pairs containing named entities that are the same in the two languages are considered to be less comparable than document pairs containing terms (and their translations) in two languages, as the latter were more useful as bilingual resources.

On the other spectrum, document pairs can be annotated to be 'comparable' yet not 'similar'. Previous work have previously shown that articles that contain different topics may still contain some translated contents (e.g., translation of terms) if they are from the same domain. These articles are referred to as 'weakly comparable' (Skadiņa et al., 2012). Since the scope of this study is on interlanguage-linked Wikipedia articles only, these documents were paired because they described the same topic. I.e., this corpus does not contain any document pairs that do not describe the same topic. Cases where 'non-similar' documents were found to be 'comparable' are, therefore, not found in this corpus.

5.6.3 Limitations

The author identified two *limitations of the Wikipedia similarity dataset* with regards to the size of the dataset and the size of the articles.

Firstly, the evaluation set only contains 100 document pairs per language pair. These document pairs were selected using a stratified sampling of the anchor word and text method (described in Chapter 6) in order to include 100 document pairs with a wide range of similarity into the evaluation corpus. As a result, this evaluation set may not represent the distribution of similarity in Wikipedia in general. Furthermore, this may introduce a bias towards the proportion of document pairs with high overlap of links and words that are included in the evaluation set, although those document pairs may not occur very frequently in Wikipedia. On the other hand, the proportion of document pairs with low similarity was also shown to be smaller. Future work is required to improve the evaluation corpus to add more non-similar instances to provide a more balanced evaluation corpus.

Secondly, most of the document pairs contain only up to 1,000 words, which meant that larger Wikipedia articles were not represented in the evaluation set. This limitation was set to reduce assessors' fatique. In the current form of the evaluation task, assessors were required to read the document contents prior to assigning a similarity and comparability score, identifying matching contents and assessing the sentence similarity in the matching contents. These tasks became extremely difficult when assessing document pairs that were too long. Including these document pairs (i.e., documents with word length over 1,000 words) may instead introduce inaccuracies in the dataset due to a higher probability of assessors' fatigue and human error. By limiting the size of documents, assessors were able to focus more time on reading the document contents in order to reliably assess them.

Taking into account the limitations, however, this evaluation set still represents a valuable resource for measuring similarity methods. This evaluation set includes inter-language-linked articles with different similarity degrees to provide better resources for training and evaluation of different approaches. Moreover, the set also captures various issues that affect similarity of documents. These findings can be used for further understanding the similarity in Wikipedia articles, improving automatic methods to measure similarity and performing an automatic evaluation of the methods.

5.7 Conclusion

This chapter has described the work in creating an evaluation corpus specifically for Wikipedia. In this section, the author answers the research questions presented earlier in this chapter.

RQ1. What are the characteristics of similar interlanguage-linked articles in Wikipedia? The evaluation corpus has identified that Wikipedia articles with different scores exhibit different similarity characteristics. Similar document pairs were shown to contain *similar structure, overlapping named entities, overlapping fragments,* and *translated contents.* A high proportion of translated contents characteristic was only found in documents with the highest similarity scores (Q1 scores of 5). However, the first three characteristics (i.e., similar structure, overlapping named entities and fragments) could still be found in document pairs with lower similarity scores (Q1 scores of 3 or above). A high proportion of non-similar document pairs (Q1 scores of 2 or below) was shown to contain *different information,* although a small proportion of these documents may still contain similarity at the sub-document level, e.g., same entities and similar sentences or phrases.

RQ2. Can we create an evaluation benchmark for Wikipedia? I.e., do human assessors agree on Wikipedia similarity? The author has proposed an evaluation scheme to gather human judgments on 800 Wikipedia documents in 8 language pairs. These documents were selected using the anchor text and word overlap method and a stratified sampling in order to include documents containing a wide range of similarity. This corpus allows Wikipedia characteristics to be investigated in more detail, and for future similarity measures to be evaluated against the human judgments. Overall, a moderate agreement was achieved between the assessors across all the evaluation questions (mean weighted Cohen's Kappa between 0.46 and 0.56). Since no specific guidelines were created to define the different scores in the evaluation questions, it was expected that assessors' answers would differ slightly due to the assessors' different point of views behind the different scores. This was proven by the increased weighted Cohen's Kappa score when cases where assessors' answers differed by one were considered as an agreement (mean weighted Cohen's Kappa between 0.62 and 0.77), showing a good agreement between assessors for each evaluation question.

Future work should explore including more document pairs to increase the size of the corpus, and increase the size of the evaluation documents to make the similarity corpus more representative of the state of Wikipedia. Extending this corpus is out of the scope of this work, but would be a promising step to move forward to strengthen the current evaluation benchmark.

Related publication

- Paramita, M., Clough, P., Aker, A. and Gaizauskas, R. 2012. Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012),* Istanbul, Turkey, pp. 790-797.

Chapter 6

Anchor Text and Word Overlap Method

The findings in Chapters 4 and 5 have further confirmed previous work: that interlanguage-linked articles in Wikipedia exhibit varying degrees of similarity. Furthermore, different characteristics that contribute to the similarity between a document pair have also been identified. Based on this information and related literature, this thesis investigates four different methods (as shown in Figure 1.1 in page 11) to measure cross-lingual similarity in Wikipedia. This chapter reports the first of four experiments that have been carried out to develop and analyse methods to identify similarity in Wikipedia interlanguage-linked articles. In this experiment, the author developed a method to measure similarity between a document pair using the similarity of Wikipedia links between both articles. Similar information across languages are identified using the interlanguage links information in Wikipedia.

6.1 Background

Information derived from interlanguage links has previously been used to identify similar information across different languages in Wikipedia. One approach is the *link-based bilingual lexicon approach* (Adafre & de Rijke, 2006), previously described in Section 2.4.2. This approach is language-independent and does not require any translation resources. Instead, it creates its own translation resources (further referred to as a *bilingual lexi-*

con) for a language pair by extracting titles of all Wikipedia interlanguage-linked articles in that language pair. This bilingual lexicon is then utilised for identifying similar content in different languages. Adafre and de Rijke (2006) showed that this approach was able to identify similar sentences in Dutch-English with high precision, although low recall was observed. Its performance in other language pairs, especially under-resourced languages, has to date not been studied.

In this experiment, the author investigated the use of this approach in identifying cross-lingual similarity in 8 different language pairs. This method was selected because it relied only on information within Wikipedia and therefore could be applied to other language pairs (that were available in Wikipedia) without requiring any external linguistic resources. An adaptation of this method is proposed to identify similarity in Wikipedia at the document level.

The proposed method, referred to as *the anchor text and word overlap method* (or $anchor + word$), differs to the link-based bilingual lexicon approach (Adafre & de Rijke, 2006) in four ways. Firstly, prior to measuring similarity, Adafre and de Rijke (2006) represented each sentence using the links only, i.e., any words that are not linked to any Wikipedia article are discarded (see example in Table 2.5 in page 48). However, the author suggests that some of these non-linked words (such as numbers or named entities) may appear the same across languages and should be taken into account when measuring similarity. Therefore, the proposed $anchor + word$ method represents each sentence using both the anchor texts (i.e., clickable texts or linked words in the articles) and the remaining words (i.e., non-linked words in the articles). This approach is proposed to increase the recall of the method.

Secondly, Adafre and de Rijke (2006) carried out a sentence alignment by allowing only a one-to-one correspondence between similar sentences. However, as identified in Section 4.4.3, similar contents appear in the document pair, but do not correspond to a one-to-one alignment at the sentence level. I.e., contents described in one sentence in one article may be represented in more than one sentences in the other article. To accommodate this, the $anchor + word$ method allows a many-to-one correspondence when aligning similar sentences.

Thirdly, Adafre and de Rijke (2006) also used information extracted from the redirection pages to build the bilingual lexicon. Further research, however, has shown that this significantly decreased the accuracy of the extracted bilingual terms (from 92.3% to 23.1% in German-English as reported in Erdmann et al. (2009)). Therefore, in this study, the author only extracted the titles from the interlanguage-linked articles when building the bilingual dictionary.

Finally, the link-based bilingual lexicon approach identifies similarity at the sentence level. The $anchor + word$ method, on the other hand, identifies similar sentences and further aggregates the information to measure similarity at the document level.

This experiment aims to answer the third research question:

RQ3. Can language-independent approaches be used to identify cross-lingual similarity in Wikipedia?

(a) How does the method compare to approaches using linguistic resources, such as MT systems?

(b) How does the performance for the approach vary for different language pairs?

(c) What language-independent features are best for measuring cross-lingual similarity in Wikipedia?

First, the author describes the method in Section 6.2 and the experiments in Section 6.3. The method is evaluated against a baseline that utilises a MT system (in this case, Google Translate). The results are reported in Section 6.4. Finally, the author discusses the results and concludes the experiment in Section 6.5 and Section 6.6.

6.2 Method

The $anchor + word$ method contains four processes (shown in Figure 6.1). Firstly, a pre-processing step is carried out to extract article contents from the Wikipedia dumps.

Secondly, a bilingual lexicon is created by utilising the titles of interlanguage-linked articles. The bilingual lexicon is then used to perform the anchor text translation process. Similarity is calculated for all possible sentence pairs and the highest scoring sentence pairs were aligned. Finally, similarity is calculated between the document pair by aggregating the similarity scores for the aligned sentences. These processes are described in more detail below.

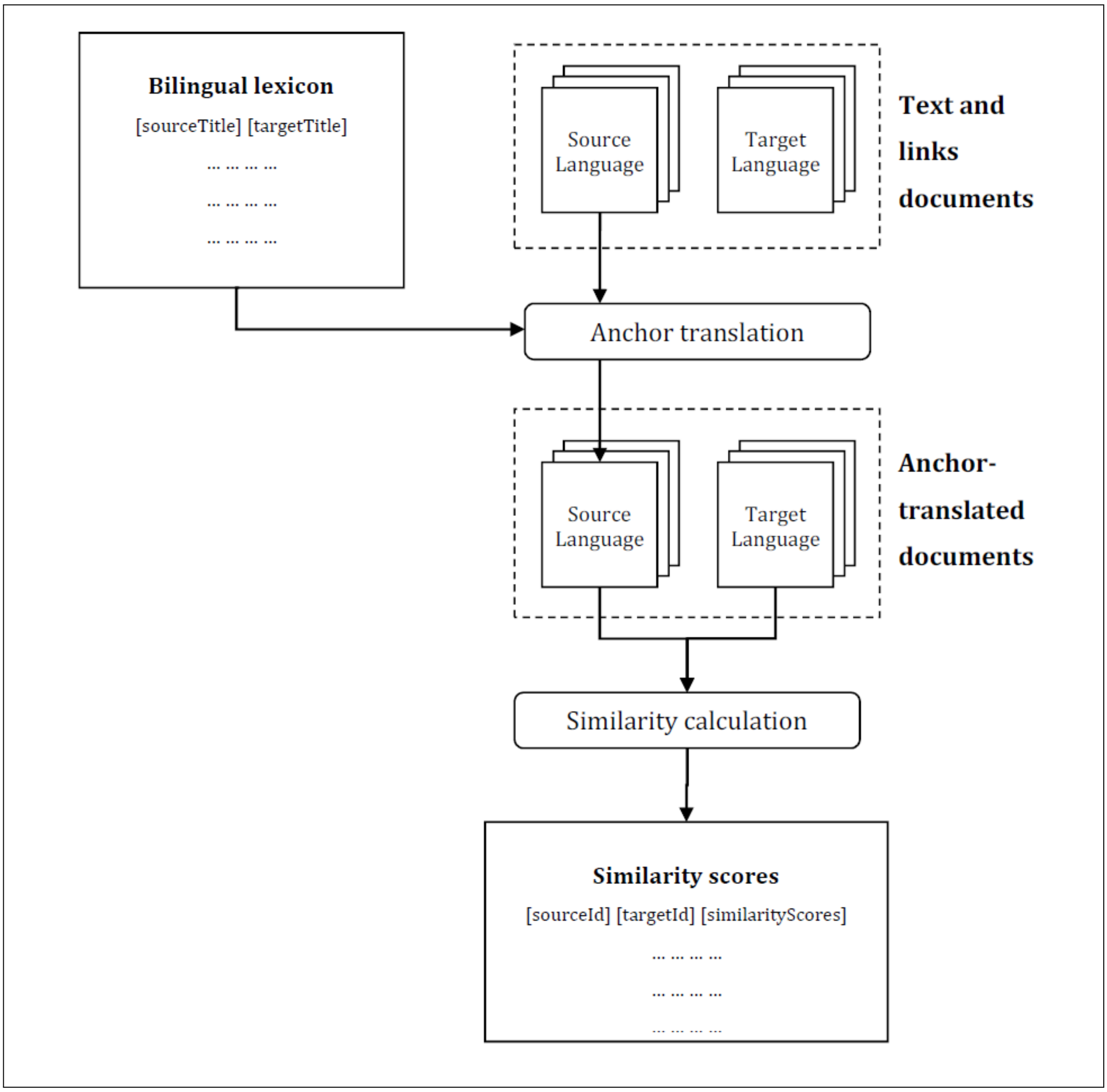


Fig. 6.1 Anchor text and word overlap approach

...
A large proportion have orbital elements similar to those of 4 Vesta, either close enough to be part of the [[**vesta family**]], or having similar [[**eccentricity (orbit)**]] and [[**inclination**]]s but with a [[**semi-major axis**]] lying between about 2.18 [[**astronomical unit**]] and the 3:1 [[**Kirkwood gap**]] at 2.50 AU.
...

Fig. 6.2 Excerpt of the English Wikipedia article of “V-type asteroid”

6.2.1 Pre-process documents

Firstly, given a Wikipedia dump for a language pair (referred to as a source language and target language), the documents are pre-processed to filter out contents in the infoboxes, tables, images and reference lists. The remaining text and links are preserved at this stage. The contents are further split into sentences. The output of the pre-processing stage is referred to as the *text-and-link version of the article*. More information about these processes have previously been described in Section 3.6.1.

An example of a text-and-link English article is shown in Figure 6.2.[1] The links are shown inside the brackets (i.e., [[links]]).

6.2.2 Create a bilingual lexicon

The *anchor+word* method utilises information from the interlanguage links to build a bilingual dictionary (referred to as the bilingual lexicon) to assist with the translation process. The bilingual lexicon is created by extracting titles of interlanguage-linked articles in both source and target languages. These titles often have translation relations between them and are valuable as bilingual resources (Adafre & de Rijke, 2006). Duplicate titles were removed prior to creating the bilingual lexicon.

Each entry in the resulting bilingual lexicon contains two items: the title of article in the source language and the title of the paired interlanguage-linked article in the target language. For example, an excerpt of the Slovenian-English bilingual lexicon is shown

[1] This excerpt is taken from a Wikipedia article titled “V-type asteroid’; the most current version of this topic is available in Wikipedia: `https://en.wikipedia.org/wiki/V-type_asteroid`.

Astronomija	Astronomy
Bitka pri Trafalgarju	Battle of Trafalgar
Velika nagrada Velike Britanije	British Grand Prix
Tehnika	Engineering
Anglija	England
Reokavski preliy	English Channel
Halucanija	Hallucination
...	...

Fig. 6.3 Excerpt of a Slovenian-English bilingual lexicon extracted from Wikipedia

in Figure 6.3. More information on building bilingual lexicons has been described in Section 3.6.1.

6.2.3 Anchor text translation

The text-and-links version of the documents contain the text contents of the article and information about clickable links in the documents, further referred to as the *anchor texts*. Each anchor text is linked to a Wikipedia article. In some cases, the anchor text may be the same as the title of the linked article. E.g., an example of this is "[[astronomical unit]]" (shown in Figure 6.2), which means the anchor text "astronomical unit" is linked to an article titled "astronomical unit". In some cases, the anchor text may be different to the title of the linked article to make the text more readable in the plain text version (i.e., the version shown to the readers). An example of this case is "[[List of countries by area|78th-largest sovereign state in the world]]", shown in Table 6.1. In this case, the anchor text (clickable link) was "78th-largest sovereign state in the world" and it links to an article titled "List of countries by area".

To translate the anchor text, first, it is transformed into the title of its linked article (if different). If the title exists in the bilingual lexicon, the title is translated into the target language using the bilingual lexicon previously created for that particular language pair. This process allows overlapping information across languages to be identified. If the title does not exist in the bilingual lexicon, it is kept in its original form (i.e., in the source language).

Table 6.1 Example of different links and anchor texts

Text and links	... With an area of 242,500 square kilometres (93,600 sq mi), the United Kingdom is the [[**List of countries by area|78th-largest sovereign state in the world**]]. ...
Plain text	... With an area of 242,500 square kilometres (93,600 sq mi), the United Kingdom is the 78th-largest sovereign state in the world. ...

For example, Figure 6.4a shows a Slovenian article prior to the anchor text translation; the anchor texts are shown in bold. Using the bilingual lexicon (in this example, a Slovenian-English bilingual lexicon), the anchor texts are replaced with their equivalent English links. The resulting article after the translation is shown in Figure 6.4b.

All anchor texts found in the target articles are also transformed into their corresponding links (i.e., titles of linked articles). However, no translation is performed in the target articles. In the example above, the paired English article for the "V-type asteroid" topic is shown in Figure 6.2. After translating the Slovenian links, the translated Slovenian article (Figure 6.4b) and the English article (Figure 6.2) are shown to share the following links: [[vesta family]], [[astronomical unit]] and [[Kirkwood gap]]. This example shows how the $anchor + text$ method can identify overlapping information across languages by utilising the Wikipedia interlanguage links.

6.2.4 Similarity calculation

After all anchor texts in the source documents have been translated using the bilingual lexicon, similarity is calculated in two stages, at the sentence level, and at the document level.

...
Večinajih je v bližini [[**družina Vesta|asteroidne družine Vesta**]].

Imajo podobne [[**izsrednost|izsrednosti**]], toda njihova [[**elipsa|velika polos**]] leži v območju od 2,18 [[**astronomska enota|a. .e.**]] do 2,50 a. e. (kjer je [[**Kirkwoodova vrzel|Kirkwoodova vrzel**]] 3 : 1).
...

(a) Before anchor text translation

...
Večinajih je v bližini [[**vesta family**]].

Imajo podobne [[**eccentricity**]], toda njihova [[**ellipsis**]] leži v območju od 2,18 [[**astronomical unit**]] do 2,50 a. e. (kjer je [[**Kirkwood gap**]] 3 : 1).
...

(b) After anchor text translation

Fig. 6.4 Slovenian text of “V-type asteroid”

Similarity at the sentence level

Given a document pair d_1 and d_2 written in language L_1 and L_2, respectively, where $sentenceCount(d_1) \leqslant sentenceCount(d_2)^2$, all sentences $(s_1, s_2, ..., s_m)$ in d_1 are paired to all sentences $(t_1, t_2, ..., t_n)$ in d_2. Similarity between a sentence pair s_i and t_j is calculated using the Jaccard coefficient:[3]

$$sentAlignScore(s_i, t_j) = \frac{words_{s_i} \cap words_{t_j}}{words_{s_i} \cup words_{t_j}} \tag{6.1}$$

where $words_{s_i}$ and $words_{t_j}$ represents a set of unique words in sentence s_i and t_j, respectively.

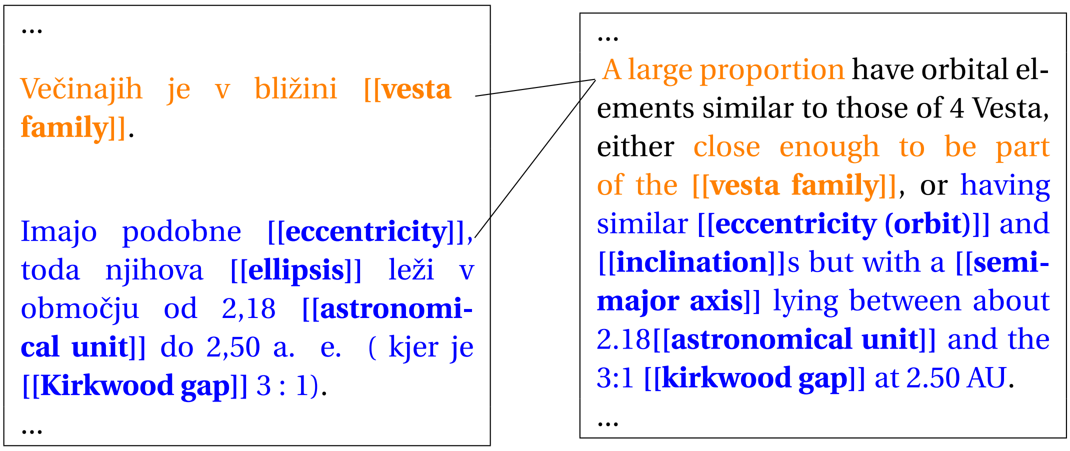
After all sentence pairs have been scored, similar sentences are identified by aligning each sentence s_i in document d_1 with the highest scoring sentence in document d_2:

$$sentAlignScore(s_i) = \max_{1 \le j \le n} sentAlignScore(s_i, t_j) \tag{6.2}$$

I.e., for a sentence s_i, the highest scoring sentence t_j is selected as its alignment. The process then continues to align the next sentence s_{i+1} in d_1. This process is carried out recursively until all sentences in d_1 have been aligned. As mentioned in the previous

[2] I.e., d_1 has the same or fewer number of sentences than d_2.

[3] Jaccard similarity was also used to measure similarity between sentences in Adafre and de Rijke (2006).



...

Večinajih je v bližini [[**vesta family**]].

Imajo podobne [[**eccentricity**]], toda njihova [[**ellipsis**]] leži v območju od 2,18 [[**astronomical unit**]] do 2,50 a. e. (kjer je [[**Kirkwood gap**]] 3 : 1).

...

(a) SL article

...

A large proportion have orbital elements similar to those of 4 Vesta, either close enough to be part of the [[**vesta family**]], or having similar [[**eccentricity (orbit)**]] and [[**inclination**]]s but with a [[**semi-major axis**]] lying between about 2.18[[**astronomical unit**]] and the 3:1 [[**kirkwood gap**]] at 2.50 AU.

...

(b) EN article

Fig. 6.5 Example of SL-EN sentences paired by the *anchor* + *word* method

section, many-to-one correspondences between sentences are allowed; an example of this is shown in Figure 6.5.

The author implemented a minimum similarity threshold to filter out irrelevant sentence pairs. If the score of the sentence pair is below the minimum threshold, the pairing information between both sentences is discarded. In this experiment, the author used a minimum threshold of 0.1, which was empirically determined by manually evaluating the similarity of sentence pairs in the evaluation corpus scored by this method.

A maximum threshold is also used in this method to reduce the noise caused by document pairs containing the same contents. As discussed in Chapter 5, although these duplicate information may be perceived to be similar, they do not represent high cross-lingual similarity, nor contain valuable cross-lingual resources because the contents were the same in both languages. In this study, the author used a maximum threshold of 1.0, i.e., exact sentences are discarded in this study. The remaining sentence pairs are then used to measure the similarity of the document pair at the document level, described in the next section.

Similarity at the document level

In the second stage, the scores of the remaining aligned sentence pairs are aggregated to represent the similarity of the document pair at the document level ($docSimilarityScore$). This section describes how this method was created.

Firstly, a document pair (d_1, d_2) containing sentence pairs with higher aligment scores are considered to be more relevant than a document pair (d_3, d_4) containing the same number of sentence pairs with lower scores. Therefore, similarity can first be identified by aggregating the alignment scores of the aligned sentences ($sentAlignScore$). This is referred to as the $totalSentAlignScores$.

Secondly, a method is required to normalise the $totalSentAlignScores$ as Wikipedia article lengths may vary significantly. Normalisation is often performed by taking account the lengths of both documents. However, as shown in Section 5.3, Wikipedia article may differ in length. Furthermore, the shorter article may still contain content that strongly corresponds (i.e., content that is either highly similar or in a translation relation) to a sub-content of the larger article. A normalisation using the length of the larger document or a combination of the two will punish these articles. Therefore, in this experiment, the sentence alignment scores is normalised using the length of the shorter article instead.

The algorithm to measure the document similarity in this experiment is shown in the following:

$$docSimilarityScore = \frac{totalSentAlignScores}{n} = \frac{\sum_{i=1}^{n} sentAlignScore_i}{n} \tag{6.3}$$

where $sentAlignScore_i$ represents the sentence alignment score for a sentence s_i in the shorter document (or 0 if the sentence is unpaired or has its alignment filtered out), and n represents the number of sentences in the shorter document.

6.3 Experiments

6.3.1 Language selection

In this experiment, the $anchor + word$ method was used to measure similarity on eight language pairs: German (DE), Greek (EL), Estonian (ET), Croatian (HR), Lithuanian (LT), Latvian (LV), Romanian (RO) and Slovenian (SL); all were paired to English (EN). In the remainder of this chapter, the non-English languages are referred to as the 'source' languages, and English is referred to as the 'target' language.

6.3.2 Corpus

This experiment utilised the Wikipedia corpus gathered in November 2009-March 2010. More information about this corpus is described in Section 3.6. Table 6.2 shows the number of interlanguage-linked articles for the eight language pairs used in this study.[4]

Although extracted in the similar time period, the numbers of interlanguage-linked articles available in the corpus were extremely different between each language pair. The smallest language pair, LV-EN, has just above 21,000 pairs of interlanguage-linked articles, whilst the largest language pair, DE-EN, contains almost 30 times more document pairs, with 637,382 pairs of interlanguage-linked articles.

[4] The number of interlanguage-linked article pairs in each language pair was previously shown in Table 3.2 in page 82.

Table 6.2 Comparison of sizes across language pairs

Language pair	Total interlanguage-linked articles	Proportion of same titles	Bilingual lexicon size
DE-EN	637,382	72%	181,408
EL-EN	36,752	23%	28,294
ET-EN	42,008	46%	22,645
HR-EN	51,432	48%	26,804
LT-EN	57,954	28%	41,497
LV-EN	21,302	27%	15,511
RO-EN	97,815	63%	35,774
SL-EN	51,332	51%	25,101

The author also reports the proportion of interlanguage-linked articles that contain the same (duplicate) titles. These duplicate titles were removed prior to creating the bilingual dictionaries. The size of the resulting bilingual dictionaries are also shown in Table 6.2.

6.3.3 Evaluation

The $anchor + word$ method was evaluated using two approaches. The first approach compared the $anchor + word$ method to a similar approach utilising a machine translation system. The second approach evaluated the performance of the $anchor + word$ method against a gold standard corpus. These approaches are described below.

The $anchor+word$ method relies only on a bilingual lexicon extracted from Wikipedia to identify similarity across languages. In the first evaluation, the author analysed how well this approach performed if better translation resources were used. To investigate this, the author developed a similar method that utilised Statistical Machine Translation (SMT) to perform the translation (instead of using Wikipedia as a translation resource). This method is referred to as *the $translation$ method*. In this method, Google Translate[5] was used to translate the 800 non-English documents in the evaluation corpus (Chapter 5) into English. After all non-English documents were translated, the similarity score of each document pair was calculated using the similarity identification method described in Section 6.2.4. Similar to the $anchor + word$ method, similar sentences were aligned, and their scores were aggregated to represent the similarity scores of the document pair. The correlation between the two approaches were evaluated using Spearman's ρ (shown in Section 6.4.1).

In the second evaluation, the author evaluated how well the $anchor + word$ method performs against the gold-standard (i.e., the evaluation corpus). As previously described in Chapter 5, each of the 800 document pairs was assessed by two assessors. Q1 scores in the corpus contain the assessors' answers for the following question: "How similar are the two documents?", specified in a 5-point Likert Scale. For this analysis, the Q1 scores given

[5] http://translate.google.com

by both assessors were averaged and used to represent the human-annotated score. Spearman's ρ was then calculated between these human-annotated scores (average Q1 scores) to scores given by the $anchor + word$ method. As a comparison, the correlation between the gold-standard and the $translation$ method is also reported in this evaluation (shown in Section 6.4.2).

6.4 Results

6.4.1 Correlation between the automatic methods

The results show that the similarity scores given by the $anchor + word$ method and the $translation$ method (described in Section 6.3.3) correlated strongly with each other (ρ=0.717, p<0.01). A scatter plot showing the correlation between the two approaches for all document pairs is shown in Figure 6.6. The correlation between the two approaches for each language pair was also evaluated. The results are shown in Table 6.3 and the scatter plots are shown in Figure 6.7. The results show that the correlation between language pairs vary widely; the highest correlation (ρ=0.897) was achieved in German-English, whilst the $anchor + word$ method correlates the least with the MT approach in Greek-English (ρ=0.441).

Table 6.3 Correlation between automatic methods

Language pair	**Correlation between automatic methods**
DE-EN	0.897
EL-EN	0.441
ET-EN	0.741
HR-EN	0.683
LT-EN	0.791
LV-EN	0.593
RO-EN	0.680
SL-EN	0.576

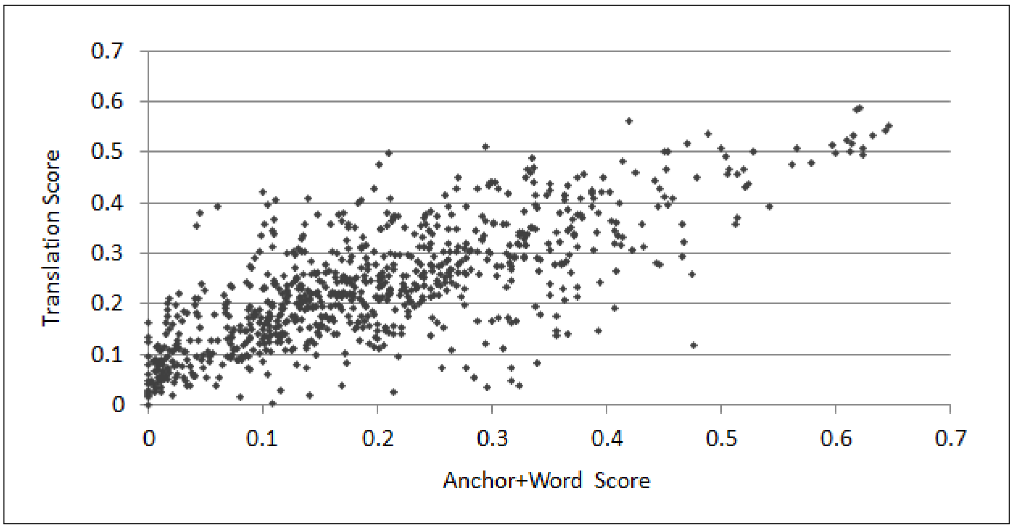


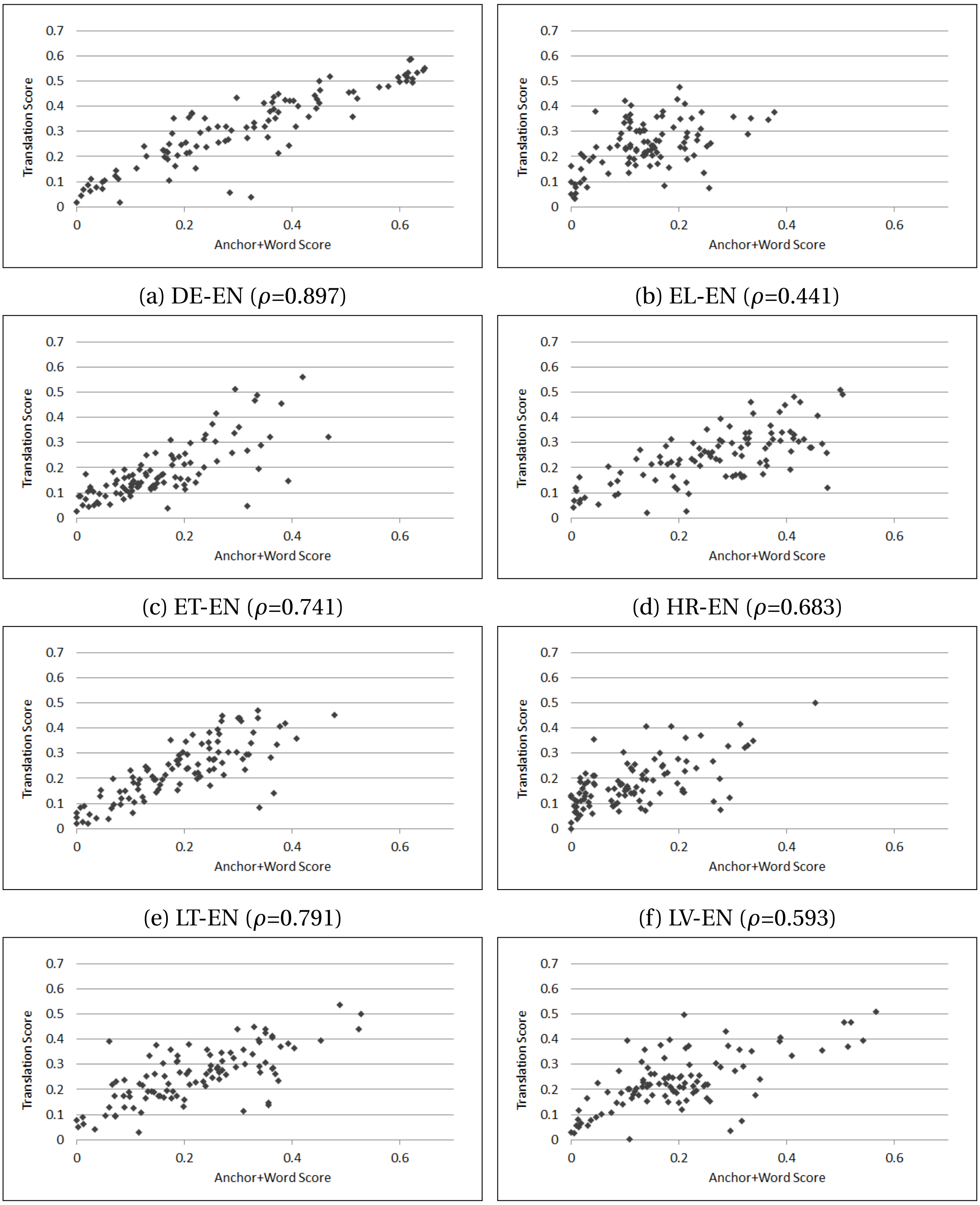
Fig. 6.6 Correlation between $anchor + word$ method and translation method (all language pairs, N=800)

6.4.2 Correlation to human-judgments

The results in Table 6.4 shows that, although both automatic approaches showed a strong correlation with each other as previously discussed in Section 6.4.1, the correlation to human judgments are somewhat lower. In general, the $translation$ method unsurprisingly achieved a higher correlation to human judgments than the $anchor + word$ method, ρ=0.481 and ρ=0.374, respectively.

Furthermore, the $translation$ method correlates better with human judgments in 6 language pairs (DE-EN, EL-EN, ET-EN, LT-EN, LV-EN and RO-EN), compared to the $anchor + word$ method. This suggests that the use of better translation resources can improve the accuracy of the method. This improvement, however, vary widely between these language pairs. Using the $translation$ method, correlation in EL-EN improved by 84.3%. RO-EN, meanwhile, only achieves a correlation score that is 0.7% higher than using the $anchor + word$ method. The remaining four language pairs achieved between 12.6%-39.7% increase in correlation scores compared to $anchor + word$ method.

Meanwhile, $anchor + word$ method is shown to be more superior in two language pairs, HR-EN and SL-EN. This is a positive result given that the result is obtained by making use of only a bilingual lexicon derived from Wikipedia.



(a) DE-EN (ρ=0.897)

(b) EL-EN (ρ=0.441)

(c) ET-EN (ρ=0.741)

(d) HR-EN (ρ=0.683)

(e) LT-EN (ρ=0.791)

(f) LV-EN (ρ=0.593)

(g) RO-EN (ρ=0.680)

(h) SL-EN (ρ=0.576)

Fig. 6.7 Correlation between $anchor + word$ method and translation method for each language pair (N=100)

Table 6.4 Correlation (Spearman Rank, ρ) between human judgments and similarity measures for 5 classes and across languages

Language	**Correlation with human judgments**	
pair	**Anchor+word**	**Translation**
All	0.374	**0.481** (⇑28.6%)
DE-EN	0.595	**0.670** (⇑12.6%)
EL-EN	0.261	**0.481** (⇑84.3%)
ET-EN	0.556	**0.687** (⇑23.6%)
HR-EN	**0.475**	0.415 (⇓12.6%)
LT-EN	0.365	**0.507** (⇑38.9%)
LV-EN	0.348	**0.486** (⇑39.7%)
RO-EN	0.301	**0.303** (⇑0.7%)
SL-EN	**0.520**	0.447 (⇓14.0%)

6.5 Discussion

6.5.1 Correlation between automatic methods

Results discussed in Section 6.4 show that the $anchor + word$ method correlates highly to the $translation$ method (ρ=0.717). Considering that the $translation$ method represents state-of-the-art translation resources, this finding is very promising. On the other hand, the translation qualities for under-resourced languages often vary widely due to the lack of bilingual resources to train the machine translation system (Skadiņa et al., 2012). If this was the case for this study, the high correlation scores cannot be used to determine the performance of $anchor + word$ method in general.

To investigate this in more detail, the author examined the translation qualities of a set of documents in the evaluation set. The author found that the quality of Google Translate for the under-resourced language pairs used at the time of the study varied widely.[6] In general, the translation quality of the under-resourced language pairs is poorer than the highly-resourced language pair, which confirmed the findings from previous literature (Skadiņa et al., 2012). As an example, an excerpt of a translated Estonian article in the evaluation corpus about the "Estonian Auxiliary Police" is shown in Figure 6.8. This

[6] The evaluation corpus was translated using Google Translate in 2010. The qualities of both Google Translate and Wikipedia have improved significantly after this experiment was completed.

> ... In Estonia, the Estonian Selbstschutz members was also politseipataljonid, välipolitseinikest, Punaarmeest and ületulnutest and mobiliseeritutest aastavahetusel 1943 - 1944, formed politsepataljonideks renamed items, original name was vahipataljon protection. ... Politseipataljonid formeeriti various tasks (also known as valvepolitsei (Schutzpolizei) watchkeeping duty, rannakaitse, fighting fronts partisanidega, etc.), hence the differences in relvaüksustel: Politseirügement set up named, Schutzmannschaft, protection, Police, Schutzmannschaft Vahipataljon infantry battalion. ...

Fig. 6.8 An excerpt of the English translation (using Google Translate) of an Estonian (ET) article about the "Estonian Auxiliary Police"

example shows that many words (often domain-specific terms) were left untranslated when using the $translation$ method. Further work is needed to analyse the translation quality in these language pairs in more detail. This task was not carried out in this study as it required linguistic knowledge of each language pair, which could not be pursued in the limited time of the study.

Although the qualities of Google Translate varied widely across languages, Google Translate was, at the time of the study, a state-of-the-art translation resource and was a valid baseline to use against the proposed method. Furthermore, it specifically highlighted the challenges for under-resourced languages. The high correlation between the $anchor + word$ method and the $translation$ method shows that the use of Wikipedia as a bilingual resource for under-resourced languages is very promising. This finding also confirms that the $anchor + word$ method can be used without a significant decrease in quality compared to the state-of-the-art translation method in the language pairs.

Variations between language pairs

Although the overall correlation between the $anchor+word$ method and the $translation$ methods is high, the correlation scores between these two automatic methods vary widely across the different language pairs. The highest correlation (ρ=0.897), achieved in German-English, is more than double the lowest correlation, i.e. ρ=0.441 in EL-EN. These varying degrees of correlation across languages can be explained using the following reasons.

Firstly, the $anchor + word$ method fully relies on *the size of translation resources* extracted from Wikipedia to identify overlapping information in different languages relies. The size of these resources for each language pair is the number of interlanguage-linked articles in the language pair (see Table 6.2).[7] The more interlanguage-linked articles are available in the language pair, the larger the translation resource is for the language pair, and the more likely that a word in the source language can be translated into English.

In the study, the results indicate that the higher number of interlanguage-linked articles ("Total ILL articles") in the language pair is, the higher is the correlation between the automatic methods (i.e., the correlation between $anchor + word$ method to the $translation$ method). LV-EN and EL-EN, which have the two lowest number of interlanguage-linked articles, also have lower correlation scores compared to other language pairs. DE-EN, on the other hand, has a very high number of interlanguage-linked articles and a high correlation scores between both automatic methods.

Secondly, the performance of $anchor + word$ method may also be influenced by *the similarity of the languages*. Because the $anchor + word$ method computes the overlap of words in the article content, it is likely to perform better on source languages that are similar to English. In other words, languages that share more English words can be measured more accurately using the $anchor + word$ method, compared to those that share fewer words.

Identifying similarity across languages requires specific linguistic knowledge of all language pairs explored in this study. Since this information was not available, the number of duplicate titles in both languages are used to indicate the similarity between the languages. The proportion of the same titles for each language pair is shown in Table 6.2. These proportions were calculated using the list of ILL articles only, and therefore, only represents the proportion of duplicate titles in a small subset of the vocabulary of the language pair. However, due to the large dataset (over 20,000 titles for each language pair),

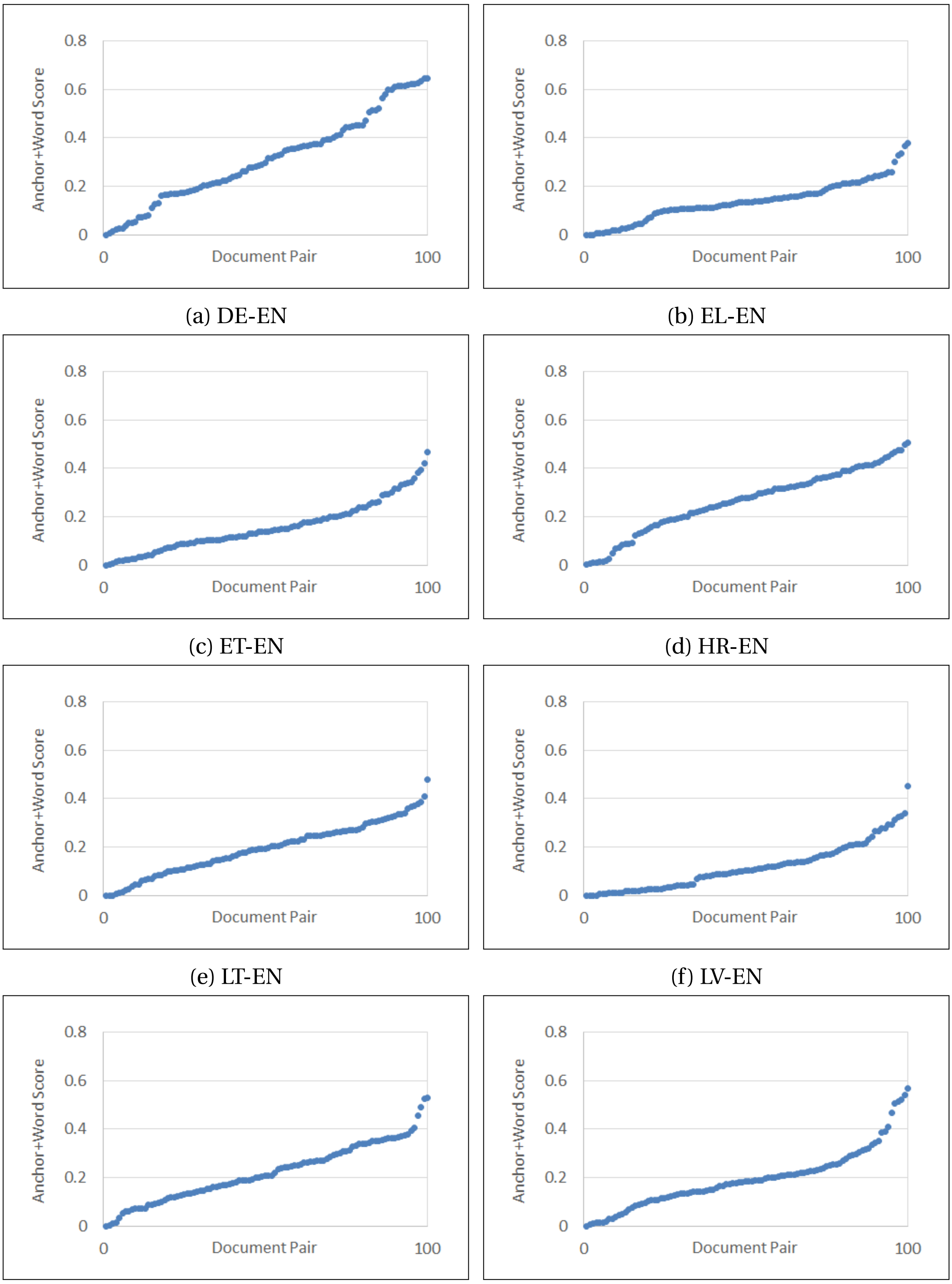
[7] As described in Section 3.6.1, duplicate titles of interlanguage-linked articles were filtered out prior to creating the bilingual lexicon. However, these duplicate titles are also valuable in identifying overlapping words in both languages. Therefore, this analysis reports the correlation between the number of interlanguage-linked articles (instead of the bilingual lexicon size) in each language pair to its correlation scores between the automatic methods.

these findings can still be used to indicate the degree of similarity of the language pairs.

Using these data, German-English was shown to achieve the highest proportion of same titles in their interlanguage-linked articles (72% of same titles). It is therefore indicated to be more likely to share other words with English in general, compared to Latvian, or Lithuanian, which share 27% and 28% same words, respectively. Unsurprisingly, Greek has the lowest proportion of same titles, 23%, undoubtedly affected by the different characters they used compared to English. The results indicate that higher correlations seem to be achieved in languages that are indicated to be similar (i.e., higher proportion of same titles), such as German-English. Greek-English, meanwhile, was shown to be the least similar languages and achieved the lowest correlation scores. This is, however, not strongly supported by the rest of the language pairs, suggesting that other factors may also affect these results.

6.5.2 Correlation to human judgments

The results presented in Section 6.4 show that the $anchor + word$ method achieved a low correlation (ρ=0.374) to human judgments. The author further analysed whether the proportion of same words in each language pair correlates with the performance of $anchor + word$ method, in the context of the correlation scores to human judgments. The data indicate that the more similar a language pair is (i.e., the higher proportion of same words that language pair has), the better $anchor + word$ method is at measuring similarity in the given language pair. The three language pairs where the $translation$ method significantly outperformed $anchor + word$ method (EL-EN, LV-EN, and LT-EN) (previously shown in Table 6.4) were found to also have the lowest proportion of same words. There is a strong negative correlation between the proportion of same words and the translation improvement (ρ=-0.72, p<0.05), suggesting that the more similar a language is, the more likely it is for $anchor + word$ method to achieve a similar performance to the $translation$ method.



(a) DE-EN

(b) EL-EN

(c) ET-EN

(d) HR-EN

(e) LT-EN

(f) LV-EN

(g) RO-EN

(h) SL-EN

Fig. 6.9 Distribution of $anchor + word$ scores across the 100 evaluation document pairs per language pair

Variations between language pairs

The author further investigated the performance of $anchor + word$ method in the different language pairs. First of all, the distribution of $anchor + word$ score for each evaluation document pair in each language pair is shown in Figure 6.9. The figure shows that most documents have linear scores with some differences in the higher score document pairs. The author further analysed the correlation between $anchor + word$ method and the human judgments in groups of documents in different scores.

Firstly, the author divided each evaluation set (i.e. each language pair) into 3 similar-size bins; 33 document pairs with the lowest $anchor + word$ score are included in the first bin, 33 of those with the highest $anchor + word$ score are included in the third bin, and the remaining 34 in between are included in the second bin. For each bin, the judgment scores given by the assessors were averaged. The results, shown in Figure 6.10, indicates that bins with document pairs that were scored lower by $anchor + word$ score also lower average judgment scores.

In the second analysis, the 3 bins are decided based on the $anchor + word$ method's score range in that language pair, i.e., each bin has the same range of scores. Given an evaluation set with a minimum score of i and a max score of j, the range of scores con-

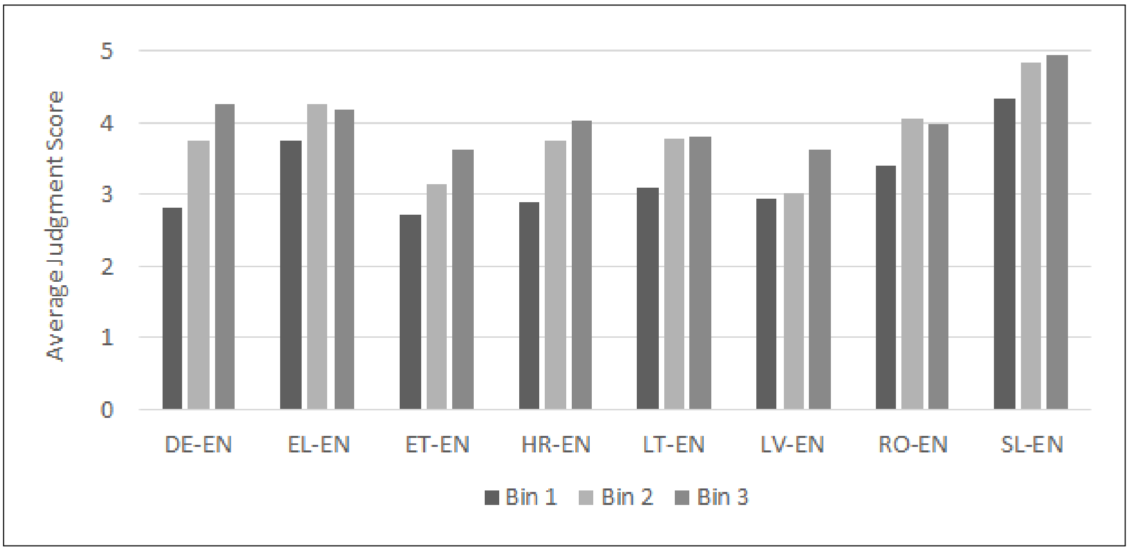


Fig. 6.10 Average judgment scores for document pairs in different bins (same size bins)

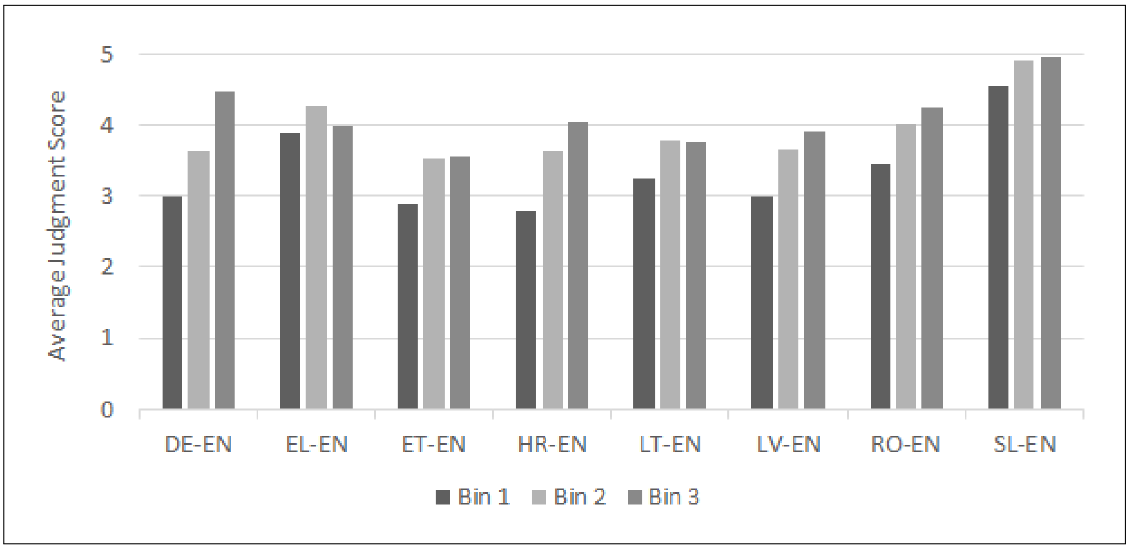


Fig. 6.11 Average judgment scores for document pairs in different bins (same score range)

tained in each bin is $(j-i)/3$. For example, if an evaluation set contains document pairs scored between 0 and 0.9, document pairs score of 0.3 or below are contained in bin 1, bin 2 contains document pairs above 0.3 and below 0.6, and the remaining document pairs (above 0.6) are included in bin 3. In this case, the number of document pairs in each bin may be different to each other. However, this analysis will show the reliability of $anchor + word$ method on the different score range. These results are shown in Figure 6.11. Similar to Figure 6.10, when dividing the evaluation set into documents with three different score bins, Figure 6.11 also shows that document pairs with higher scores are more likely to have higher similarity scores assigned by the assessors.

6.5.3 Failure analysis

Previous analyses have shown that $anchor + word$ method can be used, to some extent, to identify document pairs with different similarity. However, the low correlation scores between the $anchor + word$ method and human judgments also indicates that there are aspects of similarity not currently captured by the methods. The author inspected this further by performing a failure analysis on two cases: i) document pairs which were scored highly by the $anchor + word$ method yet were assessed to have low similarity by the assessors, and ii) document pairs which that had low scores given by the $anchor + word$ method but were manually assessed to be highly similar. The author analysed these documents and summarised the disagreement cases as shown in Table 6.5.

The first type of disagreement cases is caused by *documents having different lengths*, especially when the entire content of the smaller documents is highly similar to a subset of the longer documents. The $anchor + word$ method regarded these overlapping contents to be very useful for linguistic purposes and therefore assigned the document pairs to be highly similar. Extra information contained in the larger document does not in any way affect the usefulness of these overlapping fragments and therefore is regarded to be irrelevant when computing similarity between the document pair. On the other hand, assessors considered that the longer document contained information not available in the shorter document and hence they reflected this by assigning a lower similarity score.

Whilst the previous case describes examples where document pairs identified to be highly similar by the $anchor + word$ method were judged to be less similar by assessors, other disagreement cases occurred where document pairs not identified to be similar by the automatic methods were considered to be similar by the assessors. These disagreements appeared when assessors considered *a number of similarity aspects* that are currently not captured in the automatic method, such as the structure of the articles, or the languages that the contents were written in. The latter case caused disagreements when the article contents were written in the same language, e.g., both the Romanian[8] and English[9] Wikipedia articles of "List of American philosophers" contain contents written in

[8] http://ro.wikipedia.org/wiki/Lista_filozofilor_americani (accessed on 21 January 2014)

[9] http://en.wikipedia.org/wiki/List_of_American_philosophers (accessed on 21 January 2014)

Table 6.5 Disagreement cases between assessors and *anchor+word* method

Assigned similarity score		**Description**
Anchor+word	**Assessors**	
High	Low	Different document lengths
Low	High	Assessment of different aspect of similarity
Low	High	Lack of overlapping anchors and cognates

English. As discussed in Section 6.2, due to $anchor + word$ method developed specifically to identify *cross-lingual similarity*, it did not consider information written in the same language to be similar and therefore assigned these documents a low score. The assessors, on the other hand, simply identified that these documents had similar contents and disregarded the fact that the information was not cross-lingual, and assessed the documents to be highly similar.

Disagreement also appeared in cases where articles were genuinely similar but were not calculated properly using the language-independent method due to the *lack of overlapping words*. While this feature is considered to be useful for languages with a similar written form to English (such as German), for other languages (e.g., Greek), the alphabets are very different and subsequently the number of matching words drops significantly. This then causes the approach to rely on the availability of links; however, there are similar documents which simply do not contain enough links for the $anchor + word$ method to identify parallel sentences accurately.

6.6 Conclusion

In this experiment, the author proposed the $anchor + word$ method, which is an adaptation of the link-based bilingual lexicon approach (Adafre & de Rijke, 2006). This approach utilised the interlanguage links in Wikipedia to build resources (bilingual lexicons) to identify similarity across languages. Similarity is measured by translating anchor texts in the source languages, aligning similar sentences between the document pair (i.e., those that share high overlap of contents), and aggregating the scores at the document level. In this section, the author answered the research questions presented earlier in this chapter.

RQ3. Can language-independent approaches be used to identify cross-lingual similarity in Wikipedia? This experiment showed that the $anchor + word$ method was able to indicate the degree of similarity of a document pair. Considering that it does not require any translation resources except Wikipedia interlanguage links, the results have shown to be promising. However, it achieved a weak correlation to human judgments

(ρ=0.374), which indicated that there were characteristics of similarity that could not be captured using this method.

(a) How does the method compare to approaches using linguistic resources, such as MT systems? The $anchor + word$ method achieved a high correlation to the state-of-the-art translation method (ρ=0.717), which indicated that the language-independent approach was able to measure similarity as well as approaches that utilised a much more advanced state-of-the-art translation system.

(b) How does the performance for the approach vary for different language pairs? The performance of the $anchor + word$ method varies widely across the different language pairs. EL-EN, which has one of the smallest number of interlanguage-linked articles in the dataset (36,752 pairs), achieved the lowest correlation to human judgments (ρ=0.261) and DE-EN, which has the highest number of interlanguage-linked articles (637,382 pairs), achieved the highest correlation (ρ=0.595). These results indicate that the $anchor + word$ method relied on the size of the interlanguage-linked articles available for identifying the overlapping information across languages.

(c) What language-independent features are best for measuring cross-lingual similarity in Wikipedia? The $anchor + word$ method used two features: link overlap and word overlap in identifying similar sentences, prior to aggregating this information at the document level. Although they were able to measure similarity to some extent, the failure analysis showed that there were other characteristics that influenced annotators' scores that were not captured in this method, such as similarity of the structures, the languages of the contents, and the length of the articles. Future work should explore approaches that incorporate these features when measuring similarity.

This experiment shows that interlanguage-links information from Wikipedia can be utilised for identifying cross-lingual similarity with a comparable performance to an approach utilising a state-of-the-art translation systems. Furthermore, this approach can be applied to any languages in Wikipedia if they have the interlanguage-links information. This finding is considered to be very promising and is further explored in the next chapters of this work.

Related publications

- Paramita, M., Clough, P., Aker, A. and Gaizauskas, R. 2012. Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, pp. 790-797.

- Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M., Pinnis, M. 2012. Collecting and Using Comparable Corpora for Statistical Machine Translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, pp. 438-445.

Chapter 7

Content Similarity Features

The $anchor + word$ method described in Chapter 6 shows that Wikipedia interlanguage links information is a promising feature that can be used to measure similar information across languages. However, the results also show that there are other aspects of document similarity that are currently not captured using the overlap of links utilised in the $anchor + word$ method. In this chapter, the author reports the second experiment in this study, in which a number of features are investigated for measuring the similarity of Wikipedia document pairs. These features were derived from the findings in the evaluation corpus which listed the document characteristics that contributed to the similarity of Wikipedia interlanguage-linked articles. This experiment investigates the performance of each individual feature and a combination of multiple features for identifying similarity in Wikipedia.

7.1 Background

Language-independent features, such as word length, have been shown to be a promising feature for identifying translated sentences (Munteanu & Marcu, 2005; Patry & Langlais, 2011) and parallel documents (Resnik & Smith, 2003). Another feature, character-n-gram overlap has also been shown to be valuable for identifying cross-lingual words (McNamee & Mayfield, 2004). The use of these features in identifying similar, yet non-parallel texts, however, has not been investigated. Specific Wikipedia features, such as the link overlap, has also been utilised in previous work to identify similarity at the sentence level (Adafre & de Rijke, 2006). Utilising this feature and a word overlap in measuring similarity (i.e., the $anchor + word$ method described in Chapter 6), however, was shown to achieve a low correlation to human judgments. This suggests that there are other aspects of similarity that were not captured using this approach.

To identify a more suitable approach, the similarity characteristics that have been identified in the evaluation corpus (Chapter 5) were analysed in this study. A number of features that can be extracted automatically to capture each characteristic were identified. The author evaluated the performance of these features, both individual features and a combination of multiple features, in computing similarity of Wikipedia articles. This experiment aims to answer the following research questions:

RQ3. Can language-independent approaches be used to identify cross-lingual similarity in Wikipedia?

(a) How does the method compare to approaches using linguistic resources, such as MT systems?

(b) How does the performance for the approach vary for different language pairs?

(c) What language-independent features are best for measuring cross-lingual similarity in Wikipedia?

This chapter is structured as follows. Firstly, similarity characteristics that contributed to the similarity of documents were identified in Section 7.2. The author identifies a number of features that can be automatically extracted to capture these characteristics in Section 7.3. The experiment setup is described in Section 7.4 and the results are reported and discussed in Section 7.5. This experiment is concluded in Section 7.6.

7.2 Features to capture similarity characteristics

The evaluation corpus (previously described in Chapter 5) contained 800 document pairs that were annotated by two assessors each, to identify its similarity score and the similarity characteristics that contributed to the specified similarity score. Five main characteristics were identified which indicated whether the document pair contained similar structure, overlapping named entities (NEs), overlapping fragments, translated contents, or different information. These characteristics were aggregated between the two assessors to represent the document pair. The similarity score of the document pair is the mean of similarity scores given by the two assessors. These results, shown in Figure 7.1, identified the characteristics of document pairs that contributed to a specific degree of similarity for a document pair.

The results show that document pairs with high similarity scores were annotated to exhibit similar structure, contain overlapping named entities, fragments, or translated contents. On the other hand, document pairs that were assessed to have low similarity scores were often annotated to contain different information by the assessors. Only a small proportion of these documents were annotated to contain translated contents, similar structure and overlapping fragments.

These findings indicate that features that can be extracted automatically to identify these characteristics may be useful to measure the similarity at the document level. In this chapter, the author identified a list of features that can be used to measure each similarity characteristic as summarised in Table 7.1.

7.2.1 Characteristic: 'Similar structure'

Figure 7.1 shows that almost all document pairs assigned similarity scores between 4 and 5 were identified to also contain *similar structure.* In comparison, fewer than 30% document pairs contain this characteristic when the document pairs were given an average score of 2 or lower.

Structural similarity has previously been indicated as a relevant feature to identify

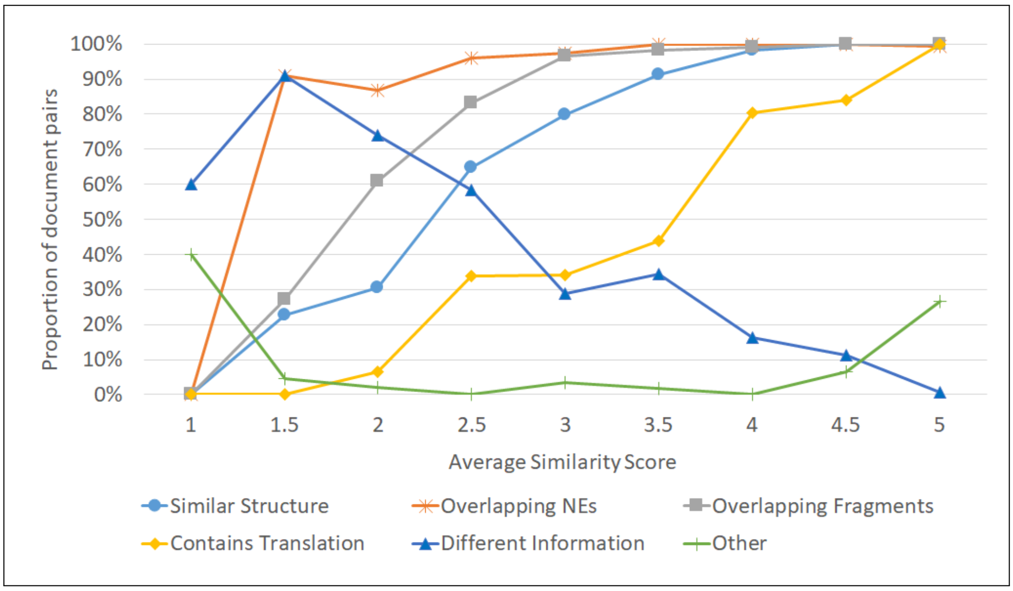


Fig. 7.1 Characteristics of document pairs with different similarity scores

parallel (translated) articles. Resnik and Smith (2003) previously indicated that features, such as HTML tags and their order of appearances in a document pair, can be used to identify whether or not the document pair is a translation of each other. These features, however, are not relevant for identifying similar (but non-parallel) documents, such as Wikipedia. Instead, 'similar structure' in Wikipedia can be represented as articles that discuss similar aspects of a topic.

The findings in Section 4.4.1 indicated that section headings in Wikipedia documents can be used to inform the aspects of the topics that are covered in the documents. Two Wikipedia articles that contain similar section headings are likely to discuss similar aspects, compared to those that contain different section headings. Therefore, features that measure the similarity between these section headings, such as the *overlapping sections* (i.e., the number of similar sections) and the *ratio of section lengths*, can be used to identify the structural similarity between these documents.

Furthermore, the findings in Chapter 5 have further shown that documents with different lengths were often punished (i.e., given a lower similarity score) by the annotators. Therefore, the *ratio of word length* can also be used to indicate the 'similar structure' characteristic.

7.2.2 Characteristic: 'Overlapping named entities'

Figure 7.1 also shows that over 95% document pairs that were assigned a mean score of 2.5 or above were indicated to have *'overlapping named entities'*. Named Entity Recognition tools, such as GATE (Cunningham, Maynard, Bontcheva, & Tablan, 2002), can be used to identify named entities in documents. These tools, however, require language-dependent resources which are not often available for under resourced languages. On the other hand, many named entities (such as persons, countries, organisations) are often included as concepts in Wikipedia, i.e., Wikipedia often contains pages describing these named entities. If the named entity is described in Wikipedia articles in more than one language, the titles of these articles can be extracted and utilised as a bilingual lexicon.

Furthermore, overlap of these named entities can be identified by utilising the bilingual lexicon to measure the *overlap of links*, as described in Chapter 6. In cases where entities are not included in the bilingual lexicon (i.e., for cases where the named entity is only described in one language), the author aims to identify the degree of overlapping named entities by measuring the overlap of characters between the texts, using *cognate overlap* (Simard, Foster, & Isabelle, 1993) and *character-n-gram overlap* (McNamee & Mayfield, 2004). More information on these features are described in Section 7.3.

7.2.3 Characteristic: 'Overlapping fragments'

The characteristic '*overlapping fragments*' was also shown to be an attribute of similar document pairs (those scoring 3 or above). To identy these overlapping fragments, the author again investigates the use of *overlap of links*. For languages that are morphologically related, similar words often have similar characters. Therefore, *cognates* and *character-n-gram overlap* are also investigated to identify the overlapping contents.

7.2.4 Characteristic: 'Contain translation'

To identify this characteristic, one requires a language-dependent resource, such as a machine translation system, to translate documents into the same language prior to finding similar contents between them. Translation can then be identified by finding an alignment of a number of words (Munteanu & Marcu, 2005). Identifying translated content is not easily achieved using language-independent methods. Translated texts, however, will also contain overlapping named entities and fragments which can be identified using the features described in the previous sections, i.e., *cognate overlap*, *character-n-gram overlap* and *overlap of links*.

7.2.5 Characteristic: 'Different information'

The characteristic 'different information' was frequently used to describe document pairs that are not similar (between 60-80% for document pairs scoring 2 or lower). In this work, the author also utilised the features specified above, i.e., overlapping links, cognate overlap, character-n-gram overlap and links overlap can be used to identify documents containing different information. More specifically, the lack of the proportion of these features may indicate that the document pairs contains a high proportion of different information.

7.2.6 Characteristic: 'Other'

As previously described in Section 5.5.2, annotators were allowed to specify other characteristics of the assessed document pairs if they were not included in the list of five characteristics described above. Most of the feedback, however, were used to specify more information about the similarity of the document pairs, such as the texts containing a large number of overlapping named entities, but not overlapping fragments. Annotators also used this characteristic to inform any language issues (e.g. content was not written in the required language), or a content error. These reasons are not particularly relevant for identifying similarity and therefore are not investigated further in this chapter.

Table 7.1 Features to identify similarity

Characteristics	Features	Extraction Level	
		Article Contents	Section Headings
Similar structure	Overlapping sections Section length ratio Word count ratio	 ✓	✓ ✓
Overlapping named entities	Overlapping links	✓	
Overlapping fragments	Cognate overlap	✓	
Contain translation	Char-n-gram overlap	✓	
Different information			

7.3 Feature extraction

The six features previously listed in Table 7.1 can be categorised into two groups based on the contents that are utilised to extract them, either using the *section heading information* or the entire *article contents*.

Two features - *overlapping sections* and *section length ratio* - are extracted using the section headings information of the article. Section headings represent the titles of sections that are contained in the document. E.g., an article about a country such as "England" may contain the following section headings: "Toponomy", "Geography", "Economy", "Demography", etc. These two features can be extracted using the section headings information only, i.e., the article contents are not used in their extraction. These features also rely on the availability of section headings in both documents. In other words, each of these documents must contain at least one section heading in order for these features to be computed. Since not all documents in the evaluation corpus contain this information, these features are investigated independently using a smaller evaluation corpus, i.e., filtering out document pairs that do not have section headings information. This experiment is described in Chapter 8.

The remaining four features, i.e., *word count ratio, overlapping links, cognate overlap* and *char-n-gram overlap*, are extracted using the article contents. Since they measure the similarity between the article contents, these are referred to as 'content similarity features'. Given a document pair, d and d' written in language L and L' respectively, methods to extract these features are described below:

- **Cross-language character n-grams ($char-n-grams$)**: To calculate char-n-grams, a simplified alphabet $\Sigma = \{a, \ldots, z, 0, \ldots, 9\}$ is considered, i.e., any other symbol, space, and diacritic is discarded and case-folding applied. The text is then codified into a vector of character n-grams ($n = \{3, 5\}$). This model is an adaptation of McNamee and Mayfield (2004). Char-n-gram overlap is calculated using Jaccard similarity.

- **Cognateness (cog)**: This concept was proposed to identify parallel sentences (Simard et al., 1993). A token t forms a *cognateness* candidate if: (a) t contains at least one digit; (b) t contains only letters and $|t| \geq 4$; or (c) t is a single punctuation mark. t and t' are pseudo-cognates if both belong to (a) or (b) and are identical, or belong to (b) and share the same four leading characters. Hence, we characterise d and d' as follows: if t accomplishes (a), it is maintained verbatim, if it accomplishes (b) it is cut down to the first four characters. Case (c) is not applied since the comparison is at the article level. Case-folding and removal of diacritics are applied. Jaccard similarity is used to measure the cognate overlap.

- **Common outlinks (lnk)**: This measure is computed by analysing the overlap of links using the link-based bilingual lexicon approach (Adafre & de Rijke, 2006). First, titles of interlanguage-linked articles are extracted as a bilingual lexicon. The links in each document are extracted. Links in document d are translated into L', prior to measuring Jaccard similarity against links in document *d'*.

- **Word count ratio (wc)**: Findings in Chapter 4 and Chapter 6 have indicated that similar documents are also likely to be more similar in length. This measure is computed as the length ratio between the shorter and the longer document (in number of words). This feature has also been shown to be valuable in identifying parallel documents (Resnik & Smith, 2003).

- **Translation + monolingual analysis ($trans_n$)**: This feature is selected as a baseline. It is a language-dependent model, in which Google Translate is used to translate d into L', generating d_t, which are then compared against d' using a standard monolingual process. Uni-gram ($trans_1$) and bi-gram ($trans_2$) word overlap are used in this experiment.

7.4 Experiment setup

To evaluate the approach the existing Wikipedia similarity corpus containing 100 document pairs from seven language pairs was used (Chapter 5). This includes German (a *highly-resourced* language), and 6 *under-resourced* languages: Greek (EL), Estonian (ET), Croatian (HR), Lithuanian (LT), Latvian (LV) and Romanian (RO); all non-English documents are paired to English (EN). The EL corpus was transliterated prior to calculating char-n-grams and cognate features. The Slovenian-English (SL-EN) dataset was filtered out in this experiment, due to the imbalance number of similar and non-similar document pairs in the set; 96 document pairs with score of 3.5 or above, and only 4 document pairs with average score of 3 or lower.[1] Each document pair was assessed by two assessors who assigned a similarity score (Q1) using a 5-point Likert scale. In this study, the mean scores are used to represent the document pair.

7.5 Results and discussion

The Spearman-rank correlation scores of each feature is shown in Figure 7.2. The results show that the language-independent features such as char-n-grams ('$c2g$', '$c3g$' and '$c4g$') were able to identify cross-lingual similarity with performance comparable to the baseline translation models using bi-gram overlap ('$trans2$'). Meanwhile, link overlap ('lnk') was shown to correlate least to human judgments. Interestingly, the results show that a simplistic language-independent model based on the word count ratio ('wc') correlates higher with human judgments compared to models using MT. This suggests that

[1] Please see Figure 5.6 in page 114 for further information.

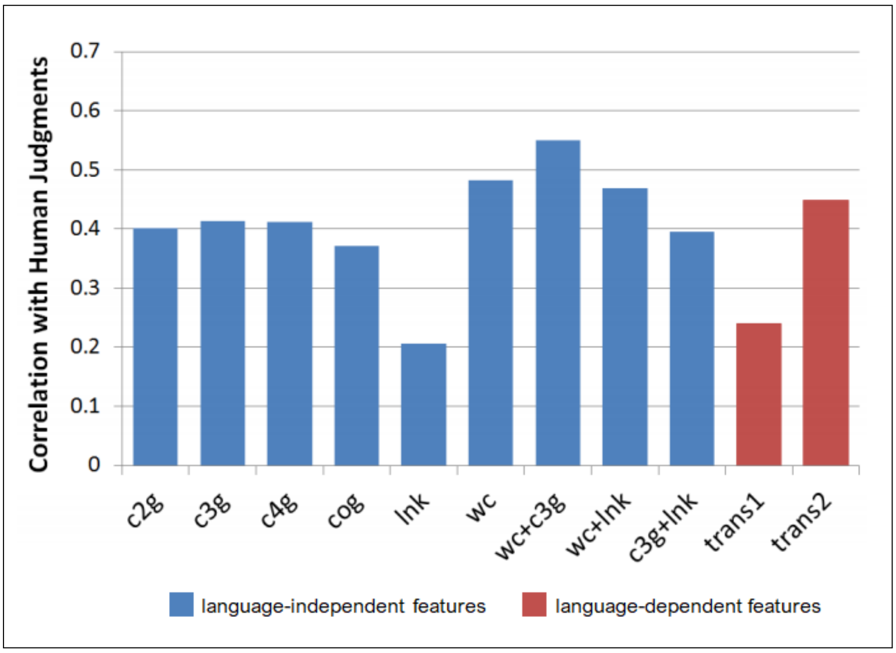


Fig. 7.2 Correlation of different models and human judgments (N=700)

interlanguage-linked Wikipedia articles with similar lengths are likely to contain similar contents.

A combination of features are also investigated in this experiment. The top three highest performing combination are shown in Figure 7.2. The highest correlation to human judgment is achieved by combining both the word length ratio and the char-3-gram overlap, '$c3g + wc$'. These findings are promising considering that these features are purely language-independent and can be calculated for many language pairs.

Table 7.2 Average correlation scores across document pairs for each language pair

Language Pair	**Lang-Independent Features**						**Lang-Dependent Features**	
	c2g	**c3g**	**c4g**	**cog**	**lnk**	**wc**	**trans1**	**trans2**
DE-EN	0.52	0.54	**0.55**	0.43	0.10	0.46	0.11	0.47
EL-EN	0.21	0.32	0.37	0.35	0.26	**0.50**	0.18	0.42
ET-EN	0.57	**0.59**	0.58	**0.59**	0.34	0.43	0.05	0.33
HR-EN	0.47	0.44	0.42	0.37	0.36	**0.56**	0.16	0.48
LT-EN	0.31	0.38	**0.39**	0.29	0.27	0.32	0.26	0.38
LV-EN	0.41	0.42	**0.44**	0.43	0.21	0.34	0.26	0.37
RO-EN	0.11	0.17	0.18	0.08	0.38	**0.45**	0.32	0.38
Mean	0.37	0.41	0.42	0.36	0.27	**0.44**	0.19	0.40

Table 7.3 Average correlation scores across document pairs for each language pair (combination of features)

Language Pair	**Lang-Independent Features**			**Lang-Dependent Features**	
	wc+c3g	**wc+lnk**	**c3g+lnk**	**trans1**	**trans2**
DE-EN	**0.63**	0.41	0.50	0.11	0.47
EL -EN	**0.54**	0.50	0.38	0.18	0.42
ET -EN	0.60	0.52	**0.65**	0.05	0.33
HR-EN	**0.61**	0.59	0.48	0.16	0.48
LT-EN	**0.46**	0.42	0.44	0.26	0.38
LV-EN	**0.47**	0.40	0.36	0.26	0.37
RO-EN	0.45	**0.57**	0.42	0.32	0.38
Mean	**0.54**	0.49	0.46	0.19	0.40

Whilst language-independent models perform well overall, their performance may differ for each language pair. Therefore, correlations in each language pair were also computed. Table 7.2 shows that whilst char-n-grams perform well on average, their correlations vary widely across different languages. The simplified outlinks (‘lnk’) model was less reliable in identifying similarity. Word length ratio, ‘wc’ is shown to be a more robust measure that perform consistently across the different language pairs.

The results also show that the use of Google Translate in identifying similarity (‘$trans1$’ and ‘$trans2$’) does not achieve a consistent result across the different language pairs. This indicates that the quality of translation between these language pairs may also vary widely. Moreover, the results show a drastic increase in correlation when using word bigrams (‘$trans2$’) compared to unigram (‘$trans1$’) overlap. The poor performance for the latter may also be caused by the weighting strategy used, i.e. term frequency. A straightforward enhancement of this model would be to remove stopwords and apply $tf - idf$ weighting.

The performance of a combination of features is also investigated for each language pair. These results are shown in Table 7.3. The combination of ‘$wc + lnk$’ results in a more stable model that perform well for all languages. The performance is, however, still lower than the combination of ‘$wc + c3g$’. The latter achieved the highest correlation for five language pairs. Meanwhile ‘$wc + lnk$’ and ‘$c3g + lnk$’ perform better in Romanian–English and Estonian–English, respectively. These results are promising: by combining language-independent models, it is possible to reliably identify cross-lingual similarity in Wikipedia with better performance (i.e., correlation to human judgment) than using MT systems.

7.6 Conclusion

Using the evaluation corpus, the author has derived a number of features that capture the characteristics of similar and non-similar documents. The correlation between the feature values and the human-judged similarity scores were reported. The author answered the research questions presented earlier in this chapter.

RQ3. Can language-independent approaches be used to identify cross-lingual similarity in Wikipedia? This experiment reported the performance of 6 different features in identifying similarity in Wikipedia: char-n-gram overlap ($c2g$, $c3g$, $c4g$), cognate overlap (cog), link overlap (lnk) and word length ratio (wc). These features can be extracted without the use of any linguistic resources, except for Wikipedia interlanguage links for measuring the link overlap measure. Overall, the results were shown to be promising (further described below).

(a) How does the method compare to approaches using linguistic resources, such as MT systems? The results show that most language-independent features (except link overlap) achieved higher correlation to human judgment compared to a simple term-frequency overlap of an approach utilising Google Translate ($trans1$ and $trans2$).

(b) How does the performance for the approach vary for different language pairs? In general, the performances of char-n-gram overlap, cognate overlap, and link overlap features vary widely across the different language pairs. However, word length ratio, wc was shown to be the most consistent feature across the different language pairs.

(c) What language-independent features are best for measuring cross-lingual similarity in Wikipedia? The best individual features vary across language pairs and are achieved when using one of the following four features, '$c3g$', '$c4g$', 'cog', or 'wc'. A combination of char-n-gram overlap and word length ratio ('$c3g + wc$') is shown to further increase the correlation scores.

This experiment has analysed the usefulness of a number of features in measuring Wikipedia similarity. The results indicate that language-independent features can be used to measure cross-lingual similarity in Wikipedia, with higher effectiveness than using features that utilised MT system and a word-n-gram overlap. These findings are very promising as these features can be calculated without the need of any translation resources. This will enable these features to be applied for measuring cross-lingual similarity to any Wikipedia language pair; transliteration may be needed for languages that used non-Latin script.

Related publications

- Barrón-Cedeño, A., Paramita, M. L., Clough, P., and Rosso, P. 2014. A Comparison of Approaches for Measuring Cross-Lingual Similarity of Wikipedia Articles. In *Proceedings of the 36th European Conference on Information Retrieval (ECIR 2014)*, Amsterdam, 13-16 April 2014, pp. 424-429.

Chapter 8

Structure Similarity Features

The work presented in the previous chapters have focused on identifying similarity of articles using features extracted from the entire content of the article. However, previous analysis (Chapter 4) and findings from the evaluation corpus (described in Chapter 5) have indicated that human annotators have identified that more similar documents often contain a more similar structure. In this chapter, the author investigates an approach to measure the similarity of Wikipedia documents by analysing the similarity of the structures (i.e., the section headings) in the Wikipedia articles. Language resources derived from Wikipedia and Wiktionary are being utilised in this approach.

8.1 Background

Measuring structure similarity between documents have been explored in previous literature. Resnik and Smith (2003) indicated that structural features, such as the order of HTML tags occurrences, can be used to identify articles that are parallel (translations of each other). Similar, yet non-parallel, articles, however, are written by different authors and often exhibit differences in the structural features that Resnik and Smith (2003) described. Due to these reasons, the use of these features in identifying similar articles, such as Wikipedia articles or news articles, is not suitable.

In Wikipedia, articles that describe more than one aspect often structure their contents into different sections (and sub-sections). E.g., the Wikipedia article of "United Kingdom" includes the following section headings (titles): *Etymology, History, Geography, Dependencies, Politics.*[1] Each section heading summarises the aspects or topics described in the section. The initial analysis of Wikipedia (described in Chapter 4) indicated that articles that correspond in translation relations may also have strong alignment between the section headings, i.e., they may share many overlapping section headings. The same analysis also indicated that articles with different section headings are more likely to be less similar at the content level, as they describe different aspects of the topic. Examples of these articles have been discussed in Section 4.4.1.

In this work, the author investigates whether an approach based on measuring the similarity between the section headings of the documents can indicate the similarity of the content of the documents. The proposed approach make use of translation resources built from Wikipedia interlanguage links and Wiktionary, a multilingual dictionary available in 152 languages.[2] This work aims to answer the third research question:

RQ3. Can language independent approaches be used to identify cross-lingual similarity in Wikipedia? Furthermore, how effective are section headings for computing article similarity compared to using the full content?

(a) How does the method compare to approaches using linguistic resources, such as MT systems? I.e., how effective is information derived from Wikipedia and Wiktionary for translating section headings compared to using high-quality translation resources?

(b) How does the performance for the approach vary for different language pairs?

(c) What language-independent features are best for measuring cross-lingual similarity in Wikipedia?

First, the author describes the characteristics of section headings of Wikipedia articles in Section 8.2. Section 8.3 describes available data that can be gathered from Wik-

[1] The use of sections and sub-sections are common in Wikipedia, however, it is not always available for short articles.

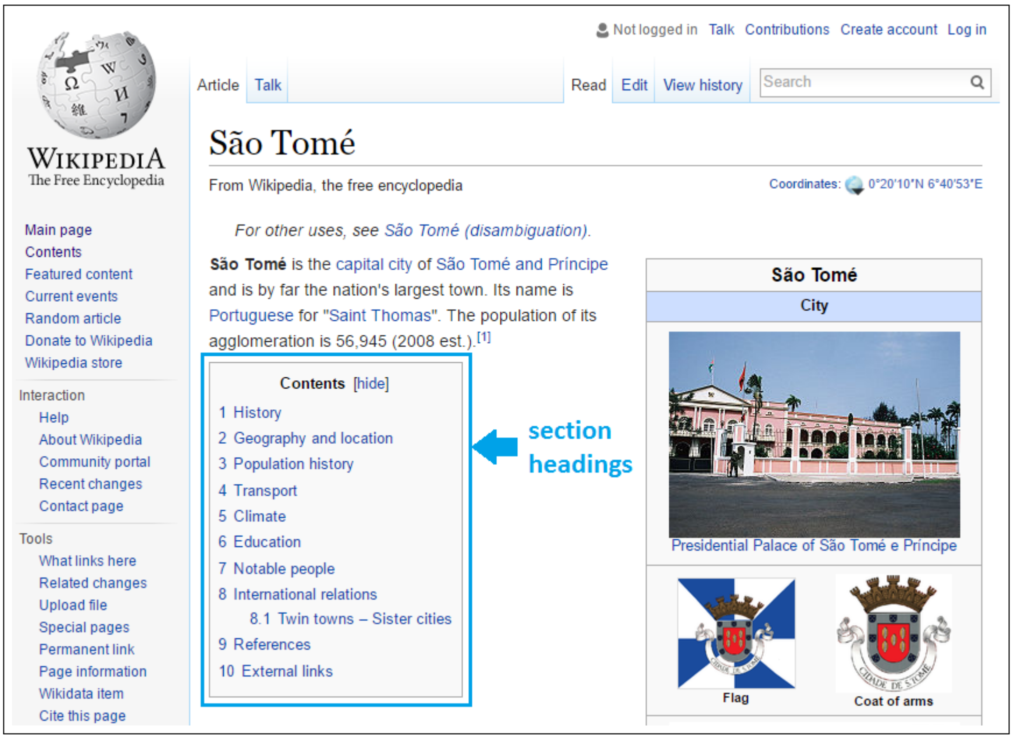
[2] https://meta.wikimedia.org/wiki/Wiktionary#List_of_Wiktionaries, accessed on 20 October 2016.

tionary, to complement the Wikipedia bilingual lexicon. The proposed approach is described in Section 8.4 and evaluation setup in Section 8.5. Finally, the results are described in Section 8.6 and discussed in Section 8.7.

8.2 Section headings in Wikipedia articles

Wikipedia articles often describe many aspects of a topic. In some cases, the content of these articles are presented in different sections, with each section describing a different aspect of the topic. The titles of sections and sub-sections in the article contents are often listed in a table titled “Contents” shown at the top of the article page as shown in Figure 8.1.

These *section headings* are valuable information to identify the main aspects described in the documents. When interlanguage-linked articles describing the same topic contain the same section heading, this indicates that they describe the same aspect of the partic-



São Tomé

From Wikipedia, the free encyclopedia

Coordinates: 0°20′10″N 6°40′53″E

For other uses, see São Tomé (disambiguation).

São Tomé is the capital city of São Tomé and Príncipe and is by far the nation's largest town. Its name is Portuguese for "Saint Thomas". The population of its agglomeration is 56,945 (2008 est.).[1]

Contents [hide]

1 History
2 Geography and location
3 Population history
4 Transport
5 Climate
6 Education
7 Notable people
8 International relations
    8.1 Twin towns – Sister cities
9 References
10 External links

section headings

São Tomé

City

Presidential Palace of São Tomé e Príncipe

Flag

Coat of arms

Fig. 8.1 Example of section headings (“São Tomé” article)

ular topic, although similarity within the aspects being described may still vary between languages. On the other hand, articles with different section headings indicate that both articles describe different aspects of the topic. These information can be used to indicate the similarity of the document content.

Unfortunately, unlike some of the links in the main contents of the articles, section headings in Wikipedia documents are not interlanguage-links. I.e., there are no direct links aligning similar sections that are written in different languages. To identify cross-lingual similarity between section headings of two articles, a bilingual resource (e.g., dictionary) is required to perform a translation prior to identifying overlapping sections in the documents.

Wikipedia interlanguage links have been utilised in the previous approach (Chapter 6) to build a bilingual resource to perform translations. Whilst it contains a large number of bilingual words and terms, they are limited to encyclopaedia-related entries. These are mostly noun or noun phrases ranging from named entities (e.g. people, companies, movies, or countries), topics (e.g. 'geography', 'medicine', 'culture'), or words that describe a particular concept (e.g. 'interaction', 'agreement', 'injustice').

Section heading titles, on the other hand, often contain entries that do not conform to the usual Wikipedia entries. This is illustrated in Figure 8.2, which shows section headings of three different articles. The article about "Trinidad and Tobago" (a coun-

1 Etymology
2 Geography
3 History
4 Politics
5 Economy
6 Demographics
7 Culture
8 Sports
9 See also
10 References
11 External links

(a) 'Trinidad and Tobago'

1 *Design and development*
2 *Variants*
3 *Operational history*
4 *Operators*
  4.1 *Civilian operators*
  4.2 *Military operators*
5 *Specifications (F.60)*
6 See also
7 References
8 External links

(b) 'Farman F.60 Goliath'

1 *Early days*
2 *Rise in literary career*
3 *Recent successes*
4 *Claims of plagiarism*
5 Bibliography
  5.1 Novels
  5.2 *Story books*
  5.3 Chronicles
  5.4 Memoirs
  5.5 Essays

(c) 'Alfredo Bryce Echenique'

Fig. 8.2 Example of section headings (non-Wikipedia concepts are shown in italic)

try in South America) (Figure 8.2a) contain the following section headings: "Etymology", "Geography", "History", etc.. Each section heading is a concept that is described in Wikipedia. This means these section titles can be translated to other languages using the Wikipedia bilingual lexicon, given that the concept is also described in the other languages. On the other hand, section headings of the other two articles: i) “Farman F.60 Goliath”, a French airliner and bomber (Figure 8.2b), and ii) “Alfredo Bryce Echenique”, a Peruvian writer (Figure 8.2c), contain non-conceptual titles, such as "Operational history", "Rise in literary career" and "Recent successes" (other similar titles are shown in italic). These titles do not exist as entries in Wikipedia, and as a result, they do not exist in the Wikipedia bilingual lexicon.

To perform translations on these out-of-domain titles, a general dictionary needs to be used instead. In this experiment, the author investigated the use of Wiktionary to carry out the translation of these titles, in addition to the bilingual lexicon extracted from Wikipedia previously utilised in the previous two experiments (Chapter 6 and Chapter 7).

8.3 Wiktionary as a bilingual resource

Wiktionary is a free multilingual dictionary that is available in 152 languages.[3] Wiktionary is investigated in this study as it contains more lexical knowledge compared to Wikipedia (Müller & Gurevych, 2009). Different to a normal dictionary, each language version in Wiktionary aims to “define all words from all languages in its own language, so that readers will be able to find definitions of all words in all languages in their own language.”[4] In other words, Wiktionary for a language version also contains entries from other languages. By July 2016, the English version contained over 4.7 million English entries from 2,500 languages.[5] Each Wiktionary entry describes the definition of a word, a term or a saying. It also contains information such as its etymologies, synonyms, antonyms, and often translations to other languages.

[3] `https://meta.wikimedia.org/wiki/Wiktionary#List_of_Wiktionaries` (20 Oct 2016).

[4] `https://en.wiktionary.org/wiki/Wiktionary:What_Wiktionary_is_not`, accessed on 11 September 2016.

[5] `https://en.wiktionary.org`, accessed on 3 July 2016.

Similar to Wikipedia, Wiktionary also contains interlanguage links that refer users to relevant pages in different languages. However, unlike Wikipedia, these interlanguage-linked pages do not contain the translation of the *word*. Instead, they represent the translation of the *word definition*. For example, the English Wiktionary page about 'etymology' (that describes the description of the word in English) has interlanguage links to more than 51 language versions.[6] These different language versions describe the definition of the word 'etymology' in different languages, without translating the word 'etymology' itself, i.e. the entry title remains to be the original English word rather than the translation. Extracting entry titles in Wiktionary are therefore not useful for extracting bilingual resources.

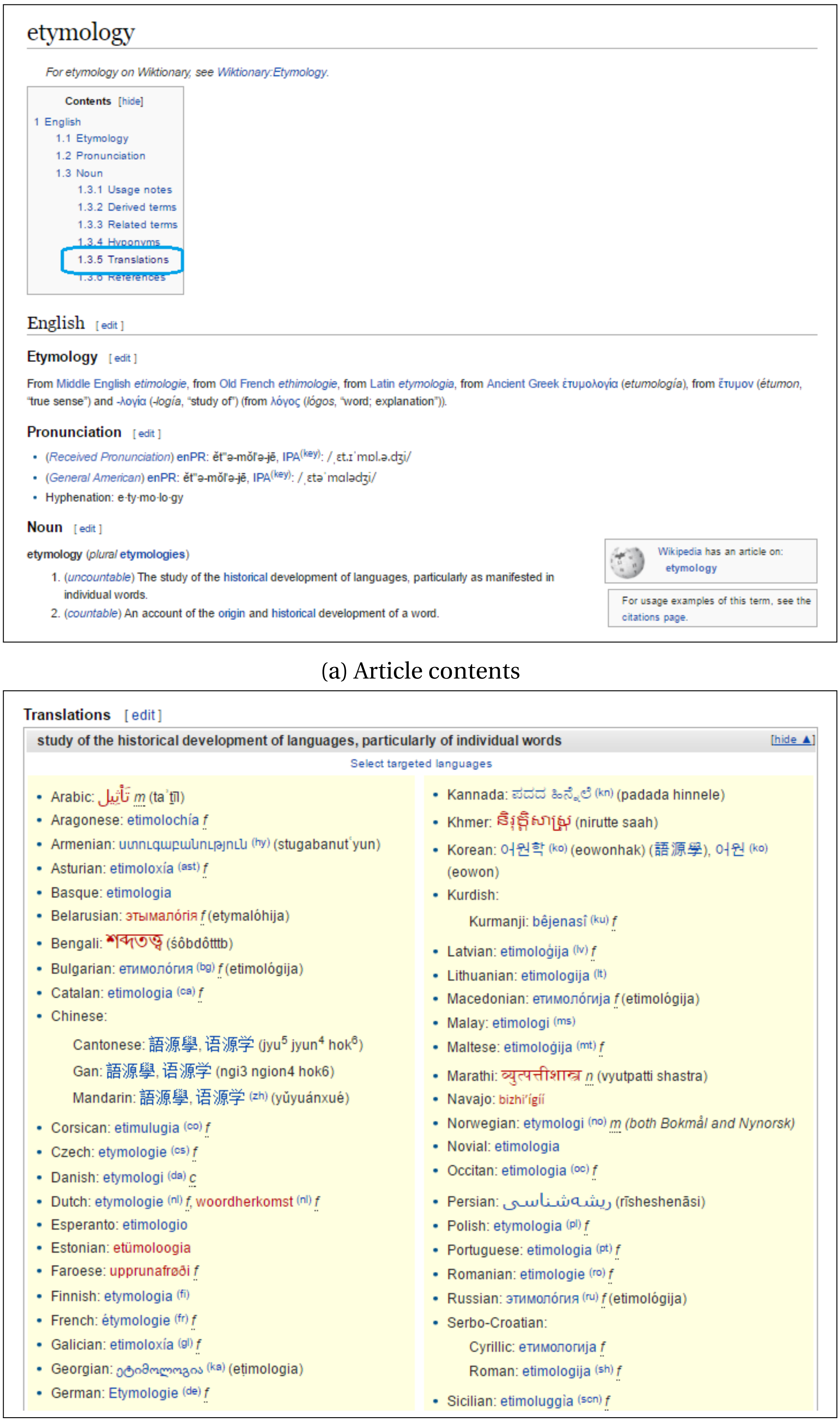
Wiktionary does contain the translation information for each word, however, this is often contained in the source language version. E.g., in the English version of Wiktionary, the translation of each entry is instead located in the 'Translation' section of the article. An example is shown in the 'table of content' of the article shown in Figure 8.3a, and the content of the translation section is shown in Figure 8.3b. Other language versions in Wiktionary, however, often structure their contents differently, e.g., the 'translation' section is named differently. This translation information is extracted to create a bilingual dictionary for this study.

8.4 Method

The proposed method aims to measure cross-lingual similarity between a document pair d_1 and d_2 in a non-English language (L_1) and English (L_2) by measuring the similarity between their section headings. These section headings are referred to as H_1 and H_2, respectively.

Firstly, bilingual dictionaries are built by extracting translation information from Wikipedia and Wiktionary. Using these dictionaries, the translation of section headings is carried to translate H_2 into L_1; the translated headings are referred to as H_2'. Structure

[6] https://en.wiktionary.org/wiki/etymology, accessed on 27 June 2017.



etymology

For etymology on Wiktionary, see Wiktionary:Etymology.

Contents [hide]

1 English
 1.1 Etymology
 1.2 Pronunciation
 1.3 Noun
  1.3.1 Usage notes
  1.3.2 Derived terms
  1.3.3 Related terms
  1.3.4 Hyponyms
  1.3.5 Translations
  1.3.6 References

English [edit]

Etymology [edit]

From Middle English *etimologie*, from Old French *ethimologie*, from Latin *etymologia*, from Ancient Greek ἐτυμολογία (*etumología*), from ἔτυμον (*étumon*, "true sense") and -λογία (*-logía*, "study of") (from λόγος (*lógos*, "word; explanation")).

Pronunciation [edit]

- (*Received Pronunciation*) enPR: ĕt″ə-mŏlʹə-jē, IPA(key): /ˌɛt.ɪˈmɒl.ə.dʒi/
- (*General American*) enPR: ĕt″ə-mŏlʹə-jē, IPA(key): /ˌɛtəˈmɑlədʒi/
- Hyphenation: e‧ty‧mo‧lo‧gy

Noun [edit]

etymology (*plural* **etymologies**)

1. (*uncountable*) The study of the historical development of languages, particularly as manifested in individual words.
2. (*countable*) An account of the origin and historical development of a word.

Wikipedia has an article on:
etymology

For usage examples of this term, see the citations page.

(a) Article contents

Translations [edit]

study of the historical development of languages, particularly of individual words [hide ▲]

Select targeted languages

- Arabic: تَأْثِيل *m* (taʾṯīl)
- Aragonese: etimolochía *f*
- Armenian: ստուգաբանություն (hy) (stugabanutʿyun)
- Asturian: etimoloxía (ast) *f*
- Basque: etimologia
- Belarusian: этымалёгія *f* (etymalóhija)
- Bengali: শব্দতত্ত্ব (śôbdôtttb)
- Bulgarian: етимоло́гия (bg) *f* (etimológija)
- Catalan: etimologia (ca) *f*
- Chinese:
  Cantonese: 語源學, 语源学 (jyu^5 jyun^4 hok^6)
  Gan: 語源學, 语源学 (ngi3 ngion4 hok6)
  Mandarin: 語源學, 语源学 (zh) (yǔyuánxué)
- Corsican: etimulugia (co) *f*
- Czech: etymologie (cs) *f*
- Danish: etymologi (da) *c*
- Dutch: etymologie (nl) *f*, woordherkomst (nl) *f*
- Esperanto: etimologio
- Estonian: etümoloogia
- Faroese: upprunafrøði *f*
- Finnish: etymologia (fi)
- French: étymologie (fr) *f*
- Galician: etimoloxía (gl) *f*
- Georgian: ეტიმოლოგია (ka) (eṭimologia)
- German: Etymologie (de) *f*
- Kannada: ಪದದ ಹಿನ್ನೆಲೆ (kn) (padada hinnele)
- Khmer: និរុត្តិសាស្ត្រ (nirutte saah)
- Korean: 어원학 (ko) (eowonhak) (語源學), 어원 (ko) (eowon)
- Kurdish:
  Kurmanji: bêjenasî (ku) *f*
- Latvian: etimoloģija (lv) *f*
- Lithuanian: etimologija (lt)
- Macedonian: етимоло́гија *f* (etimológija)
- Malay: etimologi (ms)
- Maltese: etimoloġija (mt) *f*
- Marathi: व्युत्पत्तीशास्त्र *n* (vyutpatti shastra)
- Navajo: bizhiʼígíí
- Norwegian: etymologi (no) *m (both Bokmål and Nynorsk)*
- Novial: etimologia
- Occitan: etimologia (oc) *f*
- Persian: ریشه‌شناسی (rīsheshenāsi)
- Polish: etymologia (pl) *f*
- Portuguese: etimologia (pt) *f*
- Romanian: etimologie (ro) *f*
- Russian: этимоло́гия (ru) *f* (etimológija)
- Serbo-Croatian:
  Cyrillic: етимологија *f*
  Roman: etimologija (sh) *f*
- Sicilian: etimoluggìa (scn) *f*

(b) Translations of the word 'etymology' in other languages

Fig. 8.3 An extract of the English Wiktionary page of 'etymology'

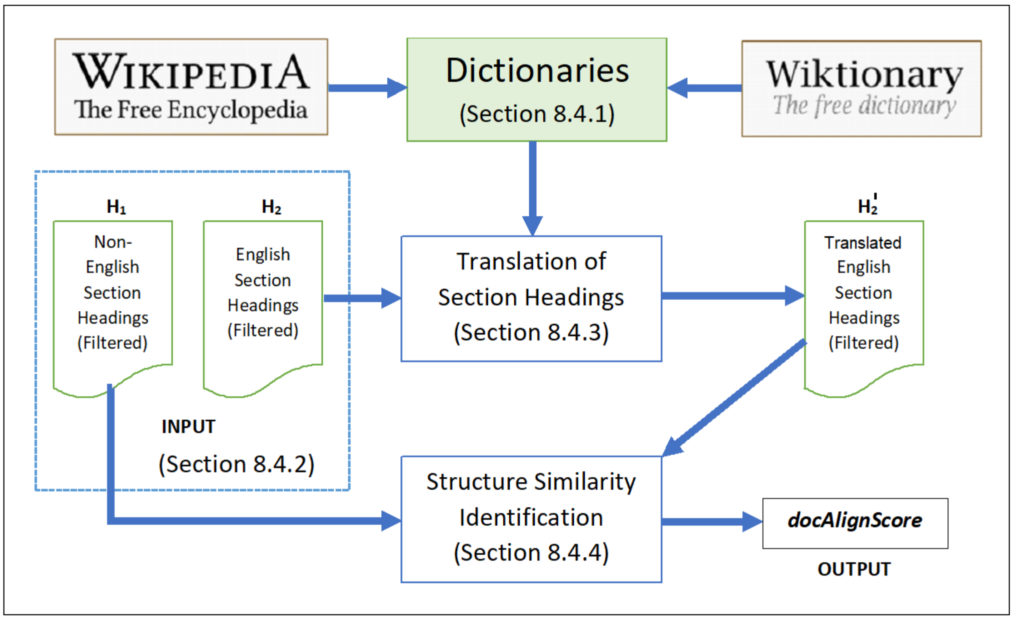

WIKIPEDIA
The Free Encyclopedia
Dictionaries
(Section 8.4.1)
Wiktionary
The free dictionary
H₁
Non-English Section Headings (Filtered)
H₂
English Section Headings (Filtered)
INPUT
(Section 8.4.2)
Translation of Section Headings
(Section 8.4.3)
H₂'
Translated English Section Headings (Filtered)
Structure Similarity Identification
(Section 8.4.4)
docAlignScore
OUTPUT

Fig. 8.4 Method to compute overlapping sections

similarity is then computed by measuring the similarity between H_1 and H'_2 (both are now written in L_1). These processes (Figure 8.4) are described in more detail below.

8.4.1 Identifying dictionaries for translation

The first stage is to to build dictionaries to perform the translation of section headings using two sources: Wikipedia and Wiktionary.

Wikipedia interlanguage-links have been shown to be promising bilingual resources. To extract this information, the author utilises the link-based bilingual lexicon method (Adafre & de Rijke, 2006) to extract the titles of Wikipedia interlanguage-linked articles for each language pair. The titles of these paired articles are used as dictionary entries for the language pair.

Information from Wiktionary is extracted by collecting all Wiktionary entries and their translations. For cases where multiple translations for an entry exist, all possible translations are included in the dictionary. The resulting dictionary contains translations of words (e.g., ‘etymology’), phrases (e.g., ‘digestive system’), and sayings (e.g., ‘raining cats and dogs’). For example, an extract of the English-German dictionaries is shown in

English entries	**German entries**
a friend in need is a friend indeed	ein Freund in der Not ist ein Freund in der Tat
a journey of a thousand miles begins with a single step	eine Reise von tausend Meilen beginnt mit dem ersten Schritt
a leopard cannot change its spots	die Katze lässt das Mausen nicht
abdicate	abdanken; verweigern; ausstoßen; abdanken
archaeological	archäologisch
digestive system	Verdauungssystem
etymology	Etymologie
raining cats and dogs	Bindfäden regnen

Fig. 8.5 Example of English-German Wiktionary translations

Figure 8.5.

8.4.2 Extraction of section headings

In the second stage, all titles of the sections and sub-sections in the articles are extracted from the *table of contents* of the article. For simplicity, the level of the sections are not preserved at this stage, i.e., all sections and sub-sections are assumed to be in the same level. Afterwards, common section headings are filtered out as they are often not useful in the similarity process. Common section headings are identified by counting the frequency of headings that appear in the Wikipedia set, prior to filtering out the three most frequent section headings (this threshold was determined empirically). E.g., for the English Wikipedia, the three most frequent headings were ‘External Links’ (appearing in 54% English documents), ‘References’ (52% English documents) and ‘See also’ (26% English documents). These section headings are filtered out as they do not make useful contributions when computing article similarity. Finally, stopwords in each section heading are identified and removed; stopwords were identified using a list of frequent words gathered from the Wikipedia corpus.

8.4.3 Translation of section headings

In this stage, the English section headings (H_2) are translated into L_1 (the non-English language), resulting in H'_2. For each section heading (h_1, h_2, ..., h_n) in H_2, the translation process is as follows:

1. If h_i exists in the dictionary, then extract its translation t_i. E.g., for English-German, 'folk etymology' is translated into 'Volksetymologie'. If the dictionary contains > 1 possible translation then extract all translations.

2. If h_i does not exist as an entry in the dictionary:

   (a) If it includes > 1 word, split the heading h_i into each word (w_1, w_2, ..., w_n), translate each word separately, and concatenate the results.

   (b) If after word-splitting, the translation is not found, trim one character from the end of the word and search for its translation. This process is performed to imitate the idea of stemming, as dictionaries are more likely to contain the stem of the words rather than the long form.[7] This process is performed recursively until either a translation is found, or stops if the original word has four characters left (previously determined through empirical investigation).

   (c) If no translation is found, h_i is left in its original form.

3. Both h_i and t_i (if found) are then included in H'_2.

4. Steps 1-3 are repeated until all headings have been translated.

8.4.4 Identifying overlapping sections

Once all the section headings have been translated, the similarity between both lists of section headings (i.e., H_1 and H'_2) were measured to identify the proportion of overlapping sections. In this stage, similar section headings in both documents are identified and aligned. Firstly, every source heading $s_i \in H_1$ is paired to every target heading $t_j \in H'_2$. For each s_i, the most similar target heading t_n (allowing many-to-one alignments) is identified using the following alignment and section similarity scoring (*secSimScore*) methods:

[7] Performing lemmatisation would be more appropriate, however, this will involve the need of using a language-dependent resources, e.g., a lemmatiser.

1. If s_i is included in t_n, both headings are aligned; $secSimScore(s_i, t_n) = 1$.

2. If not, split heading s_i into each word $(w_1, w_2, ..., w_p)$:

   (a) For each word w_m, identify if w_m is included in t_n. If not, trim the word by one character (up to a minimum word length of 4) and recursively search for the trimmed word:

   $$wordScore(w_m, t_n) = \begin{cases} 1, & \text{if } w_m \text{ (original or trimmed) is included in } t_n \\ 0, & \text{otherwise} \end{cases} \tag{8.1}$$

   (b) After all words have been aligned, $secSimScore$ is calculated by measuring the proportion of words in s_i that are found in t_n:

   $$secSimScore(s_i, t_n) = \frac{\sum_{m=1}^{p} wordScore(w_m, t_n)}{p} \tag{8.2}$$

   where p is the number of words in s_i. E.g., if all words in s_i are found in t_n, $secSimScore$ equals 1; if only one of two words can be aligned, $secSimScore$ equals 0.5.

3. Step 1-2 are performed between s_i and the remaining sections in H_2'. After which, the highest scoring pair is selected as the alignment for s_i.

8.4.5 Measuring similarity of the document pair

After the alignment process between H_1 and H_2', the scores are aggregated to derive a score for the document pair ($docSimScore$). The aligned sections in both documents are referred to as A_1 and A_2', respectively ($A_1 \subseteq H_1$ and $A_2' \subseteq H_2'$). Three different methods to measure the $docSimScore$ are investigated:

1. *align1*: This method does not take the $secSimScore$ into account, but instead re-

lies on the number of aligned sections in both documents only:

$$docSimScore = \frac{(|A_1| + |A'_2|)}{(|H_1| + |H'_2|)} \tag{8.3}$$

where $|A_1|$ and $|A'_2|$ represent the number of aligned sections in H_1 and H'_2, respectively, and $|H_1|$ and $|H'_2|$ are the number of sections in H_1 and H'_2.

2. *align2*: This method takes the $secSimScore$ into account:

$$docSimScore = \frac{(S_1 + |A'_2|)}{(|H_1| + |H'_2|)} \tag{8.4}$$

where S_1 represents the sum of $secSimScore$ for each section in A_1.

3. *align3*: In this method, aligned sections with $secSimScore < 1$ are filtered out, prior to calculating the $docSimScore$ using Equation 8.3.

Finally, previous experiment (Section 6.2) also shows that articles with different lengths are less likely to be similar. Therefore, the author also investigates an additional feature, *section length ratio* (*sl*). This feature represents the length ratio between the smaller number of sections and the bigger number of sections in the two documents.

8.5 Experimental setup

An in-house tool was developed to extract the translation information from the English Wiktionary. As described in Section 8.3, extracting the translation information from a source (non-English) language to the target (English) language is not straightforward and requires language-dependent knowledge to adapt the approach to suit the requirements of each language. To simplify the task and to focus the approach in a language-independent way, only the English Wiktionary is used to extract the translations of words and build the dictionary.

To identify stopwords, the author extracted a list of frequent words from each language version of Wikipedia to be utilised as a stopword list (an average size of 871 words per language).

Due to the unavailability of the Croatian-English translation in Wiktionary, only 7 language pairs are used in this study: German (a *highly-resourced* language), and 6 *under-resourced* languages: Greek (EL), Estonian (ET), Lithuanian (LT), Latvian (LV), Romanian (RO) and Slovenian (SL); all paired to English (EN).[8]

The approach is evaluated using the Wikipedia similarity corpus (described in Chapter 5) in the 7 language pairs. Two annotators assessed the similarity (Q1) of each document pair using a 5-point Likert Scale; the mean rating between the two assessors is used to represent the similarity of the document pair.

Documents without section headings were removed for these experiments, resulting in 600 document pairs across the 7 language pairs. The proposed method is compared to *c3g*, the tf-idf cosine similarity of the *char-3-gram overlap* between the entire article contents.[9] To investigate the effectiveness of Wikipedia-Wiktionary as translation resources, Google Translate was used as a state-of-the-art comparison.

8.6 Results

The author reports the Spearman-rank correlations between similarity scores computed using methods from Section 8.4 and the average human-annotated similarity scores from the evaluation corpus in Table 8.1 (“Individual Features”). Results show that features based on section headings (ρ=0.36 for $align1$) were able to achieve comparable overall correlations compared to using char-3-gram overlap ($c3g$) on the entire article content (ρ=0.34). Results using $align2$ was similar (ρ=0.35). The $align3$ method, however, achieved significantly lower score (ρ=0.23), suggesting that the strict alignment process may have lost valuable cross-lingual information. Section length (sl) was shown to perform consistently across most language pairs (ρ=0.35). The $c3g$ method, however, per-

[8] An investigation was carried out on substituting Croatian with Serbo-Croatian as they represent the same language family. However, this introduced many non-translated words and inaccurate translations that decrease the accuracy of their method. Therefore, Croatian-English is not investigated in this work.

[9] This feature is identified as the best language-independent feature to identify cross-lingual similarity in this corpus set.

Table 8.1 Correlation scores (Spearman's ρ) of individual and combined features

Lang	Individual Features					Combined Features		
	Section Headings (SH)				Article	SH	SH + Article	
	align1	**align2**	**align3**	**sl**	**c3g**	**align1_sl**	**sl_c3g**	**align1_sl_c3g**
DE	0.33*	0.28	-0.01	0.45*	**0.46***	0.42*	**0.67***	0.59*
EL	0.17	0.19	0.19	**0.42***	0.38*	0.36*	**0.56***	0.47*
ET	0.27*	0.29*	0.29*	0.37*	**0.57***	0.37*	**0.58***	0.54*
LT	0.43*	**0.44***	0.39*	0.40*	0.34*	0.54*	0.51*	**0.58***
LV	0.31*	0.33*	0.18	**0.34***	**0.34***	0.40*	0.46*	**0.49***
RO	**0.54***	**0.54***	0.51*	0.14	0.20	**0.40***	0.20	0.39*
SL	**0.41***	0.32*	0.00	0.33*	0.03	**0.44***	0.33*	0.42*
Mean	**0.36**	0.35	0.23	0.35	0.34	0.42	0.49	**0.50**
Note: $*p < 0.01$; the best results for the "Individual Features" and "Combined Features" are shown in bold; "Avg" score is calculated using Fisher transformation.								

formed poorly for RO-EN and SL-EN (ρ=0.20 and ρ=0.03, not statistically significant), possibly due to dissimilar surface forms between languages. Section heading features were shown to achieve either the same or better correlation scores than $c3g$ in 5 of the 7 language pairs.

Further results indicate that a combination of features produces a more robust similarity measure. Table 8.1 ("Combined Features") reports the three best feature combinations. Firstly, a combination of only Section Headings (SH) features, $align1\_sl$, increases the correlation score to 0.42 ($\uparrow$16.67% compared to $align1$, the best individual feature). Correlation can further be increased by combining both SH and article features. Combining the section length and the char-3-gram overlap (sl_c3g) further improved the performance (ρ=0.49; $\uparrow$36.11%); considering that this combination can be computed without the need of a dictionary, this result is very promising. Lastly, the combination of three features, $align1\_sl\_c3g$, achieves the highest correlation score (ρ=0.50; $\uparrow$38.89%).

The results reported previously were derived by utilising Wiktionary and Wikipedia. As a comparison, the approach is carried out with a higher-quality translation resource. In this case, Google Translate was used to translate all the section headings. The same structure alignment algorithm and document similarity measure ($align1$, $align2$, $align3$) were then carried out after the translation. These approach (utilising Google Translate) are referred to as $gAlign1$, $gAlign2$ and $gAlign3$. Table 8.2 shows the correlation scores between these approaches to human judgments.

The results show that the correlation scores when utilising Wikipedia and Wiktionary are very similar for most language pairs. In two language pairs (DE-EN and RO-EN), $align1$ was shown to outperform $gAlign2$, the best algorithm when utilising Google Translate for these two language pairs. Similar performances were achieved in ET-EN (ρ_{align2}=0.29 and $\rho_{gAlign3}$=0.31), LT-EN (ρ_{align2}=0.44 and $\rho_{gAlign2}$ = 0.48), LV-EN (ρ_{align2} =0.33 and $\rho_{gAlign2}$=0.41) and SL-EN (ρ_{align1}=0.41 and $\rho_{gAlign2}$ = 0.46). The major difference when utilising these two resources was observed only in one language pair (EL-EN), where $gAlign2$ significantly outperforms $align2$ ($\rho_{gAlign2}$=0.46 and ρ_{align1}=0.17).

8.7 Discussion

Results from the previous section have shown how structural similarity features could be used to identify similarity in the documents. Three different alignment algorithms were proposed ($align1$, $align2$ and $align3$). The use of section length was also investigated for measuring similarity. In this section, the results across languages are analysed. Cases where the algorithm failed to analyse the alignment are identified along with the performance of the bilingual resources used in this study.

Table 8.2 Correlation scores (Spearman's ρ) for individual feature for each language pair

Lang	Wikipedia-Wiktionary			Google Translate		
	align1	**align2**	**align3**	**gAlign1**	**gAlign2**	**gAlign3**
DE	**0.33***	0.28	-0.01	0.27*	0.27*	-0.08*
EL	0.17	0.19	0.19	0.46*	**0.48***	0.43*
ET	0.27*	0.29*	0.29*	0.27*	0.29*	**0.31***
LT	0.43*	0.44*	0.39*	0.46*	**0.48***	0.41*
LV	0.31*	0.33*	0.18	0.38*	**0.41***	0.39*
RO	**0.54***	**0.54***	0.51*	0.51*	0.52*	0.49*
SL	0.41*	0.32*	0.00	0.45*	**0.46***	0.42*
Mean	0.36	0.35	0.23	0.40	0.42	0.35
Note: $^*p < 0.01$						

8.7.1 Results analysis

First, the author analyses the *different alignment methods that were proposed.* Firstly, the results in Table 8.1 show that the accuracy of performance using $align1$ and $align2$ is similar; $align1$ has a higher average correlation score (ρ=0.36 and ρ=0.35 for $align1$ and $align2$, respectively); however, $align2$ performs the same or better than $align1$ in more language pairs, i.e. EL-EN, ET-EN, LT-EN, LV-EN, and RO-EN. The use of $align3$, which require all words in the EN section headings to appear in the non-English section headings prior to alignment, exhibits poorer performance. This shows that the stricter alignment might have missed some alignment information.

Interestingly, the results in Table 8.2 shows that the $align2$ method outperforms the $align1$ method on all language pairs when using Google as the translation resource (average ρ=0.42 for $gAlign2$ and ρ=0.40 for $gAlign1$). Moreover, the average correlation score for $gAlign3$ is shown to be as high as $align2$ (ρ=0.35). This shows that the stricter alignment methods ($align2$ or $align3$) can be improved significantly if high quality resources are being utilised in the translation process. On the other hand, when noisier or smaller resources are being used (e.g. Wikipedia and Wiktionary), a less strict alignment method will perform better, in this case $align1$ and $align2$.

The results also show that the section length ratio, sl, can also be used to identify similarity in Wikipedia articles. Comparing the results across languages, sl was also shown to perform more consistently across languages (with ρ ranging between 0.33 and 0.45 for 6 out of 7 language pairs). Furthermore, this feature achieves better correlation to human judgments to the best $align[n]$ method in four language pairs, DE-EN, EL-EN, ET-EN and LV-EN.

Comparing all the individual features, $c3g$ achieves more consistent results across languages, except for RO-EN and SL-EN (ρ=0.20 and ρ=0.03, respectively); meanwhile, the correlation scores of $align1$ were shown to vary widely across language pairs. The $align1$ method achieved a poor correlation (ρ=0.17 for EL-EN, not statistically significant), but a much higher correlation (ρ=0.54) for RO-EN. This finding shows that the performance of this method may be affected by other issues, further investigated in Section 8.7.3.

In total, the use of section heading information (i.e. $align1$, $align2$, $align3$, and sl) was able to achieve higher correlation to human judgments compared to the use of char-3-gram overlap, $c3g$. Considering that these features are computed using only section headings and not taking the entire article content into account, this finding is very promising as they can measure similarity using much less data (i.e., the length of section headings is considerably smaller than processing the entire WIkipedia articles).

These findings further indicate that the section heading similarity approach should be combined with other features to produce a more robust measure for identifying similarity. The results in Table 8.1 produce an interesting finding: even after using a feature that measures similarity using the entire article (in this case $c3g$), the results can still further be improved by combining this feature with structural similarity features, although they were based only on the section headings. Comparing the correlation scores of the combined features to $c3g$, the author observed a 44% increase when combining $c3g$ and sl, or a 47% when combining $c3g$ and $align1$ and sl; each having equal weight.

Overall, the highest correlation scores in all 7 language pairs were achieved by utilising the proposed structure similarity approaches, either as individual features or in combination with other features.

8.7.2 Dictionary analysis

In this section, the author analyses the dictionaries being used in this study. Figure 8.6 shows the dictionary size derived from Wikipedia and Wiktionary used in this study, highlighting low numbers of entries for all under-resourced languages. The average entries for Wiktionary were 13,751 (min=5,484 and max=39,150), whilst Wikipedia has an average entry of 128,723 (min=24,357 and max=640,730).

Although much smaller in size, an average of 66% of Wiktionary entries were not available in the Wikipedia lexicon. The proportion of new Wiktionary entries, i.e., those not already included in the Wikipedia lexicon, for each language pairs is shown in Figure 8.7. The lowest proportion of new entries is found in LT, with only 52% of Wiktionary

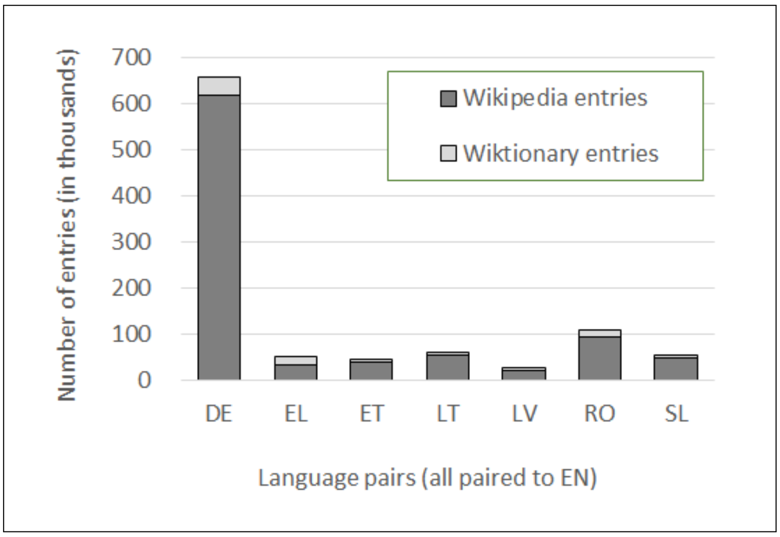


Fig. 8.6 Size of dictionaries

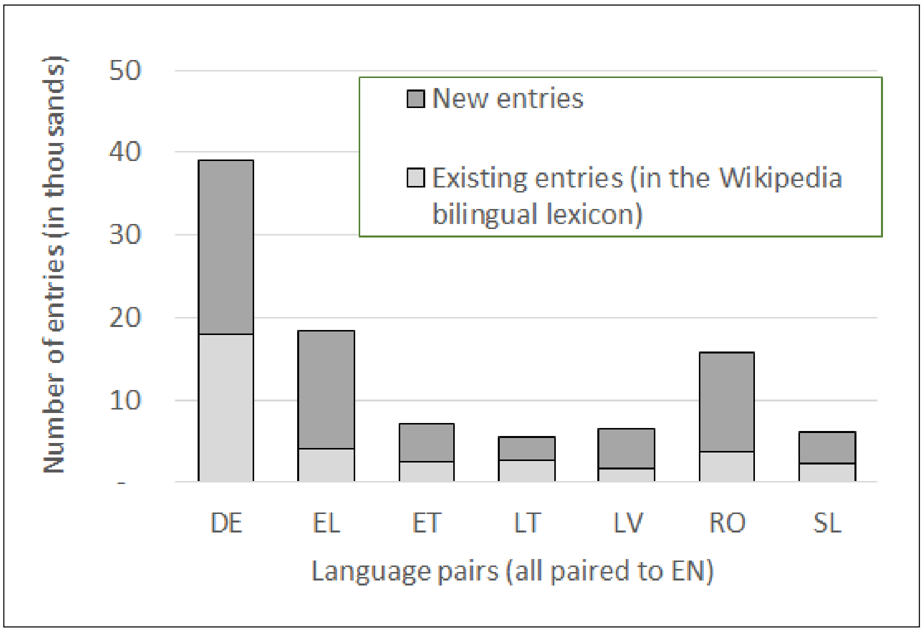


Fig. 8.7 Coverage of Wiktionary (compared to Wikipedia)

entries are not already included in Wikipedia. Meanwhile, RO and EL have the highest proportion of new entries of 77% and 78%, respectively. This shows the importance of Wiktionary in complementing the Wikipedia lexicon.

8.7.3 Failure analysis

Although the structural similarity methods have been shown to perform similarly to a feature based on the entire article content ($c3g$), its correlation across languages varies widely. In this section cases where the proposed methods failed to identify similarity accurately are investigated.

Firstly, the use of Wikipedia and Wiktionary often introduced *low quality of translation*. These were found more in EL-EN document pairs, which explains the poorer performance in this language pair, in which some section titles were not properly translated.

In cases where the section headings were accurately translated, an alignment may be missed for some cases because the proposed method had *a limited capability to identify synonyms*. I.e., section headings that do not have any overlapping words will not be identified to be similar. For example, in a Greek-English article pair about Scribus (a desktop publishing software), two sections contain the exact same content but with different titles. The section in English was titled "Milestones", whilst the corresponding Greek section was titled "Important Times" (see Table 8.3).[10] These synonyms are missed in the alignment process due to the lack of overlapping words.

More examples of these were also found in the EL-EN dataset, such as "working hours" (EL) and "opening hours" (EN), "collection" (EL) and "examples display" (EN), etc. Use of word embedding to identify the relatedness of these titles should be explored as a future work.

8.7.4 Limitations

The previous analysis have shown the possibility of using structural similarity features in identifying similarity. Although shown to be promising, this approach has a few limitations.

Firstly, this approach requires both articles to contain section headings. If one article does not contain at least one section heading, then this method cannot be utilised

[10] The Greek title was translated using Google Translate to increase the readability.

Table 8.3 Examples of Greek section headings

Article Title	**Greek Headings***	**English Headings**	**English Headings (translated)****
Central Air Force Museum	Working Hours (`Ωρες λειτουργίας)	Opening Hours	ανοιχτός; ανοίγω; ώρα;
	Collection (Συλλογή)	Examples Display	παράδειγμα; παράσταση; οθόνη; παροσιάζω;
Scribus	Features (Δυνατότητες)	Capabilities	*no translation found*
	Important Times (Σημαντικές χρονικές στιγμές)	Milestones	σηιλιομετροδείκτης

* The English translations of Greek headings is provided here for ease of reading. They are not used in the alignment process.
** The Greek translations of English headings is carried out using Wikipedia and Wiktionary. They are displayed to show the lack of overlapping words between the examples

to identify the structural similarity between the documents. This information is mostly available for medium to large Wikipedia articles, however, is often unavailable for very small articles. In this case, different features or methods will need to be used instead.

Secondly, although this approach utilises freely available resources, both Wikipedia and Wiktionary (or other general dictionaries) must be available for the languages. The use of the Wikipedia bilingual lexicon only will reduce the performance of this approach due to its lack of general lexicon words available. Since the sizes and coverage of languages in Wikipedia and Wiktionary are increasing rapidly on a daily basis, the coverage of languages that this method can process is expected to increase further in the future.

Another limitation, is that some document pairs may have similar structure but the content within the sections may be very different. In this case, the author recommends that the structure similarity features should not be used as an individual feature. Instead, it should be combined with other features to produce a more robust measure.

Finally, more investigation is required to analyse the performance of this method for similar document pairs that are richer in content (i.e., containing more than 1,000 words) as they may exhibit different characteristics to those described in this work. These documents are not included in the evaluation corpus and therefore were not investigated in this experiment. Furthermore, some features that did not perform well with short documents may work better with longer documents due to the richer data (e.g., a larger number of section headings). Due to the limited number of available section headings in the dataset, the order of the aligned section headings is not taken into account in this work. More work is therefore needed to further evaluate these features in larger documents.

8.8 Conclusion

This chapter describes an approach for identifying cross-lingual similarity of Wikipedia articles by measuring the structural similarity (i.e., similarity of section headings) of the articles. The approach utilises dictionaries derived from Wikipedia and Wiktionary to identify translated contents in the different languages. In this section, the author answered the research questions presented earlier in this chapter.

RQ3. Can language-independent approaches be used to identify cross-lingual similarity in Wikipedia? Results show that the section heading similarity feature ($align1$) and ratio of section length (sl) achieve a positive but weak correlation with human judgments (ρ=0.36 and ρ=0.35, respectively). Interestingly, the results show that identifying similarity using only the section headings managed to achieve a comparable performance to using the best content similarity feature, i.e., the overlap of char-3-grams ($c3g$) on content from the entire article, which achieved ρ=0.34.

(a) How does the method compare to approaches using linguistic resources, such as MT systems? The use of Wikipedia and Wiktionary resources in this approach was shown to be as efficient as utilising Google Translate (for many language pairs). The results are promising given that these resources are freely available for a large number of languages.

(b) How does the performance for the approach vary for different language pairs? The performance of the section heading similarity features show a considerable range. It achieves a high performance for RO-EN and SL-EN. However, it performs poorly in three language pairs, EL-EN, ET-EN and LV-EN; these three languages also have the smallest dictionaries which are likely to affect the performance of the method.

(c) What language-independent features are best for measuring cross-lingual sim-

ilarity in Wikipedia? The results indicated that a combination of three features: the section heading similarity feature ($align1$), ratio of section length (sl) and char-3-grams overlap ($c3g$) achieves the best correlation to human judgments (ρ=0.50).

This experiment shows that structure similarity features are promising features to measure similarity, as they can indicate the similarity of a document pair without evaluating the article contents. The correlation between these approaches to human judgments, however, can be further improved by combining structure similarity features and content similarity features and should be investigated further.

Related publications

- Paramita, M. L., Clough, P., and Gaizauskas R. 2017. Using Section Headings to Compute Cross-Lingual Similarity of Wikipedia Articles. In *Proceedings of the 39th European Conference on Information Retrieval (ECIR 2017)*, Aberdeen, *10-13 April 2017, pp. 633-639.*

Chapter 9

Classification of Similar Documents

In Chapters 6-8, the author has investigated the use of a number of language-independent features to aid the computation of similarity between articles in Wikipedia at the document level. The findings suggest that none of the features individually can accurately measure similarity at the document level. However, a combination of multiple features has been shown to correlate better with human judgments. In this section, the author investigates a different approach by turning the task into a supervised learning problem by combining these features into a document similarity classifier. Given a pair of Wikipedia articles, the classifier will predict the level of similarity between the document pair.

9.1 Background

The findings from previous approaches have identified a list of features that can be used for measuring cross-lingual similarity in Wikipedia. Some of these features, i.e., the overlap of section headings and section length ratio, measure the *structural similarity*. Oher features, such as the overlap of anchors/links, overlap of char-n-gram, overlap of cognates, and word length ratio, measure the *content similarity*. All of these features can be extracted using only information derived from Wikipedia and Wiktionary. In this chapter, these features are combined and investigated as part of a classification approach.

The use of a classifier has previously been explored to identify parallel content written in different languages. For example, Resnik (1999) developed a classifier to identify Web sites that are translations of each other. A classification approach has also been investigated for identifying parallel sentences in comparable corpora (Chu, Dabre, & Kurohashi, 2016; Munteanu & Marcu, 2005). However, a classification approach for predicting similarity in Wikipedia documents has not been investigated before.

This experiment aims to answer the following research question:

RQ3. Can language-independent approaches be used to identify cross-lingual similarity in Wikipedia?

   (a) How does the method compare to approaches using linguistic resources, such as MT systems? In this experiment, the author also aims to investigate how well the approach works for new language pairs.

   (b) How does the performance for the approach vary for different language pairs?

   (c) What language-independent features are best for measuring cross-lingual similarity in Wikipedia?

First, the author utilised the similarity corpus to create the features for the classification approach, described in Section 9.2. The experimental setup, including the different classification problems being investigated are described in Section 9.3. Finally, the results overall languages and for each language pair are reported in Section 9.4 and Section 9.5, respectively. Discussion about the results and their relation to the literature are reported in Section 9.6 and finally, the study is concluded in Section 9.7.

9.2 Identification of similarity features

The insights gathered from the similarity corpus (described in Chapter 5) have identified five characteristics of document pairs that contribute to the similarity degree of the document pairs: similar structure, overlapping named entities, overlapping fragments, contain translation, and different information. The experiment reported in Chapter 7 (see Section 7.2) has listed a number of feature types that can be used to capture these five similarity characteristics. These feature types are again summarised in Section 9.2.1. Features that have been extracted for each of these feature types are then described in Section 9.2.2.

9.2.1 Feature types

The author identified a number of feature types that can be used to capture the five characteristics of similar and non-similar article pairs:

1. **Similar structure**: A few features have been identified that can capture the similarity between the document structures. Firstly, the *word length ratio* can be used to indicate whether the document pairs contain similar content sizes. Structure similarity can also be identified by analysing the section headings, as they indicate the aspects of the topic that are discussed in the article pair. E.g., article about England may contain the following section heading titles: “History”, “Geography”, “Politics”, “Economy”, “Demography”. The *section heading similarity* is measured between the section headings of both documents. The *section length ratio*, i.e., the number of sections between the two articles, is also used to capture this characteristic.

2. **Overlapping named entities**: Named entities, such as name of people, locations, etc., are very likely to be included as concepts in Wikipedia. As a result, these named entities are often written as links which refer to the Wikipedia articles describing these named entities. Therefore, the *links overlap* feature is extracted to measure the degree of overlapping named entities. There may be cases where non-popular named entities are not included in Wikipedia, and therefore are not captured using the overlap of links. In this case, we use *word overlap* to capture this due to the high similarity of named entities across some languages. For other cases, *char-n-gram overlap* is used to capture similarity between words written in different languages.

3. **Overlapping fragments**: Fragments that overlap between different articles can be

identified by measuring the following three feature types: *links overlap*, *char-n-gram overlap* and *word overlap*.

4. **Contains translation**: This characteristic captures a stricter form of content similarity, i.e., that the content describes the same named entities and contains overlapping fragments, but also correspond to each other as translations. Therefore, the same three feature types were used to capture these characteristics: *links overlap*, *char-n-gram overlap* and *word overlap*.

5. **Different information**: Identification of different information has been somewhat captured by the lack of similarity or overlap of the first four features, e.g., document pairs that have low numbers of overlapping named entities or fragments are likely to contain different information, compared to those that have a high number of overlapping named entities or fragments. Additionally, the notion of different information can also be captured using *word length ratio*: the ratio between the lengths of document pairs. Longer articles are very likely to contain additional information not included in the shorter ones.

In total, this section has identified six different feature types. The first two feature types, i.e., *overlap of section headings* and *section headings length ratio*, measure the similarity of how information is presented between the two documents. Meanwhile, the remaining four feature types, i.e., *links overlap*, *char-n-gram overlap*, *word overlap* and *word length ratio*, measure the content similarity between two articles.

9.2.2 Feature engineering

The author further identified a number of features for each feature type, as described in the following:

1. **Section length ratio**: The first feature measures the ratio of section lengths in both documents. Since not all Wikipedia pairs contain section headings, three binary features are extracted to capture whether section headings exist in the source document (`SourceSectionExists`), target document (`TargetSectionExists`), or both

(`BothSectionsExist`). These contain 1 if they exist, or 0 if they do not. The following features are also calculated to capture this feature type: the number of sections in the source document (`SourceSectionLength`), the number of sections in the target document (`TargetSectionLength`), the absolute difference between number of sections in both documents (`SectionLengthDifference`) and the ratio of section length with respect to the document with the larger number of sections (`SectionLengthRatio`). *[7 features]*

2. **Section heading similarity**: This feature group measures the similarity between section headings in both articles using the section heading similarity method described in Chapter 8. Firstly, the non-English section heading titles are translated into English using Wikipedia and Wiktionary (where available). All possible pairs of section headings between document pairs are created, prior to aligning similar section headings (based on Jaccard similarity of word overlap between the section headings). The scores are then aggregated using three different algorithms: `align1` representing the proportion of aligned section headings, `align2` representing the average similarity score of aligned section headings, and `align3` representing the proportion of aligned section headings whose similarity scores are above a certain threshold. *[3 features]*

3. **Links overlap**: Wikipedia articles are enriched with links to articles in the same languages. E.g., an article about “Barack Obama” contains links to Wikipedia articles about “Honolulu” (where he was born), “Harvard Law School” (where he studied), and “List of Presidents of the United States.” Link overlap is computed using the link based lexicon approach (Adafre & de Rijke, 2006). Firstly all links in the body of the Wikipedia article are extracted. Given document d_1 and d_2 written in languages l_1 and l_2 respectively, the extracted wiki links are represented as W_1 and W_2. A Wikipedia-based bilingual lexicon is created by listing titles of the interlanguage-linked article pairs and using them as translation resources. This lexicon is then used to translate all links in W_1 from l_1 to l_2, prior to measuring its similarity to

W_2 (written in L_2). The link overlap score is calculated using Jaccard coefficient of the links (`Links_BinaryJaccard`), Jaccard coefficient of the frequency of links (`Links_Jaccard`) and the cosine similarity of the term frequency (TF) of the links (`Links_TF`). Cosine similarity of links is also calculated using the TF-IDF scores. In this study, the author computed TF and IDF scores using the formulas shown in Equations 2.2 and 2.3, respectively. Since the availability of corpora to calculate the IDF scores may be limited for some under-resourced languages, this study investigates the use of different corpora with varying sizes to calculate the IDF scores. Three IDF scores are calculated using three different corpora: a corpus of 100 document pairs written in the language pair, a corpus of 10K document pairs written in the language pair, and a larger corpus of 50K document pairs written in the language pair. These features are referred to as `Links_TFIDF_100Corpus`, `Links_TFIDF_10KCorpus` and `Links_TFIDF_50KCorpus`, respectively. *[6 features]*

4. **Character *n*-gram overlap**: This feature group explores the use of char-*n*-grams (McNamee & Mayfield, 2004) to capture the cross-lingual similarity between Wikipedia document pairs (n=[2,3,4]). To extract the features, firstly, non-Latin alphabet documents (such as Greek) are transliterated, followed by the removal of diacritics, case folding and removal of punctuation marks. The author extracts the char-*n*-gram overlap using the Jaccard coefficient of unique char-n-grams (C*n*G_BinaryJaccard), Jaccard coefficient of the frequency of the words (C*n*G_Jaccard), the cosine similarity based on term frequency (TF) of the *n*-grams (C*n*G_TF). Similar to the previous feature, the cosine similarity of the TF-IDF scores are also calculated using three different corpora (described above), resulting in the following features: C*n*G_TFIDF_100Corpus, C*n*G_TFIDF_10KCorpus, and C*n*G_TFIDF_50KCorpus. *[18 features]*

5. **Word overlap**: This feature group computes the overlap of words (such as shared numbers or named entities). No translation is performed for this feature group; therefore, document pairs will only achieve a score more than 0 if any part of the

contents contain an exact overlap. Features are extracted using Jaccard similarity and cosine similarity, resulting in the following features: `WordOverlap_BinaryJaccard`, `WordOverlap_Jaccard`, `WordOverlap_TF`, `WordOverlap_TFIDF_100Corpus`, `WordOverlap_TFIDF_10KCorpus`, and `WordOverlap_TFIDF_50KCorpus`. *[6 features]*

6. **Word length ratio**: The following features are extracted to capture the difference in the word length between both articles: the number of words contained in the source document (`SourceLength`), the target document (`TargetLength`), the length difference in words (`LengthDifference`) and the ratio of number of words with respect to the longer document (`LengthRatio`). *[4 features]*

In summary, 44 different features are extracted from each document pair. The majority of these features can be extracted directly from the Wikipedia article content without requiring further linguistic resources. The only features requiring an additional resource are the section similarity features (`align1`, `align2`, `align3`) as they make use of Wikipedia and Wiktionary resources, which are available in a large number of languages. The list of the features and the resources they need are summarised in Table 9.1.

9.3 Experiment setup

9.3.1 Selection of languages

This experiment was carried out in eight language pairs: German (DE), Greek (EL),[1] Estonian (ET), Croatian (HR), Lithuanian (LT), Latvian (LV), Romanian (RO) and Slovenian (SL), which were all paired to English (EN).

[1] Greek transliteration process was performed using ILSP Transliteration Tool, which is available in: `http://nlp.ilsp.gr/soaplab2-axis/`, accessed on 2 April 2019.

Table 9.1 Features used to capture similarity with their required resources

Feature types	**Features**	**Required resources**
Section length	SourceSectionExists	-
	TargetSectionExists	-
	BothSectionsExist	-
	SourceSectionLength	-
	TargetSectionLength	-
	SectionLengthDifference	-
	SectionLengthRatio	-
Section similarity	align1	Wikipedia, Wiktionary
	align2	Wikipedia, Wiktionary
	align3	Wikipedia, Wiktionary
Links overlap	Links_BinaryJaccard	Wikipedia
	Links_Jaccard	Wikipedia
	Links_TF	Wikipedia
	Links_TFIDF_100Corpus	Wikipedia
	Links_TFIDF_10KCorpus	Wikipedia
	Links_TFIDF_50KCorpus	Wikipedia
Char-n-gram overlap (n=[2,3,4])	CnG_BinaryJaccard	-
	CnG_Jaccard	-
	CnG_TF	-
	CnG_TFIDF_100Corpus	Wikipedia*
	CnG_TFIDF_10KCorpus	Wikipedia*
	CnG_TFIDF_50KCorpus	Wikipedia*
Word overlap	Word_BinaryJaccard	-
	Word_Jaccard	-
	Word_TF	-
	Word_TFIDF_100Corpus	Wikipedia*
	Word_TFIDF_10KCorpus	Wikipedia*
	Word_TFIDF_50KCorpus	Wikipedia*
Word length	SourceLength	-
	TargetLength	-
	LengthDifference	-
	LengthRatio	-
* In this study, Wikipedia is used as a corpus from which IDF is computed. However, any corpus can be used.		

9.3.2 Dataset

The author utilised the Wikipedia similarity corpus (described in Chapter 5) as the training and evaluation dataset. In each of the 8 language pairs, two annotators have assessed 100 document pairs to annotate four aspects of the document pairs, namely the similarity (Q1), the proportion of overall document contents (Q2), the sentence similarity between the shared content (Q3) and the comparability level (Q4); each was evaluated using a 5-point Likert Scale. The author investigates the performance of the classification approach mainly to identify similarity (Q1), however, the classifier's performance in predicting other aspects (Q2-Q4) is also discussed. For each document pair, the mean scores between the assesssor's annotations were used in the evaluation dataset.

As previously described in Section 5.5.1 (specifically Figure 5.5 in page 114), this dataset contains document pairs with varying degrees of similarity. The number of document pairs with low similarity scores, however, is considerably lower than those with higher similarity scores. This causes the size of the classes used in the classification experiments to be imbalanced, with the lowest similarity degree class to be the minority. The imbalanced dataset may cause issues as the classifier is not able to properly learn the characteristics of the minority class due to the lack of training instances (Japkowicz & Stephen, 2002).

Therefore, in this experiment, the author explores the classification performance using three different datasets:

1. *Dataset 1: Original dataset.* In this case, 800 document pairs were used in the 10-fold cross-validation. No other data were used in the training and evaluation. This dataset is used for all of the experiments.

2. *Dataset 2: Original dataset + SMOTE (for training only).* SMOTE (Synthetic Minority Over-sampling Technique) is a technique to create new instances based on the existing instances and their nearest neighbours (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). First, for each minority instance, the five nearest neighbours were selected. Afterwards, a subset of neighbours were randomly chosen from this set, and a new

instance was created along the distance between the original minority instance and each of the chosen neighbours. The proportion of the instances added using the SMOTE approach varies between the different classifications; the number of added instances is decided empirically to achieve a more balanced dataset whilst limiting the risk of over-fitting. For the binary classification across all languages, SMOTE was used to create 50% more instances in the minority class, while for three-class classification, SMOTE was used to add 200% more instances in the minority class.The added instances were used only in the training dataset in the 10-fold cross validation. Only the original instances (800 document pairs) were used in the test dataset.

3. *Dataset 3: Original dataset + additional data (for training and testing).* In this experiment, for each language pair, the author increased the size of the minority class by adding 50 new document pairs that describe the different topics. These pairs were created by randomly pairing documents from the original dataset for each language pair, ensuring that they did not describe the same topic. Since these document pairs describe different topic, it is unlikely that they share similar contents within them. Therefore, these document pairs could be used as more training data for the ‘non-similar’ document pairs set. In total, 400 new document pairs were added into the dataset as ‘non-similar’ document pairs. In total, there are 1,200 document pairs in the dataset that were used in the 10-fold cross-validation for both binary and three-class classification. In the five-class classification, however, fewer ‘non-similar’ document pairs (12 document pairs per language pair, resulting in a total of 96 document pairs) were added into the dataset for the five-class classification in order to avoid over-fitting. A total of 896 document pairs are used in this experiment.

9.3.3 Classifiers

Different machine learning algorithms were explored in this study: Naïve-Bayes, Decision Tree, Random Forest, Neural Networks (Multilayer Perceptron) and SVMs. These algorithms were selected based on their uses in previous studies (Chen, Huang, Tian, & Qu, 2009; S. Dumais, Platt, Heckerman, & Sahami, 1998; Joachims, 1998; S. Lee & Choeh, 2014; Wäschle & Fendrich, 2012; Xu, Guo, Ye, & Cheng, 2012). Each classifier was tested using all 46 features. Missing features were left empty and no manual parameter tuning was carried out. The author used the WEKA machine learning toolkit[2] (version 3.6) for classification. The performance of the classifiers are investigated for solving different classification problems: binary classification, multi-class classification, and a regression, further described below.

Binary classification

The first classification problem is a binary classification to determine whether a document pair is 'similar' or 'not similar'. The document pairs were grouped into two classes based on the average similarity score from the two human annotators: 'similar' (score ≥ 3.5) and 'not similar'(score ≤ 3). In total, 536 document pairs (67%) were assigned to be 'similar', and 264 (33%) document pairs were assigned to be 'non-similar'. The distribution of similar and non-similar documents for each language pair in the evaluation dataset is shown in Figure 9.1.

3-class classification

The second approach is to build a *multi-class classifier* that categorises document pairs into three different levels of similarity. In this classification problem, the author divided the document pairs into three levels of similarity based on the mean similarity scores: 'high similarity' ($score \geq 4$), 'low similarity' ($2 < score < 4$), or 'no similarity' ($score \leq 2$). These thresholds result in 420 document pairs of 'high similarity' (53%), 307 'low

[2] http://www.cs.waikato.ac.nz/ml/weka/

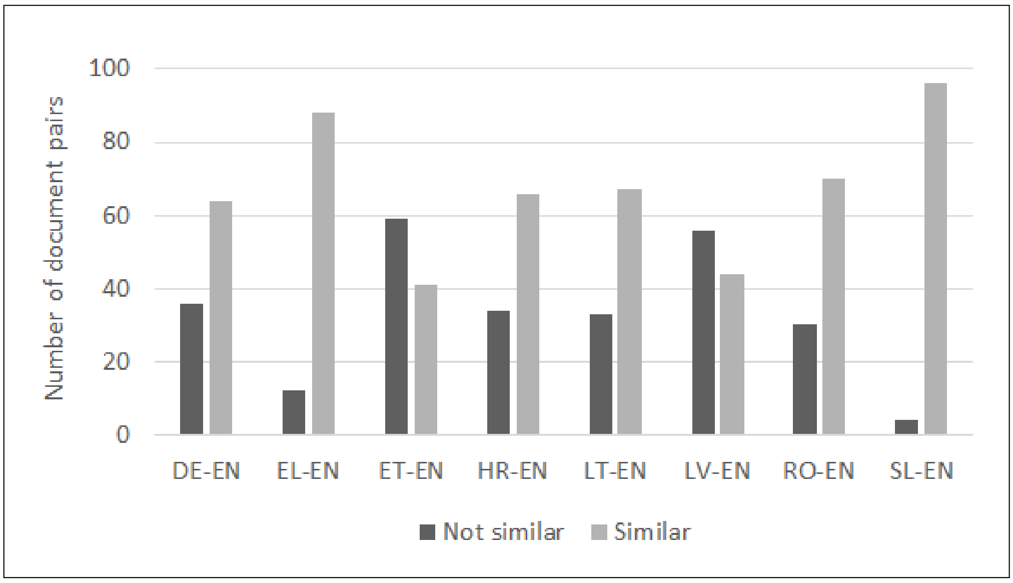


Fig. 9.1 Distribution of similar and non-similar documents (N=100 for each language pair)

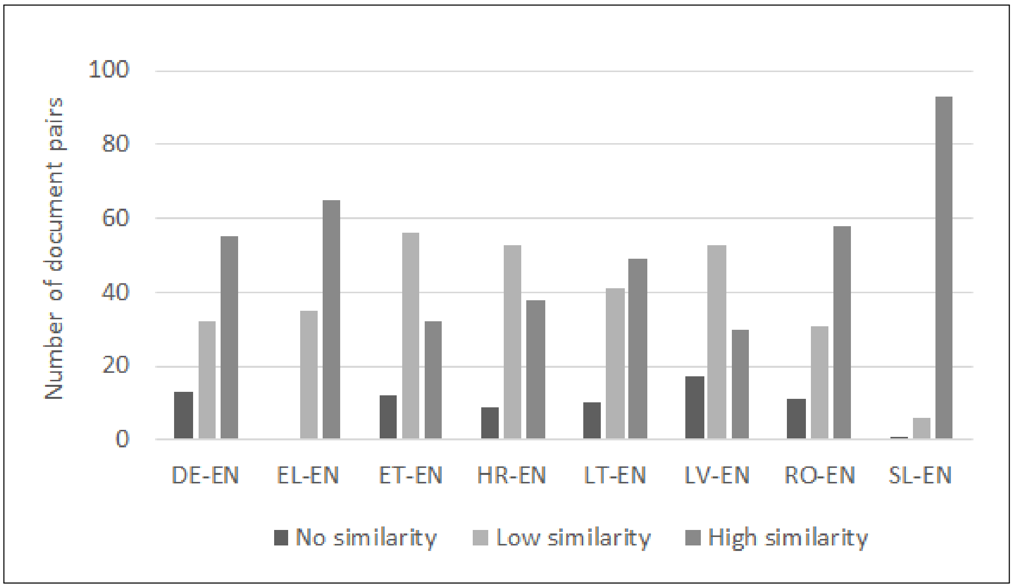


Fig. 9.2 Distribution of document pairs for three similarity groups (N=100 for each language pair)

similarity' pairs (38%) and 73 document pairs with 'no similarity' (9%). Figure 9.2 shows the distribution of these three classes in each language pair.

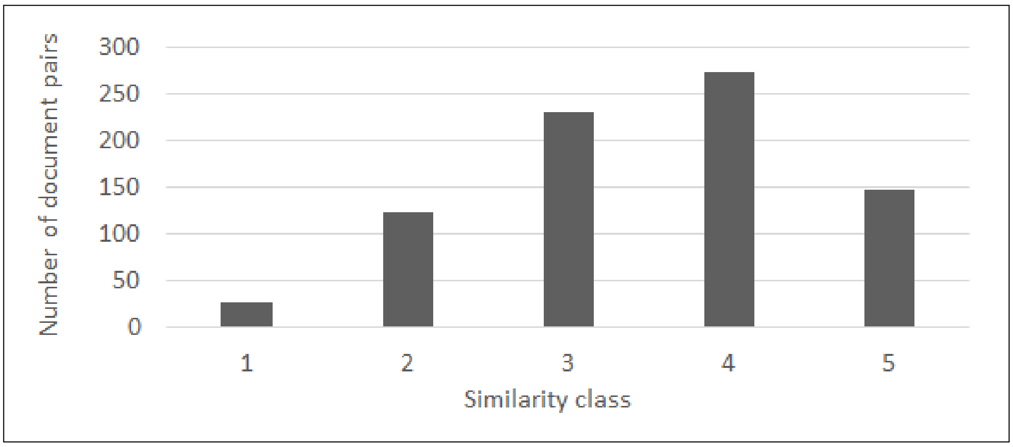


Fig. 9.3 Distribution of document pairs for five similarity classes (all language pairs, N=800)

5-class classification

The classifier's performance is also investigated in a *5-class classification* problem (1='not similar' and 5='very similar'). In this experiment, the document pairs are classified into five classes by rounding down the average score for each document pair to the nearest integer. This results in 147 document pairs of score 5 (18%), 273 document pairs of score 4 (34%), 230 document pairs of score 3 (29%), 123 document pairs of score 2 (15%), and 27 document pairs of score 1 (3%); a summary is shown in Figure 9.3.

Regression

Finally, the last problem is to predict the degree of similarity of the document pair using regression algorithms. Instead of classifying document pairs into a distinct number of classes, the regression algorithm predicts a continuous output (of the similarity score). To train the classifier, the average score given by the two annotators were used as the expected similarity score. The distribution of the similarity scores for all document pairs are shown in Figure 9.4.

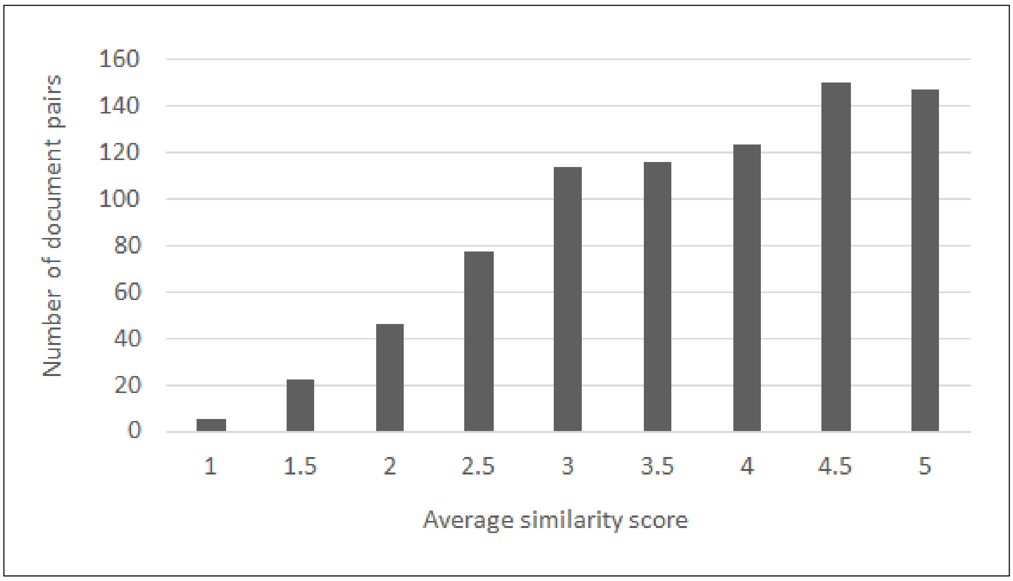


Fig. 9.4 Distribution of document pairs for all average similarity scores (all language pairs, N=800)

9.3.4 Evaluation

The classifier is evaluated using 10-fold cross validation using a number of metrics since they capture different aspects of performance. First, the author reported the classification accuracy (i.e., the percentage of correctly classified instances). Unfortunately, due to the imbalanced dataset (discussed in Section 9.3.3), classifiers may show high accuracy scores although the performance in identifying the minority class is poor. Therefore, the author also reported the F-measure score for each class (also known as the F_1 score); this score represents the harmonic mean of its precision and recall. The unweighted macro-average F-measure for all classes is also evaluated to capture the overall performance in classifying all classes, treating all classes (e.g., 'similar' and 'non-similar') with the same importance. AUROC (Area Under Receiver Operating Characteristic) is also reported, where the metric varies between 0.50 (random classifier) to 1.0. Values above 0.80 are considered good classification. Finally, for evaluating the five-class classifier and regression, the author also reported the Root Mean Squared Error (RMSE) scores.

Two baselines were used for this study: one incorporating a *language-independent feature* (`Baseline_LI`), and one incorporating a *language-dependent feature* (`Baseline_LD`):

1. `Baseline_LI`: The first baseline utilises the `LengthRatio` feature. This feature was chosen as it was shown to be the language-indepedent feature with the highest correlation to Q1 scores (Chapter 7). Logistic Regression is utilised as the algorithm for the language-dependent classifier, and Linear Regression algorithm is utilised as the regression baseline.

2. `Baseline_LD`: The second baseline incorporates a feature that requires a language-dependent resource. In this case, Google Translate[3] was utilised to translate the documents into the same language (i.e., in this case, non-English documents were translated into English), prior to measuring the cosine similarity of TF-IDF of the word overlap. Logistic Regression is utilised as the algorithm for the language-dependent classifier, and Linear Regression algorithm is utilised as the regression baseline.

Statistical significance is evaluated against the baseline using a two-tailed paired t-test with correction ($p<0.05$).

Finally, a feature selection was also performed to investigate the contribution of each feature to the classification approach. Four different approaches were used to identify these features, by i) calculating the information gain between different metrics, ii) computing the correlation between each feature and similarity, iii) finding the most useful subset of attributes, and iv) analysing the weights of features used in regression algorithms (Guyon & Elisseeff, 2003; C. Lee & Lee, 2006).

9.4 Results across all language pairs

9.4.1 Binary classification

Firstly, the performance of the binary classifier was evaluated across all language pairs using 10-fold cross validation on the entire dataset (800 annotated document pairs). Af-

[3] The evaluation documents were translated using Google Translate in February 2012. By the time this study was carried out in August 2018, this service had become a fully-paid service, and therefore, no new translations were carried out.

Table 9.2 Binary classification results (all language pairs; Dataset 1: 800 document pairs)

Classification	**Features**	**Accuracy**	**F-measure**[1]	**AUROC**
Baseline_LI	LengthRatio	74% (5.16%)	0.68 (0.06)	0.74 (0.07)
Baseline_LD	MT-WordOverlap[2]	71.63% (3.34%)	0.63 (0.05)	0.74 (0.05)
Random Forest	44 features	**81.38%*** (4.02%)	**0.79*** (0.05)	**0.87*** (0.05)
Logistic Regression	44 features	79.25%* (4.65%)	0.76* (0.05)	0.85* (0.05)
Multilayer Perceptron	44 features	79.38% (4.30%)	0.77 (0.06)	0.84* (0.05)
SVM	44 features	79.63%* (5.68%)	0.77* (0.06)	0.77 (0.06)

[1] Unweighted macro-average F-measure

[2] TF-IDF word overlap of the translated D1 (into English) and D2 (originally written in English). IDF was computed using a corpus of 50K EN documents.

* Statistically significant ($p<0.05$) compared to Baseline_LI and Baseline_LD.

terwards, the binary classifier's performance was investigated for the different language pairs.

Using the 44 features listed in Table 9.1, classifiers with different algorithms were trained and evaluated. The results show that the Random Forest classifier outperformed all other algorithms for all evaluation metrics (see Table 9.2). Its performance is also significantly higher than both baselines; it correctly classified 81% of document pairs as similar or non-similar, and achieved an unweighted macro-average F-measure=0.78 and AUROC=0.87. The Random Forest classifier performed better in classifying similar document pairs than non-similar document pairs (F-measure=0.86 for 'similar' class and F-measure=0.70 for 'non-similar'). The confusion matrix, shown in Table 9.3, further shows that the classifier was able to classify 87.31% 'similar' document pair correctly, compared tco classifying 68.18% 'non-similar' documents correctly.

The results above show that the binary classifier shows a promising result in classifying similar and non-similar document pairs in Wikipedia. However, the accuracy in identify 'non-similar' document pairs is poorer. A reason for this is the low proportion of 'non-similar' document pairs in the dataset. Addition of more 'non-similar' document

Table 9.3 RandomForest binary classification: confusion matrix (Dataset 1: 800 document pairs)

	Classified as		F-measure
	Non-similar	Similar	
Non-similar	180 (68.18%)	84 (31.82%)	0.70
Similar	68 (12.69%)	468 (87.31%)	0.86

pairs into this dataset will allow the classifier to learn the characteristics of these document pairs in order to be able to classify them better. In this experiment, the author tested two approaches to increase the number of 'non-similar' document pairs into the dataset.

The first approach is to use SMOTE to add 50% new data into the 'non-similar' classes in the training process (i.e., Dataset 2). The evaluation set remains the same (i.e., 800 document pairs). Using 10-fold cross-validation, the results show that the classifier was able to classify 'non-similar' document pairs better (i.e., correctly identifying 75.38% 'non-similar' document pairs, compared to 68.18%) and achieving an F-measure of 0.73 (compared to 0.70). Its performance in classifying 'similar' document pairs slightly decreased to 83.96%, compared to 87.31%; the F-measure, however, remains as high at 0.86.

The second approach is to add a new set of document pairs both in the training and testing dataset (i.e., Dataset 3) by pairing documents that describe different topics. For each language, 50 pairs of documents were added into the minority class, resulting in 400 new document pairs added into the dataset as 'non-similar' document pairs, bringing the data to be more balanced: 664 'non-similar' document pairs, compared to 536 'similar' document pairs. In total, there are 1200 document pairs in the dataset. The Random Forest classifier was re-trained using the new dataset and evaluated using 10-fold cross-validation. The results (Table 9.4) show that the additional of the new data was able to increase the classifier's performance in classifying non-similar data, accurately classifying 85.54% all non-similar document pairs (compared to 68.18%), and increasing the F-measure to 0.87 (compared to 0.70). Overall, the classifier achieves an unweighted macro-average F-measure of 0.86 and classifies 86% document pairs accurately.

Table 9.4 RandomForest binary classification: confusion matrix (Dataset 3: 1200 document pairs)

	Classified as		**F-measure**
	Non-similar	**Similar**	
Non-similar	568 (85.54%)	96 (14.46%)	0.87
Similar	72 (13.43%)	464 (86.57%)	0.85

9.4.2 Three-class classification

Different algorithms were compared to solve the three-class classification problem and the results show that, similar to the binary classification problem, the Random Forest classifier performed best (see Table 9.5). The confusion matrix (Table 9.7) shows that the classifier was able to identify the majority of ‘low similarity’ and ‘high similarity’ cases (F-measure=0.65 and 0.80, respectively). However, it achieved much lower F-measure in identifying document pairs with ‘no similarity’ (F-measure=0.20). Due to the small amount of training data for this class (73 out of 800 document pairs), 80.82% of the ‘no similarity’ document pairs were misclassified into the ‘low similarity’ class instead.

Using SMOTE to add more training data for the minority class (‘no similarity’) was able to improve the classifier’s performance. The unweighted macro-averaged F-measure

Table 9.5 3-class classification results (all language pairs)

Classification	**Features**	**Accuracy**	**F-measure**[1]	**AUROC**
`Baseline_LI`	`LengthRatio`	63.50% (2.69%)	0.43 (0.02)	0.68 (0.12)
`Baseline_LD`	`MT-WordOverlap`	59.25% (4.50%)	0.39 (0.04)	0.77 (0.07)
Random Forest	44 features	**70.13%*** (4.06%)	0.53* (0.06)	0.84 (0.07)
Logistic Regression	44 features	66.00% (4.71%)	**0.56*** (0.05)	**0.85** (0.07)
Multilayer Perceptron	44 features	60.63% (6.24%)	0.50 (0.07)	0.79 (0.07)
SVM	44 features	66.63% (5.30%)	0.46 (0.04)	0.74 (0.12)

Note: Standard deviation scores are shown in brackets.

[1] Unweighted macro-average F-measure

* Statistically significant (p<0.05) compared to `Baseline_LI` and `Baseline_LD`.

Table 9.6 RandomForest 3-class classification: confusion matrix (Dataset 1)

	Classified as			**F-measure**
	No similarity	**Low similarity**	**High similarity**	
No similarity	9 (12.33%)	59 (80.82%)	5 (6.85%)	0.20
Low similarity	8 (2.61%)	211 (68.73%)	88 (28.66%)	0.65
High similarity	0 (0%)	75 (17.86%)	345 (82.14%)	0.80

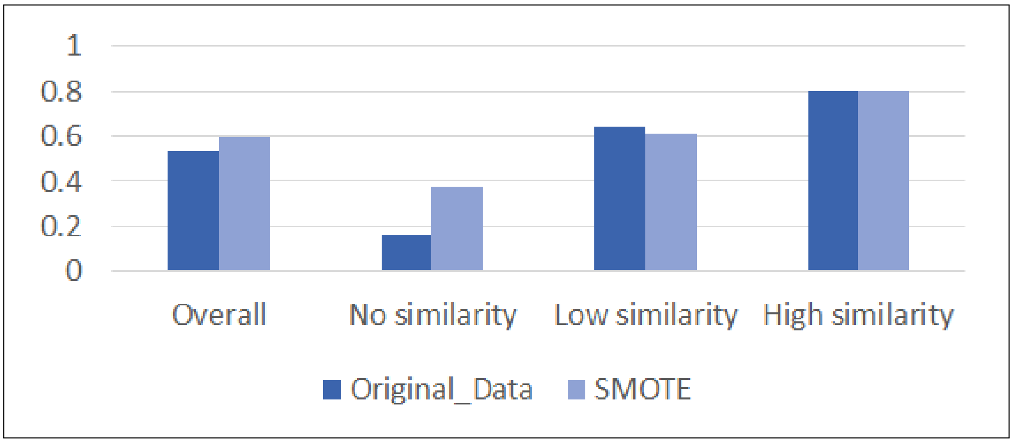


Fig. 9.5 RandomForest 3-class classification with SMOTE (F-measure)

increased to 0.60, compared to using the original dataset (F-measure=0.53; statistically significant, p<0.05). The results for a Random Forest classifier trained on the original data (without SMOTE) and the populated data with SMOTE, shown in Figure 9.5, show that the additional instances in the 'no similarity' training data were able to significantly increase the F-measure of this class (F-measure=0.38), without a significant decrease in the F-measure scores for the remaining two classes.

The author also evaluated the classifiers' performance using the extended dataset (Dataset 3), containing 1,200 document pairs. The results, shown in Table 9.7, show that the additional 'non-similar' document pairs enabled the classifier to perform significantly better in identifying these 'non-similar' documents (F-measure=0.91, compared to 0.20), without much loss in performance in classifying the remaining two classes (i.e., 'low similarity' and 'high similarity' classes).

Table 9.7 RandomForest 3-class classification: confusion matrix (Dataset 3: 1200 document pairs)

	Classified as			F-measure
	No similarity	**Low similarity**	**High similarity**	
No similarity	413 (87.31%)	52 (10.99%)	8 (1.69%)	0.91
Low similarity	14 (4.56%)	200 (65.15%)	93 (30.29%)	0.63
High similarity	4 (0.96%)	74 (17.62%)	342 (81.43%)	0.79

9.4.3 Five-class classification

In the five-class classification problem, each class represents the degree of similarity between article pairs (1='not similar' and 5='very similar'). This experiment utilised a meta-classifier[4] to allow the application of different classification algorithms for ordinal class problems (Frank & Hall, 2001). The results, reported in Table 9.8, show that the Random Forest classifier again outperforms all other algorithms.

Table 9.9 shows the confusion matrix for the five classes. It shows that the performance in classifying the lowest similarity class (Class 1) is very low, due the low availability of document pairs in this class (i.e., 27 document pairs only). In classifying Class 2-Class 4, the classifier was able to classify between 48.78%-62.64% document pairs into the correct classses. Only 35.37% document pairs of Class 5, however, was accurately classified, while over half of these document pairs were misclassified into Class 4. This indicates that the characteristics between these document pairs are very similar and that the classifier cannot differentiate between them. Furthermore, although the percentage of correctly classified instances seem lower than the binary and 3-class-classifier, when classifying documents of Class 2-Class 5, 88.62%-96.33% document pairs were classified to the correct classes or other classes differ by 1.

Due to the limited size of the minority class, SMOTE was not investigated in this approach in order to avoid overfitting the evaluation set. Instead, the author experimented with the additional of 'non-similar' document pairs in this 5-class classification. However, adding 400 document pairs will add more problems with the data imbalance, as the non-similar document pairs will be over-represented in the dataset. Therefore, the

[4] `OrdinalClassClassifier` in Weka

Table 9.8 Five-class classification results (all language pairs)

Classification	Features	Accuracy	F-measure[1]	AUROC	RMSE
`Baseline_LI`	`LengthRatio`	40.25% (2.75%)	0.24 (0.02)	0.62 (0.23)	0.37 (0.00)
`Baseline_LD`	`MT-WordOverlap`	39.88% (3.79%)	0.23 (0.03)	0.77 (0.12)	0.38 (0.00)
Random Forest	44 features	**50.88%*** (5.56%)	**0.42*** (0.07)	0.71 (0.11)	**0.35*** (0.01)
Logistic Regression	44 features	49.25% (7.08%)	**0.42*** (0.06)	**0.78** (0.09)	0.36 (0.01)
Multilayer Perceptron	44 features	45.75% (4.61%)	0.39* (0.07)	0.72 (0.12)	0.43* (0.02)
SVM	44 features	48.88%* (4.27%)	0.37* (0.04)	0.50 (0.00)	0.45* (0.02)

Note: Standard deviation scores are shown in the brackets.

[1] Unweighted macro-average F-measure

* Statistically significant ($p<0.05$) compared to `Baseline_LI` and `Baseline_LD`.

Table 9.9 Random Forest 5-class classification: confusion matrix (Dataset 1)

	Classified as					**F-measure**
	1	**2**	**3**	**4**	**5**	
1	2 (7.41%)	10 (37.04%)	11 (40.74%)	4 (14.81%)	0 (0%)	0.14
2	0 (0%)	60 (48.78%)	49 (39.84%)	14 (11.38%)	0 (0%)	0.50
3	0 (0%)	35 (15.22%)	122 (53.04%)	64 (27.83%)	9 (3.91%)	0.50
4	0 (0%)	10 (3.67%)	62 (22.71%)	171 (62.64%)	30 (10.99%)	0.56
5	0 (0%)	2 (1.36%)	13 (8.84%)	80 (54.42%)	52 (35.37%)	0.44

author added a smaller number of ‘non-similar’ document pairs into the dataset, i.e., 12 ‘non-similar’ document pairs for each language pair. This results in 96 ‘non-similar’ document pairs across 8 language pairs being added into the original 27 ‘non-similar’ document pairs available in the dataset, bringing the total of the Class 1 document pairs to 123 document pairs, the same number of instances as Class 2. The results are shown in Table 9.10.

The results show that the classifier was able to achieve a macro-average F-measure of 0.57 using Dataset 3, compared to 0.42 using the original dataset (Dataset 1). This suggest that the classifier can perform better if there were more data to train the different classes. The performance for Class 2-Class 4 were slightly poorer with the additional data. The

Table 9.10 Random Forest 5-class classification: confusion matrix (Dataset 3: 896 document pairs)

	Classified as					**F-Measure**
	1	**2**	**3**	**4**	**5**	
1	101 (82.11%)	8 (6.50%)	11 (8.94%)	3 (2.44%)	0 (0%)	0.87
2	4 (3.25%)	52 (42.28%)	54 (43.90%)	13 (10.57%)	0 (0%)	0.47
3	3 (1.30%)	29 (12.61%)	118 (51.30%)	74 (32.17%)	6 (2.61%)	0.49
4	0 (0%)	9 (3.30%)	58 (21.25%)	169 (61.90%)	37 (13.55%)	0.56
5	0 (0%)	2 (1.36%)	13 (8.84%)	77 (52.38%)	55 (37.41%)	0.45

F-measure for Class 5, however, was slightly higher (0.45 compared to 0.44), although majority of these instances were still mostly classified into Class 4.

9.4.4 Regression

Finally, the same set of features were investigated in a regression problem. For the baselines, two classifiers were developed using a Linear Regression algorithm; one used a language-independent feature (`LengthRatio`) and one used a language-dependent feature (`MT-WordOverlap`). Different regression algorithms are being explored in this study: Random Forest, Linear Regression, Multilayer Perceptron, SVM and K-Nearest Neighbours were investigated in this work. The results are shown in Table 9.11. In this scenario, Random Forest algorithm achieved the best performance (RMSE=0.73) and significantly outperformed both baselines. Similar performance was also achieved when using Linear Regression and SVM using all 44 features. The results also show that, although `LengthRatio` was previously shown to be useful in predicting similarity in binary classification problem, it is not sufficient when used solely in a regression problem.

9.4.5 Feature comparison

Findings suggest that language-independent features can be used to predict the degree of similarity between document pairs. The usefulness of these features, however, may be different to each other. This section investigates which features are the most important in classifying similarity in Wikipedia. The aim of this section is not to perform a feature

Table 9.11 Regression results (all language pairs)

Algorithm	**Features**	**RMSE**	**Pearson's Correlation Coefficient***
Baseline_LI (Linear Regression)	LengthRatio	0.89 (0.05)	0.47 (0.06)
Baseline_LD (Linear Regression)	MT-WordOverlap	0.90 (0.06)	0.45 (0.09)
Random Forest	44 features	**0.73*** (0.06)	**0.69*** (0.07)
Linear Regression	44 features	0.74* (0.08)	0.67* (0.08)
Multilayer Perceptron	44 features	1.15* (0.22)	0.51 (0.11)
SVM	44 features	0.74* (0.08)	0.67* (0.07)
K-Nearest Neighbour	44 features	1.05* (0.07)	0.44 (0.07)

Note: Standard deviation scores are shown in the brackets.
* Statistically significant (p<0.05) compared to Baseline_LI and Baseline_LD.

selection to improve the classifier, but to explore which features or a subset of features are the most useful in indicating Wikipedia similarity.

First, the author evaluated the information gain of the different metrics using 10-fold cross validation. This was applied to the binary, three-class and five-class classification problems. The following features were found to be the most useful in all three classification approaches: word overlap features (Words_BinaryJaccard and Words_Jaccard), followed by length ratio features (LengthDifference and LengthRatio). Character n-gram overlap features of all possible n=[2,3,4] were also found in the top 10 (C2G_BinaryJaccard, C2G_Jaccard, C3G_BinaryJaccard, C3G_Jaccard, C4G_BinaryJaccard, C4G_Jaccard). Finally, the SectionLengthDifference feature was rated as one of the top 10 useful features in identifying similarity in three-class and five-class classification problem, but did not feature in the top 10 features in solving the binary classification problem.

Feature analysis was also carried out by evaluating the usefulness of each feature by measuring the Pearson's correlation between each feature to the class. Similar results were found using this approach. In this case, LengthRatio, Words_BinaryJaccard and LengthDifference were rated as the top 3 features that most correlate with the class in all three classification problems. For regression, Words_BinaryJaccard, LengthRatio, and C4G_Jaccard, were shown to correlate the most. Interestingly, LengthDifference was not shown to have the least correlation with the average similarity scores in a regression problem. In both classification and regression, structure similarity features and link overlap were shown to have very weak correlation to the similarity scores of the document pairs.

The third analysis of features was performed by finding the most useful subset of attributes, taking into account the redundancy caused by interaction between features (Hall, 1998). In general, the following feature types: length ratio, character-n-gram overlap and word overlap were found to be useful for all the classification approaches. The resulting subset of attributes also include link overlap (Links_Jaccard) and section similarity (align2). For regression, six features were selected as a subset: LengthRatio, StructureLengthRatio, C4G_Jaccard, Links_BinaryJaccard, Links_Jaccard, and Words_BinaryJaccard.

Finally, the author analysed the weight of features of a Linear Regression algorithm by predicting the degree of similarity (i.e., in a regression problem). The results show that some features, such as the length of articles and its length difference (in number of words), length of section headings, were not found to be useful in predicting similarity. Meanwhile, LengthRatioSB, align2, C3G_Jaccard, C3G_TFIDF_10KCorpus, C4G_BinaryJaccard, C4G_TFIDF_100Corpus, Links_Jaccard, Words_BinaryJaccard and Words_TFIDF_50KCorpus are the features that are most positively weighted.

The author also investigated the performance of the classifier when using only features from each feature type. This is performed in a binary classification approach (Table 9.12) and regression approach (Table 9.13). In both cases, the results show that the char-4-gram overlap feature types perform best in predicting similarity, shortly followed by char-2-gram and char-3-gram overlap. In regression, word overlap feature types were also shown to perform best; significantly higher than both baselines.

Table 9.12 Binary classification results using each feature group (RandomForest)

Feature types	**F-measure**[1]	**AUROC**
Baseline_LI[2]	0.68	0.74
Baseline_LD[3]	0.63	0.74
Section length	0.62	0.69
Section similarity	0.46*	0.57*
Links overlap	0.55*	0.61*
Char-2-gram overlap	0.71	0.80
Char-3-gram overlap	0.69	0.78
Char-4-gram overlap	**0.72**	**0.80**
Word overlap	0.68	0.78
Word length	0.66	0.73

[1] Unweighted macro-average F-measure
[2] Logistic Regression of the length ratio feature
[3] Logistic Regression of the word overlap (TF-IDF) on the EN-translated contents
* Statistically significant ($p<0.05$) compared to Baseline_LI and Baseline_LD.

Table 9.13 Regression results using each feature group (RandomForest)

Feature types	**RMSE**	**Pearson's correlation coefficient**
Baseline_LI[1]	0.89	0.47
Baseline_LD[2]	0.90	0.45
Section length	0.92	0.42
Section similarity	1.00*	0.22*
Links overlap	0.99*	0.24*
Char-2-gram overlap	0.85	0.54
Char-3-gram overlap	0.85	0.54
Char-4-gram overlap	**0.83**	**0.56***
Word overlap	**0.83***	**0.56***
Word length	0.93	0.42

[1] Linear Regression of the length ratio feature
[2] Linear Regression of the word overlap (TF-IDF) on the EN-translated contents
* Statistically significant ($p<0.05$) compared to Baseline_LI and Baseline_LD.

9.4.6 Classification of other evaluation aspects

The author also investigated the classifier's performance in identifying other aspects of the documents that were annotated by the assessors, namely the proportion of similar sentences (Q2), the similarity between these aligned sentences (Q3), and the comparability score (Q4). In this experiment, Random Forest classification algorithm was utilised since it was shown to achieve the highest results in the previous experiments. The classifier's performance in classifying these aspects (Q2-Q4) are compared to the classifier's performance in identifying similarity (Q1) in a binary classification problem. The author used the same threshold that was used in the binary classification approach (see Section 9.3.3) to divide the dataset into two classes based on the mean annotation score for each aspect: Class 1 ($score \leq 3$) and Class 2 ($score > 3$). The results are shown in Table 9.14.

The results show that the classifier achieves the highest F-measure when used to classify similarity of the document pairs (unweighted macro-average F-measure=0.79), and the proportion of similar sentences in the documents (unweighted macro-average F-measure=0.79). Its performance in identifying comparability is significantly lower (unweighted macro-average F-measure=0.69) and it achieves the lowest average F-measure in identifying the similarity between the aligned sentences (unweighted macro-average F-measure=0.65). This result is to be expected since the classifier uses features extracted at the document level rather than the sentence level.

The author also investigated and compared the performance of the classifier in a regression problem. I.e., in this experiment, the average score given by the assessors are used as the expected score of the classifier. The Root Mean Squared Error (RMSE) for each evaluation aspect is reported in Table 9.15. When applied to solve a regression problem, the classifier achieved the best accuracy in predicting the comparability score (Q4)

Table 9.14 Binary classification results in identifying different evaluation aspects (Random Forest; Dataset 1)

Aspects	**Number of Docs**		**Accuracy**	**F-measure**			**AUROC**
	Class 1	**Class 2**		**Class 1**	**Class 2**	**Unweighted macro-average**	
Similarity (Q1)	264	536	81.38%	0.71	0.86	0.79	0.87
Proportion of similar sentences (Q2)	290	510	80.25%	0.72	0.85	0.79	0.87
Similarity of aligned sentences (Q3)	133	667	85.25%	0.39	0.92	0.65	0.83
Comparability (Q4)	544	256	76.38%	0.84	0.58	0.71	0.81

Table 9.15 Regression results in identifying different evaluation aspects (Random Forest; Dataset 1)

Evaluation Aspects	**RMSE**	**Pearson's Correlation Coefficient**
Similarity (Q1)	0.73	0.69
Proportion of similar sentences (Q2)	0.69	0.69
Similarity of aligned sentences (Q3)	0.73	0.61
Comparability (Q4)	0.56	0.65

(RMSE=0.56). Moderate to strong correlation scores were achieved in all the four evaluation aspects (Pearson's r range between 0.61 and 0.69). Again, the lowest performance was achieved in identifying similarity between the aligned sentences.

The results between the binary classification and the regression approach show contradictory results with regards to predicting the comparability aspect of the document. Whilst it achieves a poor performance in the binary classification result, its performance in predicting comparability in regression is much higher. These results are further discussed in Section 9.6.4.

9.5 Results for each language pair

The results described in the previous section have explored the performance of the classifier when trained and tested using the entire dataset, regardless of the language. In this section, the author analysed how the classification approach performs for each language pair. Due to the limited dataset (i.e., 100 document pairs for each language pair), this experiment is only performed for the binary classification problem and regression.

9.5.1 Binary classification

The author created two Random Forest classifiers that utilised different training datasets:

1. *Classifier 1* was trained and evaluated using document pairs from the evaluation dataset from the respective language pair (i.e., 100 document pairs). E.g., DE-EN

classifier is trained using DE-EN data, EL-EN classifier used EL-EN data, etc. 10-fold cross-validation was utilised in this evaluation setup.

2. *Classifier 2* was trained using document pairs from the evaluation dataset from all languages *except* the respective language. For example, a DE-EN classifier was built using dataset from the remaining 7 language pairs (EL-EN, HR-EN, LT-EN, LV-EN, RO-EN, and SL-EN), i.e., 700 document pairs. The performance is then tested on the DE-EN data (i.e., 100 document pairs). This experiment also analysed whether language-independent features trained on a set of language pairs can be applied to identify similarity in a different language pair.

The classifier's performance for each language pair is shown in Table 9.16.

In general, the classifier performs better at classifying 'similar' document pairs (mean accuracy=84.46%, F-measure=0.84) compared with 'non-similar' pairs (mean accuracy= 53.21%, F-measure=0.56). The highest performance for identifying 'similar' document pairs was achieved in EL-EN (accuracy=100%, F-measure=0.95); the lowest in LV-EN (accuracy=56.82%, F-measure=0.63). When identifying 'non-similar' document pairs, the classification performance varies significantly between languages. The best 'non-similar' performance is achieved in DE-EN (accuracy=89.93%, F-measure=0.87). Its ability to classify non-similar document pairs was particularly poor in two language pairs: EL-EN (accuracy=16.67%, F-measure=0.29) and SL-EN (accuracy=0%, and F-measure=0.00), due to the imbalanced dataset in these two languages (12 and 4 'non-similar' document

Table 9.16 Results for Classifier 1 (per language pair)

Lang pair	**Correctly Classified Instances**			**F-measure**			**AUROC**
	Similar	**Non-similar**	**Macro-Avg**	**Similar**	**Non-similar**	**Macro-Avg**	
DE-EN	93.75%	86.11%	89.93%	0.93	0.87	0.90	0.97
EL-EN	100%	16.67%	58.34%	0.95	0.29	0.62	0.87
ET-EN	70.73%	88.14%	79.44%	0.75	0.85	0.80	0.85
HR-EN	80.30%	58.82%	69.56%	0.80	0.60	0.70	0.82
LT-EN	86.57%	45.46%	66.01%	0.81	0.53	0.67	0.77
LV-EN	56.82%	80.46%	68.64%	0.63	0.75	0.69	0.74
RO-EN	88.57%	50%	69.29%	0.84	0.57	0.71	0.86
SL-EN	98.96%	0%	49.48%	0.97	0.00	0.49	0.97
Mean	84.46%	53.21%	68.84%	0.84	0.56	0.7	0.86

pairs respectively, see Figure 9.1). This resulted in the inability of the classifier to be properly trained to identify the 'non-similar' document pairs.

The author investigated the use of SMOTE to generate more instances in the minority classes. However, this did not improve the results, which was likely due to the very small number of instances in the dataset.

Classifier 2, on the other hand, was trained using data from 7 language pairs (i.e., not including the respective language pair). It allowed the classifier to learn from more instances, although they were not written in the same language pair. We compared the unweighted macro-averaged F-measure between Classifier 1 (trained on the language pair) and Classifier 2 (trained on other language pairs) in Figure 9.6. A comparison between F-measure for the similar and non-similar classes was also investigated (shown in Figure 9.7 and Figure 9.8, respectively).

Results show that, although not trained on the respective language pair, Classifier 2 was able to achieve better F-measure scores in 5 language pairs (EL-EN, HR-EN, LT-EN, LV-EN and SL-EN). Classifier 2 was also able to improve the ability to identify the minority class for language pairs with lack of training data for the minority class (see EL-EN, ET-EN and SL-EN in Figure 9.8). Results suggest that the classifier model trained from a set of languages can be used to identify similarity in a new language pair (i.e., a language pair that was not included in the training data).

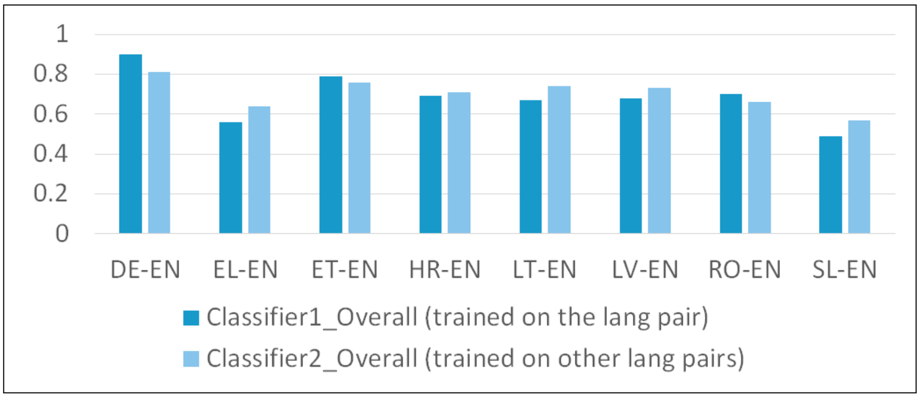


Fig. 9.6 Unweighted macro-average F-measure per language pair

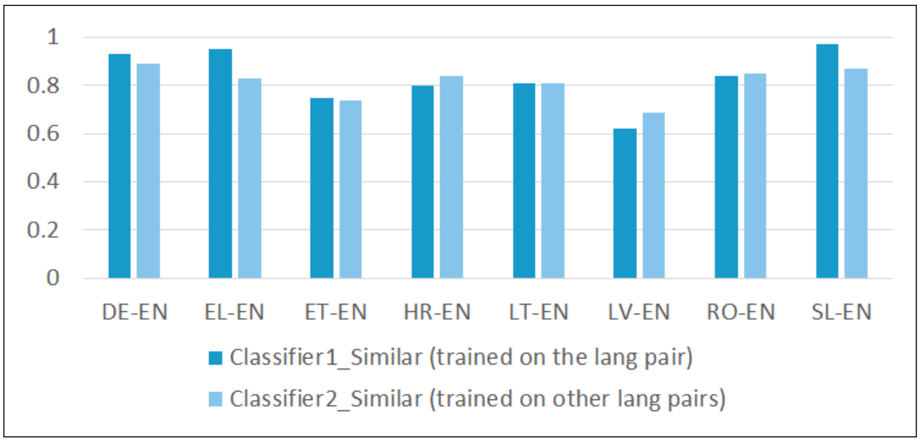


Fig. 9.7 F-measure for 'similar' document pairs per language pair

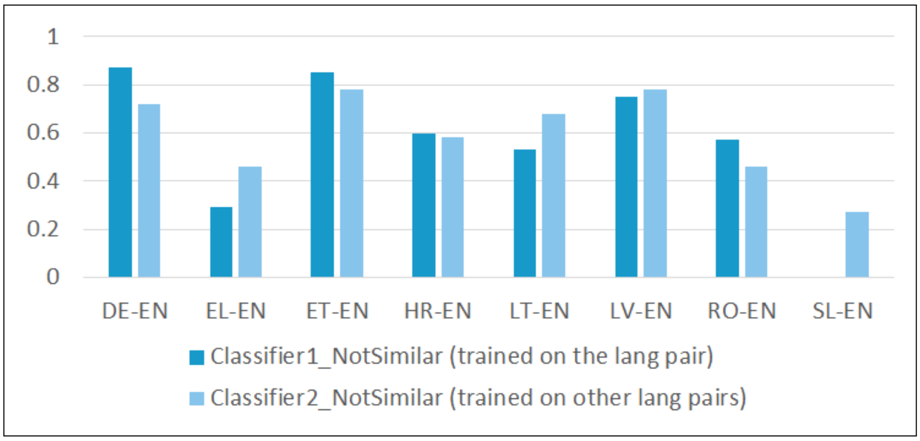


Fig. 9.8 F-measure for 'non-similar' document pairs per language pair

9.5.2 Regression

The performance of the language independent features in predicting similarity for each language pair was also evaluated in a regression problem. The results are shown in Table 9.17. The results show that the Random Forest algorithm achieves lower RMSE and stronger Pearson's correlation coefficient compared to both baselines, although only DE-EN results were found to be statistically significant ($p<0.05$). The RMSE for EL-EN and HR-EN for Random Forest and Baseline_LI were similar, suggesting that the language-independent features do not work much better than using the length ratio on its own; however, the results for the remaining language pairs show that language-independent features do achieve better results compared to using the length ratio. Furthermore, the

Table 9.17 Regression performance for each language pair (Dataset 1)

(a) Root Mean Squared Error (RMSE)

Language Pair	**Baseline_LI**	**Baseline_LD**	**Random Forest**
DE-EN	0.92 (0.22)	0.91 (0.15)	**0.53*** (0.08)
EL-EN	0.58 (0.12)	0.63 (0.14)	**0.55** (0.10)
ET-EN	0.68 (0.10)	0.68 (0.20)	**0.57** (0.13)
HR-EN	0.80 (0.21)	0.81 (0.22)	**0.76** (0.27)
LT-EN	0.81 (0.17)	0.71 (0.13)	**0.65** (0.14)
LV-EN	0.91 (0.19)	0.85 (0.21)	**0.79** (0.17)
RO-EN	0.94 (0.20)	1.07 (0.17)	**0.83** (0.20)
SL-EN	0.49 (0.21)	0.48 (0.15)	**0.40** (0.19)
* Statistically significant (p<0.05) to both baselines. *Note: standard deviation is shown in brackets.*			

(b) Pearson's Correlation Coefficient

Language Pair	**Baseline_LI**	**Baseline_LD**	**Random Forest**
DE-EN	0.47 (0.24)	0.48 (0.22)	**0.87*** (0.09)
EL-EN	0.51 (0.20)	0.37 (0.41)	**0.59** (0.18)
ET-EN	0.42 (0.19)	0.51 (0.33)	**0.62** (0.22)
HR-EN	0.56 (0.22)	0.50 (0.35)	**0.57** (0.30)
LT-EN	0.32 (0.34)	0.51 (0.21)	**0.61** (0.22)
LV-EN	0.40 (0.21)	0.43 (0.22)	**0.56** (0.28)
RO-EN	0.58 (0.27)	0.37 (0.28)	**0.62** (0.17)
SL-EN	0.24 (0.29)	0.40 (0.34)	**0.58** (0.33)
* Statistically significant (p<0.05) to both baselines. *Note: standard deviation is shown in brackets.*			

RMSE scores for 5 language pairs (DE-EN, EL-EN, ET-EN, LT-EN and SL-EN) were lower than the RMSE overall language pairs (0.73, see Section 9.4.4). Similar evaluation results was achieved by analysing the Pearson's correlation coefficient. However, in both cases, there is a big variance in the performance for each fold. This may be caused due to performing 10-fold cross validation on a small dataset (100 document pairs).

9.6 Discussion

In this work, the author has investigated the use of a number of language-independent features in classifying the degree of similarity of Wikipedia document pairs. These features were also tested for measuring other aspects of the documents, such as the proportion of similar sentences, similarity of the aligned sentences, and the comparability level of the document pairs. In this section, the author discusses these findings further.

9.6.1 Language-independent features

The features used in this classification study are simple features that are easy to extract. Most features can be extracted without the use of any linguistic resources. Whilst others benefit from multilingual information in Wikipedia and Wiktionary. One might argue that not all languages are available in Wikipedia. However, the author argues that Wikipedia is currently available in over 250 languages and this number is growing every day. Similarly, the size of Wiktionary is also growing daily. This means that the coverage of languages that the classifier will be able to process in the future will also expand with the amount of languages covered by Wikipedia and Wiktionary. Similarly, the high connectivity between Wikipedia articles in different languages also mean that the size of bilingual resources able to be utilised in this study is also increasing.

There are two language independent features that have been studied for the purpose of measuring similarity, as discussed in Chapter 2, which were not included in this work. The first one is Cross-lingual Explicit Semantic Analysis (Potthast et al., 2008; Sorg & Cimiano, 2008). There were a couple of reasons why the CL-ESA method was not investigated as a feature in this work. Firstly, to achieve a good performance, CL-ESA requires a bilingual parallel or comparable corpus containing at least 100,000 document pairs for each language pair to be used as the concept documents (Potthast et al., 2008). The corpus used in this study (i.e., Wikipedia corpus in 2010) contains fewer than 100,000 interlanguage linked document pairs for each of the under-resourced language pairs. The only language pair containing more than the required number was the highly-resourced language pair: German-English corpus.[5] Whilst CL-ESA have been used for measuring similarity in Wikipedia, the test and concept documents should be kept separate, which

[5] By 30 August 2018, all the language pairs investigated in this study contained more than 100,000 interlanguage-linked documents except Latvian-English (85,148 document pairs). However, these still indicate that the required resources limit the possible applications of CL-ESA.

introduces a problem if CL-ESA is to be used for measuring similarity overall Wikipedia. Furthermore, CL-ESA assumed that the concepts in Wikipedia documents "are described 'sufficiently exhaustive' for all languages" (Potthast et al., 2008, p. 524). Whilst these may be the case for Wikipedia versions in highly-resourced languages (due to their more advanced developments, both in quantity and quality, than the under-resourced languages), these may not be the case for under-resourced language pairs. Furthermore, the findings described in this study have further shown that the degree of similarity between these document pairs still vary widely, contradicting their assumptions about the high comparability of the corpus. Furthermore, it is unclear how the concept documents should be selected, how robust CL-ESA performs given different concept documents, or documents with different comparability. Finally, the work in CL-ESA was often carried out in a multilingual information retrieval task, i.e., utilising the CL-ESA score as means of ranking between documents of the same topic. Its use in measuring cross-lingual similarity between document pairs, however, has not been investigated before. Previous work has also not yet thoroughly studied the use of CL-ESA for under-resourced languages, possibly due to the limitations described before. Due to these issues and the limited time in the study, CL-ESA was not investigated as a feature for the classifier.

Another state-of-the-art feature for measuring similarity (mostly between words) are word embedding approaches, such as *word2vec* (Mikolov et al., 2013b). This approach utilises deep learning methods to learn representations of words and have shown good performance in various NLP problems. Although its ability to measure similarity or relatedness between words has been reported widely, its usage in measuring similarity between document content is still a developing research area. One approach is to use a document embedding, further known as *doc2vec* (Le & Mikolov, 2014). Although pre-trained document embedding models are available in English (monolingually), the author was not able to find any pre-trained bilingual document embedding suitable for use in this study. Pre-trained bilingual word embeddings (using *word2vec* method) are available for some languages and there are available tools to create these embeddings for more languages given the availability of bilingual corpora. However, how this information should be aggregated or combined to measure similarity at the document level has not been investigated yet. Due to the these limitations, the author decided that utilising word embedding and/or document embeddings is a field of work on its own and it was not an avenue that could be explored within the scope and the timeline of this study. Future work, however, should explore this topic once this approach has been shown to measure document similarity as well as word similarity.

9.6.2 Comparison of performance against MT

This study reports how a classifier supported by language-independent features can achieve significantly better performance compared to using a language-dependent feature, i.e., utilising a machine translation system to translate the non-English contents into English, prior to measuring the TF-IDF of word overlap. One possible reason why the translation method performed poorly is due to the poor translation quality for the under-resourced languages at that time; the work presented in this study utilised Google Translate to translate all the non-English documents into English in 2012. It is likely that the translation quality has since improved significantly for these language pairs. However, by the time this document was written, Google Translate was not available as a free service anymore. The author was not able to compare the translation qualities between the 2011 version and the most current version (2018), nor to calculate the corresponding performance of the classifier using the current Google Translate version. However, this highlights the issues surrounding under-resourced languages, i.e., the limited availability of a good quality translation system.

In some cases, the use of Google Translate was shown to perform worse in document pairs that already contain a high word overlap. The evaluation dataset contains a number of documents which contained contents written in the same language. E.g., one of the Estonian-English document pairs listed names of Estonian schools (both written in Estonian). After applying MT, the school names in Estonian documents were translated into English, and therefore reduced the word overlap scores between the English document and the translated Estonian document.

9.6.3 Identification of similar features

The results reported in this chapter have explored the usefulness of features in identifying similarity. When the performance of different features is analysed in different classification problems, the results show that a subset of features (i.e., *word overlap* features, and *char-n-gram overlap* features) were found to be useful for all types of classification problems.

Simple word overlap features were shown to perform very well for identifying similar documents in Wikipedia. This finding is surprising because the author expected that article content from different languages would have a small overlap of the same words. After checking the dataset, however, the author identified that there were some documents that contained the same words in each languages. E.g., an article pair about “Brown Township” in both German and English contains a list of a city named “Brown Township” in all American states. An extract of both document contents are shown in Table 9.18.

This document shows that Wikipedia contains articles that may contain overlapping words, due to the number of named entities available in both languages. Another examples found in the dataset is an article about “Michel Creton”, a French actor. Both articles in English and German contain a list showing selected filmography featuring him. Both these sections, as a result, contain a list of movies with the French titles, in both the English and German article.

The high number of word overlap in these documents does cause some bias in the evaluation dataset (compared to Wikipedia in general). This bias is caused by the way the evaluation documents were sampled. As discussed in Chapter 5, the selection of documents for the evaluation corpus was not performed randomly. Instead, it was a stratified sampling of documents with varying degree of links and word overlap. This approach forced a higher proportion of Wikipedia articles with higher degrees of word overlap to be included in the evaluation corpus, compared to Wikipedia in general. At the time, there was no other language-independent approach that could have been used to identify similarity in Wikipedia across multiple languages (hence the motivation to do this study). This approach was chosen as it was a lightweight method and was shown to be suitable to select document pairs of different similarity degrees. However, this approach has caused the feature distribution of this evaluation set to differ to the overall Wikipedia articles. In the future, a better evaluation set is needed to include more document pairs that better represent Wikipedia to reduce the word overlap bias.

Another feature that was found to be useful in classifying document pairs is the *section length difference* feature. This feature was found to contribute better in classifying document pairs into more classes of similarity (3 classes and 5 classes), however, was not found to be very useful in classifying document pairs into two classes. Meanwhile, *section similarity* features were not found to be very useful for identifying similarity when used individually. One possible reason for this is that the size of the evaluation corpus limits the usability of this feature. In the evaluation set, only short article pairs (mostly fewer than 1,000 words) were used. As a result, in 105 document pairs (13.13%), at least

Table 9.18 An extract of articles with high word overlap (Example 1)

DE document	**EN document**
Brown Township	**Brown Township**
“Brown Township” ist der Name mehrerer Townships in den Vereinigten Staaten: * Brown Township (Clay County, Arkansas). * Brown Township (Monroe County, Arkansas). * Brown Township (Champaign County, Illinois). * Brown Township (Illinois). * Brown Township (Hancock County, Indiana). * Brown Township (Hendricks County, Indiana). * Brown Township (Montgomery County, Indiana). * Brown Township (Morgan County, Indiana). * Brown Township (Ripley County, Indiana). * Brown Township (Washington County, Indiana). ...	“Brown Township” may refer to the following places in the United States: **Arkansas** * Brown Township, Clay County. Arkansas. * Brown Township, Monroe County, Arkansas. **Illinois** * Brown Township, Champaign County, Illinois. **Indiana** * Brown Township, Hancock County, Indiana. * Brown Township, Hendricks County, Indiana. * Brown Township, Montgomery County, Indiana. * Brown Township, Morgan County, Indiana. ...

one of the articles did not have any section headings to be compared. In the remaining ones (695 document pairs), on average, each article has 5.04 section headings (SD=4.04). However, the number of sections vary widely (min=1 and max=49). Having a full score of 1 in section similarity features mean that all the section headings in both articles could be aligned to each other. However, this does not differentiate cases where article pairs contain a small number of section headings, compared to those with a higher number of section headings. E.g., consider two document pairs, i) this document pair only has 1 section each which is aligned together, or ii) this document pair contains 10 section headings each, all of which are aligned. Intuitively, the latter cases should give more indication of similarity between articles than the previous ones. However, due to the limited lengths of articles in the evaluation corpus, up to 80% document pairs (555 document pairs) have at least one article with 5 section headings or fewer. As a result, only 140 document pairs were left that contained more than 5 section headings. Due to this limited number, it was not possible to further investigate the usability of the section similarity features in bigger documents. This is a promising avenue to explore for future work.

Finally, the results shown in this study provide little evidence that *links overlap features* can be used to measure similarity at the document level. This finding is interesting as links data types were shown in previous studies (Adafre & de Rijke, 2006) to be a good indicator for identifying similarity at the sentence level. One possible reason for this is due to the limited availability of the interlanguage links for the under-resourced languages. The link overlap method relies on the amount of interlanguage-links between the different language pairs, since it is utilised as bilingual resources to identify the overlapping information. For well-resourced languages (used in Adafre and de Rijke (2006)'s study, there is a significantly higher number of interlanguage-links compared to under-resourced languages. E.g., for this study, German-English has 637,382 interlanguage-links, whilst the number of interlanguage-links for under-resourced languages are much lower (i.e., below 100,000 interlanguage links for each language pair, with a minimum of 21,302 interlanguage-links for Latvian-English). In this case, if the concepts described in the interlanguage links overlap completely between German-English and Latvian-English, only 3.34% concepts available in the German-English interlanguage-links were available in the Latvian-English interlanguage-links. This significantly limits the ability of the method to identify overlapping information between both languages.

9.6.4 Identifying 'similarity' and 'comparability'

The results in this study have further indicated that language-independent features can be used to predict the degree of similarity of a Wikipedia document pair. In some cases, however, one might need to predict the degree of comparability instead. This is often needed when the document pairs are required for the purpose of extracting bilingual resources instead. How does the same features perform when it is used to classify 'comparability' instead? As shown in Table 9.14, the same features were evaluated for predicting comparability in a binary classification problem and a regression problem.

When using the same threshold for a binary classification problem (i.e., Class 1: average scores ≤ 3 and Class 2: average scores > 3), the results show that the classification features performed well in identifying non-comparable document pairs (F-measure=0.83), however, its performance in identifying comparable document pairs is much lower (F-measure=0.56). Firstly, one reason for the low performance is due to the imbalanced dataset, only 32% (256 document pairs) were assigned to be 'non-comparable'. Secondly, using a mid-point as a threshold might not be a natural way to divide these document pairs based on the comparability level. Dividing between document pairs that are parallel (i.e., exact translations of each other) and those which are not, are intuitively easier for humans to do. However, to divide comparable documents into two different classes as performed in this study may be more difficult to do as the division between the two classes may be more difficult to understand, both for humans and the classifier.

To investigate this further, the same features were tested for predicting comparability using a regression algorithm. The results show that the performance in identifying comparability is significantly better than identifying similarity (RMSE=0.56 and RMSE=0.73, respectively). This means that the classifier can be used for identifying comparable documents, which are very useful for tasks such as extracting comparable documents for comparable corpora, for the purpose of improving MT system.

The improvement for MT is not evaluated as a part of this study for the following reasons. Firstly, the size of Wikipedia document pairs in under-resourced languages is already very small, that further filtering (based on comparability) will cause the number to decrease further and become insufficient for training MT. Secondly, a number of resources are required to perform this evaluation, namely, a parallel corpus for training MT (as a baseline) and a test corpus for each of the under-resourced languages investigated in this study. If not available, manual effort should be carried out to build these resources first. Given the limited time in this study, the author was not able to investigate this area further. However, previous studies have shown that documents with higher comparability have been shown to improve the quality of extracted bilingual resources and the performance of MT, compared to those with lower comparability (Li & Gaussier, 2010; Skadiņa et al., 2012).

9.6.5 Limitations

The results shown in this study indicate that the use of language-independent features for measuring similarity across languages is promising. One limitation of this study is the small dataset used for training the classifier (800 document pairs in total; 100 document pairs per language pair) and a possible bias introduced in the dataset with regards to the high proportion of links and word overlap, and the short article lengths.

One way to address this is to implement a way to introduce more data in the training, such as the synthetic minority data (SMOTE) approach (Chawla et al., 2002), which has been explored in this study and was shown to improve the accuracy of the minority class. Another approach is to add more document pairs to increase the evaluation set. One approach that has been carried out in this study was to add more instances for the lowest-similarity class, by pairing documents that described different topics. Future work should also investigate adding more instances for the other similarity classes by using the classification approach to collect more instances for the evaluation corpus. For example, the classification approach can be used to classify articles from the current Wikipedia dump, in order to achieve a more balanced dataset and to include more current articles into the corpus. Another approach is to gather more human judgments to measure different degrees of similarity between a larger set of documents. These approaches should be investigated for future work.

Another possible limitation is that the evaluation corpus used in this study was built in 2010. Whilst the quantity of Wikipedia articles has since drastically increased,[6] the approaches used to create and modify the articles remain the same (e.g., open-edit and monitored by Wikipedia editors). Furthermore, the features included in the articles, such as links, interlanguage-links and section headings are also still used in the current version. Therefore, the author believes that the findings in this study are still applicable for the current Wikipedia version. However, future work should also explore the use of a bigger and current dataset to enable more analysis of these features, such as by utilising the classification approach to collect more current data to create a new evaluation corpus.

9.7 Conclusion

The results presented in this work have investigated the use of language-independent features in identifying the degree of similarity between Wikipedia documents. In this section the author answered the research questions presented earlier in this chapter.

RQ3. Can language-independent features be used to identify cross-lingual similarity in Wikipedia? Overall yes and with good performance for binary classification (accuracy=81.38%; F-measure=0.79) and 3-class classification (accuracy=70.13%; F-measure= 0.53) using a Random Forest classifier. Poorer performance was achieved for 5-class classification with the Random Forest classifier (accuracy=50.88%, F-measure=0.42), mostly due to the imbalanced dataset for learning the minority classes. All these results were significantly higher than the language-independent and language-dependent baselines. Furthermore, it achieves significantly higher performance (RMSE=0.73, Pearson's r=0.69) than both the language-independent and language-dependent baselines (RMSE=0.89

[6] `https://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth`, accessed on 2 April 2019

and 0.90, and r=0.47 and 0.45, respectively).

When trained on a different set of language pairs, the results show that the Random Forest binary classifier was able to achieve a comparable performance to one trained with data from the relevant language pair. The use of other language pairs in the training data was also further shown to achieve a higher F-measure score for the minority class, especially for cases where there was a lack of data in the minority class (such as EL-EN and SL-EN).

(a) How does this method compare to approaches using linguistic resources, such as MT systems? The results show that the Random Forest classifier that utilised language-independent features significantly outperforms a logistic regression classifier that utilises MT system and a TF-IDF word overlap feature. The Random Forest algorithm also outperformed a linear regression algorithm that utilised the language-dependent feature in a regression problem.

(b) How does the performance vary across language pairs? When trained using 100 pairs from each language-specific dataset, results show that the performance of the Random Forest binary classifier varies widely. It performs best in DE-EN (unweighted macro-averaged F-measure=0.90), followed by ET-EN (unweighted macro-averaged F-measure=0.80). The worst performance was achieved in SL-EN (unweighted macro-averaged F-measure=0.49), and EL-EN (unweighted macro-averaged F-measure=0.62). One reason for this is the limited training data for non-similar classes and some language pairs (e.g., 12 document pairs for EL-EN and only 4 document pairs for SL-EN). When performing regression, the lowest error rates were achieved in SL-EN (RMSE=0.40), DE-EN (0.53) and EL-EN (0.55). However, due to the different balance in the dataset for each language pairs, the performance may not be comparable and a larger dataset is required to further investigate the performance of classification and regression.

(c) What language-independent features are best for measuring cross-lingual similarity in Wikipedia? When evaluated as individual features, *length ratio* and *char-n-gram overlap* were shown to be the most useful features in identifying similarity, followed by *word overlap* and *section length ratio.* Jaccard overlap correlates higher than TF-IDF (cosine similarity). *Structural (section heading) similarity* and *links overlap* were not shown to be good features when used on their own, but are useful when used in combination with other features.

The findings presented in this work show that language-independent features can be used to identify similarity (and comparability) in Wikipedia documents. Some limitations occur due to the small number of documents to train the classifier with, especially when more than two classes are being considered. For future work, the author will investigate approaches to gather additional data (e.g., using a bootstrapping method or using human assessors to judge a larger number of document pairs) to create a more robust evaluation corpus.

Related publications

- Paramita, M. L., Clough, P., Aker, A., and Gaizauskas, R. 2012. Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, pp. 790-797.

- Paramita, M. L., Clough, P., Zhuang, M., and Gaizauskas, R. 2019. Identifying cross-lingual similarity in Wikipedia for under-resourced languages. TAL Journal, 60 (1). *(under review)*

Chapter 10

Discussion

The aim of this study is to investigate language-independent methods to measure cross-lingual similarity between interlanguage-linked Wikipedia articles. Two main contributions were produced in this thesis:

1. **The creation of a Wikipedia similarity corpus (Chapter 5)**:
   The author discusses how the similarity corpus has helped to gain more understanding on similarity in Wikipedia articles in Section 10.1.

2. **The development and evaluation of a range of language-independent methods for measuring similarity (Chapter 6-9)**:
   The findings from the experiments of developing language-independent approaches are revisited in Section 10.2. Finally, the author describes some applications that benefit from these methods in Section 10.3.

The author reflects on these contributions and their relations to existing literature in this chapter.

10.1 Creation of a Wikipedia similarity corpus

The first contribution of this work, i.e., *the creation of a Wikipedia similarity corpus* (described in Chapter 5), is aimed at gaining a better understanding of similarity in Wikipedia articles, and to assist with the evaluation of cross-lingual similarity methods.[1] This corpus contains 800 document pairs in 8 language pairs, each of which was annotated by two annotators. Each document pair was annotated (in a 5-point Likert scale) by two annotators for i) its similarity, ii) its proportion of similar content, iii) its similarity of the aligned sentences and iv) its comparability. Each document pair was also annotated with the similarity characteristics that contribute to its similarity score. The work on creating the similarity corpus has been described in Chapter 5.

In this section, the author discusses the findings in this study in increasing the understanding in three different areas: similarity (and dissimilarity) in Wikipedia, characteristics of similarity in Wikipedia documents, and the relations between similarity and comparability.

10.1.1 Similarity (and dissimilarity) in Wikipedia

Previous literature have made the assumptions that interlanguage-linked Wikipedia articles are comparable or equivalent, simply because they describe the same topics (Otero & López, 2010; Potthast et al., 2008; Sadat, 2010). A significant amount of work have further assumed Wikipedia to be a comparable corpus and used it to build multilingual embeddings (Vulić & Moens, 2015, 2016), polylingual topic models (Mimno et al., 2009), and CL-ESA (Potthast et al., 2008).

The finding in this corpus, however, suggests that *although Wikipedia interlanguage-linked articles describe the same topic, the similarity of the contents themselves may vary widely.* This finding confirms previous literature (Filatova, 2009; Patry & Langlais, 2011), which although based on much smaller studies, also found that Wikipedia articles varied differently, even sometimes containing contradictory information. A more recent finding (Gottschalk & Demidova, 2016) also confirmed that interlanguage-linked articles tended to evolve independently and might reflect different point of views, resulting in different information appearing across different languages.

[1] This corpus is freely available to download and use for future research purposes: https://ir.shef.ac.uk/cloughie/resources/similarity_corpus.html

Previous works have investigated some aspects relating to interlanguage-linked articles in Wikipedia, such as the different availability of Wikipedia topics in different languages (Bao et al., 2012) and the different editing behaviours in multilingual Wikipedia (Kim et al., 2016). However, these studies did not specifically investigate the similarity within the article content itself.

Previous studies that have annotated similarity in Wikipedia articles have focused on a small set of languages (Patry & Langlais, 2011) and a small number of topics (Filatova, 2009). The corpus presented in this work, on the other hand, contains annotations for a larger number of document pairs (800 document pairs) in 8 language pairs. As far as the author is aware, at the time of writing, there is no available corpus that has been created specifically for analysing cross-lingual similarity in Wikipedia articles.

10.1.2 Similarity characteristics of Wikipedia documents

Similarity has been described in a different way for different resources. For example, similarity between news articles were previously defined by the topics described in the stories (Braschler & Schäuble, 1998). For Web articles, similarity framework differentiate documents based on the topics described in the content. Documents that described the same topic have been judged to be 'comparable' (Fung & Cheung, 2004; Skadiņa et al., 2012), whilst those that were not in the same topic but still in the same domains referred to as 'weakly comparable' (Skadiņa et al., 2012). These schemes are again summarised in Table 10.1.[2]

Although these studies have aimed to define similarity, very limited work have been carried out to identify similarity characteristics in Wikipedia articles. In this study, the author chose not to use these established similarity frameworks as they were deemed too coarse-grained to capture the differences in similarity in Wikipedia. E.g., all interlanguage-linked Wikipedia articles describe the same topic; therefore, using the available schemes, *all* interlanguage-linked Wikipedia documents should be categorised to be 'comparable' (Fung & Cheung, 2004), 'similar stories of the same topic but may be narrower or broader'

[2] This table was previously shown in Table 2.2 in Section 2.2.

Table 10.1 A comparison of comparability levels

Skadiņa et al. (2012)	**Braschler and Schäuble (1998)**	**Fung and Cheung (2004)**
Parallel represents texts which are accurate translations, or approximate translations with some addition or omissions	Pairs of documents of the same topic or event	*Parallel* represents texts which are translated sentence by sentence.
		Noisy parallel represents texts which are mostly parallel but contain non-aligned sentences which may be caused by paragraph insertions or deletions.
Strongly comparable represents texts coming from the same source or containing the same subject		*Comparable* describes texts which do not contain aligned sentences but are about the same topic.
	Similar stories of the same topic but may be narrower or broader	
Weakly comparable represents texts in the same domain but different events	Related topics which share aspects such as location or person	*Non parallel* represents disparate bilingual documents which may or may not be in the same topic.
	Different topics but share same terminologies	
Not comparable	No similarities	

(Braschler & Schäuble, 1998), and 'strongly comparable' (Skadiņa et al., 2012). Furthermore, the findings in this thesis suggest that there are significant differences between the article contents themselves. These differences can not be captured using the established frameworks since they were not specifically developed for Wikipedia.

In this study, the author aimed to analyse the characteristics of similarity in Wikipedia documents by gathering manual judgments on these aspects. As described in Chapter 5, annotators were asked to annotate the similarity of the documents in the similarity corpus using a 5-point Likert Scale. They were then asked to specify the characteristics of each document pair that they had to annotate.

Secondly, the use of a 5-point Likert scale allowed annotators to specify their own understanding of the different characteristics that contribute to their assigned similarity scores. As a result, this allows the similarity characteristics exhibited by Wikipedia document pairs specifically to be investigated, especially how they differ between document pairs of different similarity scores (see Figure 10.1).[3] These features have been utilised in developing relevant measures for identifying similarity (further described in Section 10.2).

Using this evaluation set, it was identified that highly similar document pairs (assigned scores of 5) are very likely to have similar structure; this may indicate that the documents may describe similar aspects or sub-topics. These highly similar document pairs were also found by the annotators to contain overlapping named entities, overlapping fragments, and most contain contents that are translations of each other. These characteristics are most similar to Fung and Cheung (2004)'s definition of 'parallel' and 'noisy parallel' documents.

In contrast, most non-similar document pairs (assigned scores of 1) were found to contain very different contents. Although they describe the same topic (which is the definition of a 'comparable' document pair in Fung and Cheung (2004)), the characteristics of these documents instead suggest that the contents between these documents differ significantly; they did not share similar structure and did not contain overlapping fragments

[3] This figure was previously shown in Chapter 5 (page 118).

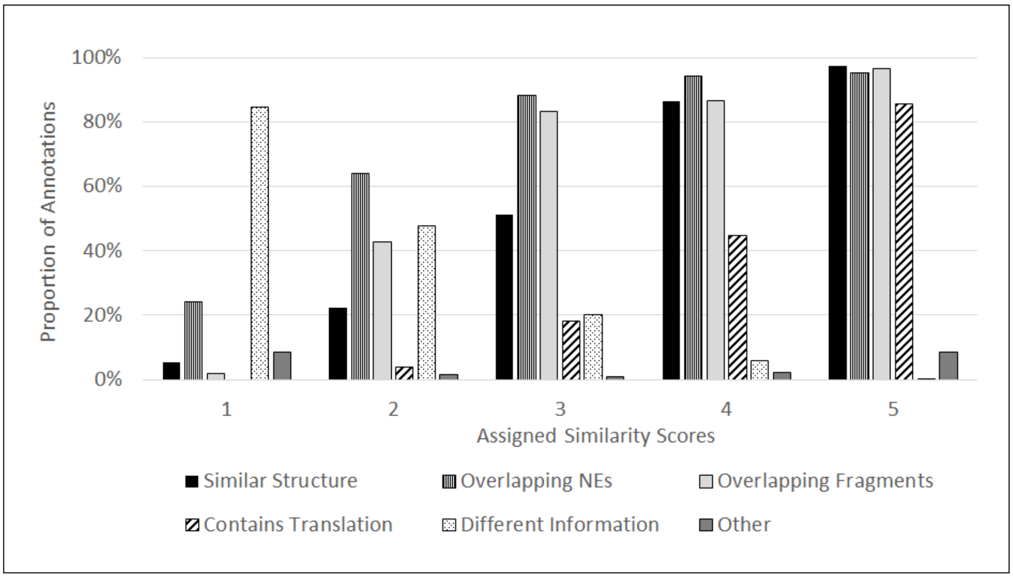


Fig. 10.1 Characteristics that capture various levels of similarity (N=1,600)

or contain translated contents. In most cases, the contents themselves were annnotated to contain different information. These document pairs, as a result, were closer to the definition of 'non parallel' documents (Fung and Cheung (2004)'s framework),'not comparable' documents (Skadiņa et al. (2012)'s framework) and 'no similarities' documents (Braschler and Schäuble (1998)'s framework).

Most document pairs with assigned similarity scores of 4 were shown to contain similar structure, overlapping named entities and overlapping fragments, but less than half were annotated to contain translations; this is most similar to the definition of 'comparable' articles (Fung & Cheung, 2004). Document pairs with assigned scores of 2-3, however, are more problematic to align to existing frameworks. These documents were shown to contain overlapping named entities and fragments, but very few contained translations. These document pairs are better represented by the following Braschler and Schäuble (1998)'s categories: 'related topics which share aspects such as location or person' and 'different topics but share same terminologies'.

These findings show that the characteristics between Wikipedia document pairs do resemble some of the categories proposed in the established frameworks. However, none of the framework accurately captured the different similarities exhibited by Wikipedia documents. These findings suggest that a Wikipedia-specific framework is required to further define the degrees of similarity in Wikipedia articles, which should be investigated as a future work.

10.1.3 Similarity vs comparability

The creation of the Wikipedia similarity corpus has also allowed us to further investigate the following two aspects: 'similarity' and 'comparability', which have previously been used interchangeably to represent the relatedness of two documents (or sub-documents) in different languages. In previous studies, the term 'similarity' has also been used to represent the relatedness of contents between two texts that are written in the same languages, such as similarity between documents (e.g., in identifying similar news articles (Barker & Gaizauskas, 2012)), similarity between sentences (e.g., semantic textual similarity (Agirre et al., 2012)), and similarity between words (e.g., using word embeddings (Mikolov et al., 2013a)). When the texts to be compared are written in different languages, the term 'cross-lingual similarity' has been used instead. The term 'comparability' has also been used to represent the relatedness of documents of different languages (McEnery & Xiao, 2007). This brings us to the question, how does cross-lingual similarity relates to comparability? As far as the author's aware, no work has further determined the relations between these two aspects.

In building the similarity dataset, the author asked annotators to annotate the document pairs with a similarity score (Q1) and comparability score (Q4) in a 5-point Likert scale. The findings in the similarity dataset (see Chapter 5) allow us to investigate on how 'similarity' relates to 'comparability'.

Firstly, previous literatures has used the amount of translated content found in the articles to represent the degree of comparability of the corresponding articles (Fung & Cheung, 2004). When analysing the annotations of the corpus, specifically the similarity scores and similarity characteristics between documents, the findings show that most documents assessed to be most similar (score of 5) were also annotated to have translated contents. Furthermore, annotators' scores of similarity scores and proportion of shared contents (Q1 and Q2, respectively) were shown to correlate very highly (r=0.88, p<0.01). This indicates that the concept of similarity was also strongly associated to the proportion of translated contents available in the articles. The similarity characteristics shown in Figure 10.1 further shows that the proportion of articles with translated content decreased significantly when their similarity scores were annotated to be 4 or below. These findings suggest that the concept 'similar' and 'comparable' do overlap as they both correlate with the amount of overlapping contents within them. This was further evidenced by a strong correlation between the similarity scores (Q1) and comparability scores (Q4) in the evaluation corpus (r=0.85; p<0.01).

The findings in this study indicate that the differences between these aspects are narrowed down to the *usefulness of the texts in supporting bilingual resources* (see Section 5.6.2). I.e., some highly similar documents in Wikipedia were given lower comparability scores due to the lack of useful bilingual resources in the contents. E.g., article pairs that contain list of named entities that are written the same in different languages, or contents written in a different language to the article (untranslated contents), may be annotated to be highly similar, yet are not considered to be useful bilingual resources, and therefore annotated with a lower score of comparability.

Despite these differences, the high correlation between these two aspects indicates that the approaches investigated in this study - although focusing on measuring similarity of document pairs - can be used for gathering comparable documents from Wikipedia. This should, as a result, increase the degree of comparability within the corpus for a more effective use in future tasks.

10.1.4 Gold-standard for evaluation

Finally, the creation of the similarity corpus has provided a gold-standard dataset to evaluate the approaches developed to measure cross-lingual similarity in Wikipedia. Other corpora are available that contain documents with different degrees of comparability. E.g., the Europarl (Koehn, 2005) corpus contains translated documents of European Parliament proceedings, and DGT-TM (Steinberger et al., 2012) contains a corpus of translated sentences in the legal domain in 22 EU languages. A comparable corpus containing news articles and documents from narrow domains were also made available by the ACCURAT project (Skadiņa et al., 2012). However, the availability of these corpora are only limited to some domains. Furthermore, since these documents come from different sources, the characteristics of similarity between these documents may not be applicable for encyclopeadic contents that are exhibited by Wikipedia documents.

As far as the author is aware, the similarity corpus produced in this work is the only available cross-lingual corpus containing Wikipedia documents with different degrees of similarity. Furthermore, this corpus allows future work to further investigate similarity in Wikipedia and to automatically evaluate methods for measuring cross-lingual similarity in Wikipedia.

10.2 Language-independent methods to measure cross-lingual similarity

The second contribution of this work is the *development of language-independent approaches to measure cross-lingual similarity in Wikipedia*. The cross-lingual similarity methods developed in this study were evaluated on eight language pairs: German (DE), Greek (EL), Estonian (ET), Croatian (HR), Latvian (LV), Lithuanian (LT), Romanian (RO) and Slovenian (SL), all paired to English (EN). The first language, German, is considered to be highly-resourced language; whilst the remaining seven are under-resourced languages. The language selections cover different language groups, Hellenic (EL), Baltic (LV and LT), Slavic (HR and SL), Romance (RO) and Germanic (DE and EN). All languages used Latin alphabets, except for EL that uses the Greek alphabet.

The approaches proposed in this work can be applied to all languages that use Latin characters. For cases where the languages are not written in Latin, a transliterator is required to transform the characters into Latin prior to using these methods. These were investigated for Greek-English, where a Greek transliterator was utilised prior to extracting the language-independent features. In this study, however, these approaches were only tested on 8 language pairs; all of which were European languages. Further study is required to evaluate how well these methods work in measuring similarity for other language pairs (such as Asian or African languages).

These findings fill the gap in the area as methods that have been investigated to measure cross-lingual similarity between documents often relied on the availability of language-dependent resources, such as an SMT system (Uszkoreit et al., 2010; Yasuda & Sumita, 2008), bilingual dictionaries (Munteanu & Marcu, 2005) or parallel corpora (Munteanu & Marcu, 2005). These resources are available for a small number of languages and domains. The use of these language-dependent methods, therefore, is not applicable for under-resourced languages. The methods described in this study, on the other hand, utilised language-independent features. The author further described how this relates to other literatures in the area.

10.2.1 Language-independent features

This study shows that language-independent features can be utilised to indicate similarity at the document level with promising results. In this section, the author relates the use of these features to previous literature.

The use of *word overlap* has been utilised in many monolingual IR studies to identify similarity between a query and a document (Baeza-Yates & Ribeiro-Neto, 1999; Manning et al., 2008; Salton & McGill, 1986). In this study, the feature was utilised in a cross-lingual setting to capture identical text (such as named entities) that might be shared across different languages. For cases where these texts were not identical, *char-n-gram overlap* was utilised to capture the overlapping characters. This feature has been used in the past to identify similar words in a cross-lingual settings, mostly in European languages, since these words may contain overlapping characters (McNamee & Mayfield, 2004).

Another feature, *word length*, has also been used in previous work to identify parallel sentences (Munteanu & Marcu, 2005; Patry & Langlais, 2011) and parallel documents (Resnik & Smith, 2003). In this study, the feature is also found to be a strong indication of the similarity of Wikipedia articles, although these articles often do not correspond in a translation manner and were written by different authors (Barrón-Cedeño, Paramita, Clough, & Rosso, 2014).

Measuring the *structural similarity* between documents has been explored in previous study (Resnik & Smith, 2003), which measured the similarity of HTML structure (i.e., the appearance and order of HTML tags in the documents) to identify parallel or translated documents in the Web. This method, however, differs substantially from the structure similarity methods previously used. Instead, the structural similarity method described in this work analyses and aligns similar section headings between the Wikipedia articles to measure similarity. This method was shown to be able to predict similarity at the document level. The *section lengths ratio* was also shown to be a good feature to identify similarity in Wikipedia articles. As far as the author is aware, these features have never been previously investigated as a feature to identify cross-lingual similarity in Wikipedia articles.

Finally, the use of *links overlap* has been investigated in previous studies as a feature to identify similar sentences in Wikipedia (Adafre & de Rijke, 2006). This approach achieved promising results in identifying similar sentences when tested in languages sharing similar roots, such as French-English (Patry & Langlais, 2011) or Dutch-English (Adafre & de Rijke, 2006). However, the findings in this study show contradictory results. This study shows that the links overlap is not a promising feature for measuring similarity at the document level for under-resourced languages. One possible reason for this is the limited number of interlanguage links available for under-resourced languages, compared to high-resourced languages, which limit the size of extracted bilingual dictionaries for identifying the overlapping information across the different languages.

Another issues that might cause the poor performance of links overlap is because Wikipedia links are unlikely to appear throughout the whole articles in Wikipedia. For example, not all relevant texts were linked to the correct concepts, or only the first occurrence of the relevant text is linked to the correct concept. Future work should therefore explore the use of Wikification, i.e., enriching Wikipedia texts with their corresponding wiki-links (Mihalcea & Csomai, 2007; Tsai & Roth, 2016) prior to measuring similarity of links at the document level.

Features investigated in this study, such as word overlap, char-n-gram overlap, ratio of section lengths, and word length can be extracted without the use of any linguistic resources. Other features, such as structural similarity overlap and link overlap, can be extracted using freely available online resources, i.e., Wikipedia and Wiktionary. These resources are investigated in this study because they are widely available in a large number of languages, i.e., 297 language pairs in Wikipedia[4] and 147 language pairs in Wiktionary.[5] These features were tested individually (Chapter 6-8), and in a classification approach (Chapter 9). The results show that language-independent features, supported with features utilising widely available resources, can be used to identify cross-lingual similarity in Wikipedia articles.

In the past decade, the work on measuring cross-lingual similarity have expanded to

[4] https://meta.wikimedia.org/wiki/List_of_Wikipedias, accessed on 11th July 2017.

[5] https://stats.wikimedia.org/wiktionary/EN/Sitemap.htm, accessed on 11th July 2017.

the use of *topic models*, such as cross-lingual LDA (Mimno et al., 2009), cross-lingual LSI (Saad et al., 2014), and cross-lingual ESA (Potthast et al., 2008). Instead of using word overlap between text, these approaches map documents into topics, and compare the distribution of topics in the documents to measure its similarity. These approaches require a parallel corpus for training (such as Europarl), which limit the scope of languages they can process (Mimno et al., 2009). Others required monolingual corpora to train the topics individually, but then required an MT system to project the two monolingual topics into the same space (Saad et al., 2014). The use of Wikipedia as bilingual training data have also been explored by Mimno et al. (2009) with the assumption that the content were equivalent across languages. The findings in this study, however, contradict this assumption and more investigation is needed to investigate the value of these approaches when trained using the entire Wikipedia. Moreover, these approaches have been used in supporting various tasks, such as to analyse topics in a corpus, or to perform text classification and clustering. Its use to measure content similarity between topically related articles (such as interlanguage-linked Wikipedia articles) should be investigated further as future work.

The use of cross-lingual ESA also shows that Wikipedia can be used to train a multilingual retrieval model, although over 100,000 Wikipedia document pairs are required to train the model to achieve a good performance (Potthast et al., 2008). Unfortunately, during the time of the experiments, this requirement was not satisfied for most of the language pairs investigated in this study. The size of interlanguage-linked articles in these languages, however, have significantly increased over the years. Therefore, a comparison of these approaches with CL-ESA is planned for future work.

Another state-of-the-art approach is the use of *word embeddings*, which has been the focus of a large number of studies in both monolingual and cross-lingual similarity over the past five years (Artetxe et al., 2017; Mikolov et al., 2013a; Ruder et al., 2017). Most of these approaches require a parallel corpus to train the bilingual word representations (Vulić & Moens, 2015). Others, similar to the work in cross-lingual topic models, trained monolingual word embeddings separately using a monolingual corpus, which is widely available. However, a parallel corpus or a bilingual dictionary is still required to project both monolingual embeddings into the same space (Mikolov et al., 2013a; Mogadala & Rettinger, 2016). A more recent work (Artetxe et al., 2017) show that it was possible to train bilingual word embeddings with the use of a very small dictionary (25 terms) or an initial alignment containing only numbers. This presents a possibility to create bilingual resources for more languages, which should be investigated in a future work. Although these approaches have shown good results in performing monolingual tasks, their performances in cross-lingual tasks have been limited to measure similarity between words (Søgaard et al., 2015) and to extract bilingual lexicon (Artetxe et al., 2017).

Further work shows that using an aggregation of word-embeddings in measuring cross-lingual similarity at a document level results in poor performance (Le & Mikolov, 2014). Better performance is achieved by training a document vector instead. However, the availability of previously trained bilingual document vectors is even more limited for under-resourced languages. Štajner and Mladenić (2018) also found that bilingual word embeddings (trained on Wikipedia interlanguage linked articles) show promising results in measuring cross-lingual similarity when given a comparable corpus of a sufficient size; however, they perform poorly for under-resourced languages.

Furthermore, these tasks are more expensive to compute than the lightweight approaches described in this work. Based on this observation, the author did not integrate the recent work in word embedding into this study. The extensive development in the area, however, show promising results that they are powerful tools to measure cross-lingual similarity. As a future work, this feature should be investigated as a language-independent method to measure cross-lingual similarity.

The focus of this study was to develop a lightweight language-independent approach that can be used to measure cross-lingual similarity in Wikipedia articles. The features proposed in this study are, therefore, lightweight and easy-to-extract. Most features can be extracted without any resources at all, whilst a few others utilise resources, such as Wikipedia and Wiktionary, which are widely available in a large number of languages.

10.2.2 Similarity at the document level

The methods developed in this study also work differently to others as they were investigated for measuring similarity in Wikipedia at different levels. Previously, similarity methods were applied to identify similar content at the *sentence level* (Adafre & de Rijke, 2006), *paragraph level* (Gottschalk & Demidova, 2017) or at the *infobox level* (Adar et al., 2009). In this study, the author proposed methods that can be used to identify similarity of Wikipedia at the *document level.*

One may argue that for the purpose of extracting bilingual resources (such as terms or sentences), extraction tools can be run in the entire corpus without the need of filtering out the non-similar documents. I.e., these tools can be used to identify parallel sentences or terms in document pairs regardless of their similarity. However, the quality of bilingual resources extracted from comparable corpus and non-comparable corpus are likely to be different; the latter is more likely to introduce some noise or inaccuracies. These results have been investigated in previous studies, such as Vulić and Moens (2013), Saad et al. (2014) and Erdmann et al. (2009), who discovered that results based on Wikipedia are considerably poorer compared to parallel or corpora. On a very large corpus, it is also useful to be able to reduce the size of the corpus to include only the similar documents in order to reduce the amount of processing of document pairs that may not contribute to the bilingual term extraction process. The availability of methods to measure similarity at the document level can assist with these issues. Furthermore, Wikipedia has been used in the past as a comparable corpus for measuring cross-lingual similarity (Potthast et al., 2008), or to extract bilingual word and document embedding (Conneau, Lample, Ranzato, Denoyer, & Jégou, 2017). Similarly, the ability to filter out the less similar documents will produce a better quality corpus to support these tasks. The author is not aware of other language-independent methods that can be used to measure cross-lingual similarity specifically for Wikipedia articles.

10.3 Applications

The previous sections have discussed the contributions of this study, including what can be learned from the similarity corpus and the developed methods. One question that arise is this: what applications can benefit from these approaches? In this section, the author discusses how the findings discovered in this study can help with a number of applications.

10.3.1 Gaining insights from Wikipedia

In previous literature, Wikipedia has often been assumed to be a comparable corpus, i.e., the interlanguage-linked articles describe the same topics, and have therefore been assumed to be similar (Potthast et al., 2008). Some literature that studied the similarity between articles, however, found that this was not always the case (Filatova, 2009; Patry & Langlais, 2011). This study further confirms the findings in previous literature that the degree of similarity in Wikipedia corpus varies widely.

Questions like, 'How many articles in Wikipedia are similar to each other?', 'Does the similarity vary between language pairs?', 'Does the degree of similarity improve in time?' or 'Are some domains more similar than other?' are just a few questions that are difficult to answer due to the lack of approaches to measure similarity in Wikipedia across a large number of languages. Approaches investigated in this study can be used to gain more insights on these questions.

At the moment, when Wikipedia readers read a topic in Wikipedia in their preferred languages, it is difficult to know if there are any additional or different information that are available in other languages. Contradictory information, that may appear between the different language versions, may mislead the readers if they assume the contents they read (in one language version) are correct.

Automatic methods that predict similarity scores in Wikipedia documents, such as the classification approach proposed in this work, are able to indicate to readers of the similarity of other interlanguage-linked articles of the same topic. A low similarity score may indicate that there may be some difference in contents in the multilingual articles describing the same topic. Furthermore, a combination of this insight with translation process (either a professional translator or an automatic machine translation) is very helpful to provide Wikipedia users with more understanding of how information differ between the different language versions of the articles.

Moreover, some of the approaches investigated here can be adapted to identify specific content within the articles that are similar or dissimilar across two language versions. For example, the structural similarity features can be used to identify omitted sections between multilingual Wikipedia articles. $Anchor + word$ method can be used to identify sentences that exist in other language versions of Wikipedia, but are not available in the language version that the readers currently read. Providing readers with these insights on the degree of similarity in Wikipedia are important to bridge information gap across languages and to provide more understanding for readers of how the cross-lingual contents may differ (Mimno et al., 2009).

10.3.2 Identifying comparable documents for MT

Wikipedia has been used as a source for gathering comparable documents as training data for statistical machine translation (SMT). The findings in this study, however, have confirmed that the degree of similarity in Wikipedia articles vary widely. Including document pairs that are not similar may cause some noise and poor quality of extracted resources (such as noisy terms or incorrect translated sentences).

The approaches explored in this study can be used to improve the quality of comparable corpus built using Wikipedia by pre-filtering documents or contents with low similarity (or comparability). Better quality comparable corpora should improve the performance of tasks and quality of resources that used Wikipedia as a comparable corpus, such as CL-ESA (Potthast et al., 2008), word embeddings (Mikolov et al., 2013a) and statistical machine translation (SMT). Due to the limited availability of training and testing data for these under-resourced languages, the author did not investigate the improvement of SMT using Wikipedia corpus of different similarity in this study. However, previous work has indicated that the improvement of MT is higher when trained with a better quality comparable corpus (Skadiņa et al., 2012).

In this section, the author identified two different ways to utilise the approaches as a pre-filtering step for improving the quality of Wikipedia corpus.

Pre-filtering non-similar documents for improving corpus quality

The classification approaches explored in this study (Chapter 9) have shown the ability to differentiate Wikipedia document pairs with different degrees of similarity. This information can be utilised to pre-filter documents that are not similar to each other. E.g., the findings in the similarity corpus has suggested that Wikipedia document pairs with similarity scores lower than 4 were found to have very limited translated contents between them. If one's aim is to use Wikipedia documents to extract translated sentences, then the 5-class classification or regression approach can be used to identify document pairs with a score of 4 and above. Documents with scores below 4 may still be useful if the aim is to extract relevant terms or phrases instead. If the aim is to use Wikipedia for terms or phrases extraction, the 3-class classification or binary classification can be used to filter out the non-similar document pairs only. The ability to differentiate between the different granularity of similarity is a useful feature of the classification approach proposed in this study.

Pre-filtering non-similar contents for improving corpus quality

Another interesting finding in this study is that document pairs with low similarity scores, may still contain similar contents. It is likely that this is caused by the nature of Wikipedia that allow articles in different languages to be continuously developed over time, with new information being added (or removed) by different people and at different rates (Gottschalk & Demidova, 2016). Document pairs that might have been developed by translating the content from one language to another, are likely to contain different information over time due to these changes.

The approaches described in this work identify similarity at the document level. It does not currently capture i) document pairs which contain medium similarity throughout all the content, or ii) those that contain parallel/highly comparable content, followed by low comparable texts.

However, this method can be easily utilised for identifying similarity at the sub-document level. This can be performed by using the approach to measure the similarity of a smaller proportion of the documents, e.g., at the level of paragraphs or the level of sections. Moreover, the section heading approach can also identify sections which are likely to be similar. This information can be used to filter out non-similar section headings prior to measuring similarity.

In the first stage, one can use the structural similarity features to identify which section headings are related to each other. This method will output an alignment of section contents which are predicted to be related based on the similarity of their section headings. This process is shown in Figure 10.2.

In the second stage (Figure 10.3), a classification approach can be used to identify how similar the contents of the sections are, disregarding the document contents in the rest of the sections. Furthermore, the `anchor+word` method can also be used in the next step to identify the translated sentences between both documents, if required.

The evaluation corpus further shows that Wikipedia articles may contain duplicate content, or content in a different language to the article (e.g., untranslated content). However, this can be filtered out by post-processing stages, such as i) performing a language identification of the document pairs prior to measuring similarity (i.e., to ensure that the contents are written in the correct languages), ii) performing a maximum threshold of the word overlap feature (to avoid including articles with duplicate contents), or iii) filtering out duplicate items that are extracted in the term extraction process. Any of these three processes will ensure that the bilingual resources gathered from the corpora are of the correct languages, whilst still maintaining the accuracy of the extraction.

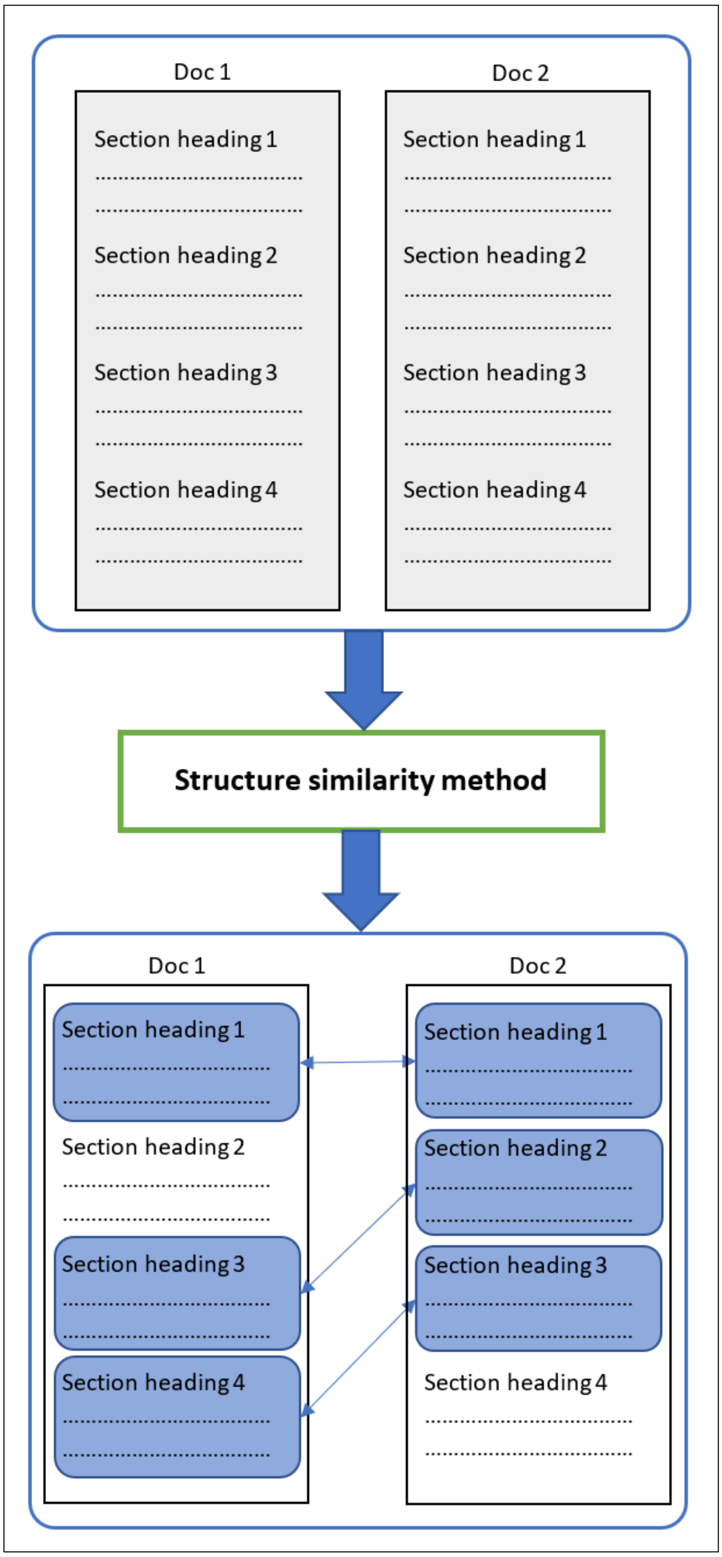


Fig. 10.2 Stage 1: Pre-filtering non-similar sections using the structure similarity approach

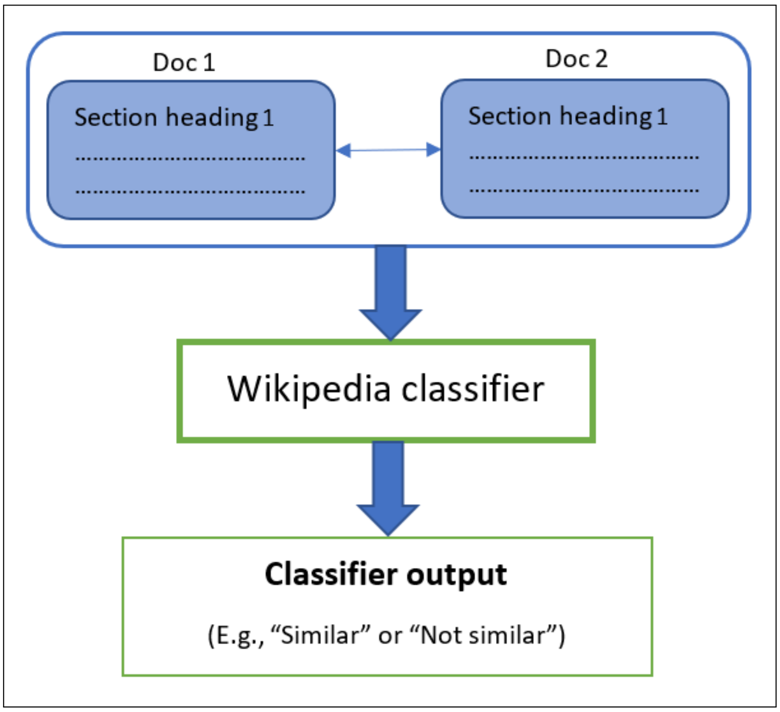


Fig. 10.3 Stage 2: Predicting degree of similarity of the aligned sections

Chapter 11

Conclusion and Future Work

This study aims to investigate methods to compute cross-lingual similarity between interlanguage-linked Wikipedia articles. Two research objectives were identified at the beginning of this study: to develop an evaluation benchmark containing human judgments on the similarity of interlanguage-linked articles, and to develop language-independent techniques to measure cross-lingual similarity across Wikipedias. These two objectives were achieved in this study: firstly, the work described in Chapter 4 and Chapter 5 have resulted in an evaluation benchmark that has allowed us to further understand the characteristics of similarity in Wikipedia; secondly, the four experiments carried out in this study (described in Chapter 6 to Chapter 9) have investigated a number of language-independent approaches and features that can be used to measure similarity in Wikipedia. In this chapter, the author summarises the two contributions of this study and answers the research questions. The limitations of the work are described and avenues for future work are also outlined.

11.1 Research contributions

There are two main contributions of work undertaken in this thesis:

1. The creation of a Wikipedia similarity corpus to understand the characteristics of similarity in Wikipedia document pairs. This corpus is available in 8 language pairs

and contains 800 document pairs with wide range of similarity degrees. As shown in this study, this corpus is also suitable for evaluating new automatic methods for measuring similarity in Wikipedia.

2. The development of language-independent approaches that can be used to measure similarity in Wikipedia articles in a large number of languages. The features used in this study are mostly language-independent, whilst a small number of features require the use of Wikipedia and Wiktionary which are widely available.

These contributions have enabled us to further understand the degree of similarity (or dissimilarity) in Wikipedia and to identify language independent methods to measure similarity in Wikipedia articles across a large number of languages.

11.2 Research questions

The findings and analysis were then used to address the following research questions.

11.2.1 RQ1: "What are the characteristics of similar interlanguage-linked articles in Wikipedia?"

Previous studies have contradicting ideas about how similar Wikipedia interlanguage-linked articles are. The findings in this study confirm that Wikipedia articles have varying degrees of similarity. These findings further reveal more information on the document characteristics that contribute to the different degrees of similarity. Highly similar document pairs were found to have similar structures and contain translated contents. They also contain a high number of overlapping named entities and overlapping fragments. Furthermore, these document pairs rarely describe different information.

The results further indicated that the lower the similarity score is for a document pair, the less similar the structure of the documents are, and the less amount of overlapping named entities, fragments, and translated contents that can be found in that particular document pair. On the other hand, the amount of different information that appear in both documents are likely to increase. These findings show that, although Wikipedia interlanguage-linked articles describe the same topic, the characteristics of these articles differ significantly. These articles, as a result, should not be assumed to be equivalent to each other.

11.2.2 RQ2: "Can we create an evaluation benchmark for Wikipedia? Do human assessors agree on Wikipedia similarity?"

In this study, the author created an evaluation corpus by gathering human judgments on 800 document pairs across 8 language pairs. For each document pair, two annotators with strong background of cross-lingual similarity assessed the contents of the documents and provided annotations on several aspects of the documents, including its similarity. The document pairs included in the evaluation corpus contain varying degree of similarity. This enables this corpus to be used as gold-standard data for various evaluation tasks, such as to evaluate the performance of automatic measures (as utilised in this study).

When measuring similarity in a 5-point Likert Scale, assessors agreed with each other with a moderate agreement (Spearman's $\rho = 0.59$; weighted Cohen's Kappa=0.38). In 84% of the cases, assessors provided the same similarity score or differing by 1, and up to 98% of the cases contained similarity scores that differed by two or less. These results show that the assessors agreed with each other to some extent, however, they might have different understanding of similarity which resulted in the varying similarity scores. Future work is required to better define the different degrees of similarities in Wikipedia in order to increase the agreement. Gathering extra annotations on the evaluation corpus should also be investigated as future work to make the judgments more robust.

11.2.3 RQ3: "Can language-independent approaches be used to identify cross-lingual similarity in Wikipedia?"

The novelty of this work lies in the investigation of language-independent features for measuring cross-lingual similarity in Wikipedia. In general, the results show that language-independent features can be used to identify cross-lingual similarity in Wikipedia articles. Most of the features described in this study can be extracted without the need of any language-specific translation resources. A small number of them utilise freely available resources, such as Wikipedia and Wiktionary, which are available for a large number of languages. This contribution has filled the gap in the literatures of the lack of language-independent approaches to measure cross-lingual similarity.

In this study, the author carried out four experiments to investigate features and approaches to measure similarity in Wikipedia. The author investigated individual features in the first three experiments, and a combination of features in a classification approach in the fourth experiment. The results for each approach are summarised below.

Anchor text and word overlap method

The anchor text and word overlap method (`anchor+word` method) (Paramita, Clough, Aker, & Gaizauskas, 2012), described in Chapter 6, was adapted from the link-based bilingual lexicon approach (Adafre & de Rijke, 2006). This approach does not require any additional language-specific resources. Instead, it creates a bilingual dictionary for each language by extracting titles of Wikipedia interlanguage-linked articles. This approach then utilises the bilingual dictionaries to identify the similar information between different sentences in the Wikipedia articles, align similar sentences together, and aggregate this information to represent the similarity at the document level.

Although this method was shown to perform well in identifying similar sentences in Dutch-English, the results in this study show that this approach correlates poorly with human judgments (ρ=0.374). However, this approach was shown to achieve a strong correlation with similarity measures using machine translation (ρ=0.717). The measure of similarity varies widely across language pairs with, for example, German-English results correlating better with human judgments than Estonian-English.

In addition to link overlap, the word overlap between the documents was also used to capture content shared between both documents. This method, as expected, was shown to work better in identifying cross-lingual similarity between languages of similar roots (e.g., German and English), as these languages are more likely to share the same content (such as named entities), compared to languages from different language groups (e.g., Greek and English).

Another insight gained from this study are the differences in the numbers of links between highly-resourced and under-resourced languages. E.g., the number of interlanguage-linked Latvian-English document pairs is 3.34% of the German-English document pairs (21,302 document pairs and 637,382 document pairs, respectively). The number of interlanguage links correlates significantly with the bilingual lexicon size, and therefore, poor availability of the interlanguage links would also affect the ability of the link overlap method to identify the similar information across languages.

One important finding that was gained from this study was that there were other similarity aspects not captured using this approach, mostly due to the varying document lengths. I.e., document pairs with varying document lengths are often punished by annotators although they contain many similar sentences. Therefore, the author investigated more features that capture different aspects of content similarity in the document pair in the next approach.

Content similarity features

In the second experiment, described in Chapter 7, the author re-visited the similarity characteristics that annotators have specified for each document pairs, and listed different similarity features that could be extracted automatically to capture these different similarity characteristics (Barrón-Cedeño et al., 2014). In this case, different features such as overlap of links, character-n-gram overlap, and word count ratio were investigated independently.

The results show that “word count ratio” is the most promising feature for measuring similarity between interlanguage-linked document pairs (ρ=0.44). This finding is interesting as the feature itself is very simplistic and can be extracted without the use of any linguistic resources. It does, however, require the contents to be tokenised before performing the word count. Languages which do not have clear separation between the words will require pre-processing before this feature can be used. It further shows that a simple linear combination of both “word count ratio” and “char-3-gram overlap” can improve the performance better (ρ=0.54). The use of “link overlap”, however, was not shown to be a good feature for capturing similarity degree at the document level (which confirmed the findings in the first experiment).

Structure similarity features

The third experiment, described in Chapter 8, investigated the possibility of using features extracted from the document structure (i.e., the section headings/titles information) to indicate the degree of similarity of the document content (Paramita, Clough, & Gaizauskas, 2017). This contribution is novel as the use of section headings to indicate the similarity at the document level has not been investigated before. Different to the article content, the section headings of the articles often do not contain interlanguage links. In this approach, both Wikipedia and Wiktionary were utilised as translation resources in order to identify and align the similar section headings. The results of this study are limited to seven language pairs due to the unavailability of Croatian-English Wiktionary data at the time of the experiment.

The results show that the similarity of section headings and the ratio of section lengths can be used to identify cross-lingual similarity of the contents with higher performance to measuring char-3-gram overlap of the content (ρ=0.36 and ρ=0.34, respectively). A combination of the section length ratio, structure similarity, and char-3-gram overlap were able to further increase the correlation scores to human judgments (ρ=0.50).

The results further show that the bilingual lexicon extracted from Wikipedia interlanguage links were not sufficient to measure the cross-lingual similarity between the section headings. Wikipedia-based bilingual lexicon contain a list of concepts that are described in Wikipedia, such as named entities or topics, e.g., “England”, “Physical therapy”. This lexicon, however, does not contain translations of common words. Some section headings, however, may use some common words to describe the contents. E.g., the English Wikipedia article of “Information Retrieval” contain section headings such as “performance and correctness measures”, “major conferences” and “awards in the field”. These words are not included in the Wikipedia bilingual lexicon and therefore cannot be translated or identified to be similar across languages. Dictionaries containing common words are therefore important to identify these overlapping information across languages. This study utilises Wiktionary, which is available in a large number of languages. However, any dictionaries may be used as a substitute if necessary.

Finally, the use of Wikipedia and Wiktionary was shown to achieve a comparable performance ot using Google Translate in translating the section headings, prior to measuring similarity on most languages, except Greek-English.

Classification approach

The previous experiments identified a number of features that can be used to measure similarity in Wikipedia. However, they indicated that a combination of features might be required to further improve the results. These findings lead us to the final approach: the Wikipedia classifier, described in Chapter 9. In this final experiment, all the features identified in this thesis were investigated for a classification and regression problem.

When developed and tested across 8 language pairs, the Random Forest binary classifier was able to correctly classify 81.38% document pairs, achieving F-measure of 0.63 (significantly higher than the two baselines, i.e., a language-independent baseline based on the word length ratio feature, and a language-dependent baseline based on TF-IDF word overlap of translated words using Google Translate). The language-independent features utilised in the classifier was also able to classify documents into three classes (accuracy=70.13%; F-measure=0.53) and five classes (accuracy=50.88%; F-measure=0.42), and in a regression approach (RMSE=0.73; Pearson’s r=0.69).

The performance of these features were also investigated for each of the language pairs, both for a binary classification problem and a regression problem. Both results showed that the performances between language pairs varied widely; the best result was achieved in DE-EN (the highly resourced language pair). The performance between under-resourced language pairs varied widely (binary classification: F-measure between 0.49 and 0.90; regression: RMSE between 0.40 and 0.83). However, these results show that in all cases, the performance of the language-independent features is better than both baselines. Please note that these results did not directly mean that the language-independent features performed better in identifying similarity in some languages than others, as the proportion of similarity in the dataset used to train each language pair differ significantly.

A comparison to MT features shows that the combination of language-independent features were shown to outperform a linear regression using the word overlap of translated contents.

Summary

The results show that these approaches can be used to indicate similarity in Wikipedia across language pairs, which are very promising considering they do not require the need of sophisticated linguistic resources. The results show that the performance of these methods are comparable to those using a state-of-the-art machine translation system. In some cases, a combination of language-independent features was also shown to outperform the use of MT system to translate the articles prior to measuring cosine similarity of TF-IDF.

The translation quality used as a baseline in this work was carried out during the time of the first experiment (carried out in 2012). Although not investigated in detail, the author found proof that the translation quality for the under-resourced languages at that time was poor, which was expected for under-resourced languages. This may explain the similar performance between the language-dependent and language-independent methods. At the end of the study, however, Google Translate is not a freely available service anymore. Therefore, no newer baseline was used in this study, however, the author expected the translation quality of these languages to have improved since then.

Some limitations of these methods require languages to use Latin alphabets, or a transliteration is required prior to using these features. Furthermore, languages which do not contain clear word separation will need to be pre-processed prior to measuring the word count. These approaches also rely on the use of Wikipedia and Wiktionary, which at the time of writing, are available in over 180 languages. Since these resources are growing all the time, the method can be expected to work for more languages in the future. Finally, the use of these approaches were tested in eight language pairs (seven of them under-resourced). The use of these approaches for languages that are, non-similar, do not contain many Wikipedia internal links, or inter-language links, however, will still need to be investigated in the future.

11.3 Limitations

11.3.1 Corpus limitations

The previous sections have illustrated how the Wikipedia similarity corpus enables a better understanding of characteristics of similarity in Wikipedia articles, and how 'similarity' relates to 'comparability', specifically in Wikipedia documents. Whilst the corpus contributed to a better understanding of similarity in Wikipedia in a larger set of topics and number of languages compared to previous works, it also has some limitations with regards to the size of dataset and how the document pairs were selected.

Firstly, as shown in Chapter 5, a stratified sampling based on the anchor text and word overlap in the document pairs were used to ensure that the selected document pairs represent different degrees of similarity. Whilst this purpose was achieved, this approach also introduced a bias into the evaluation dataset as the distribution of similarity of the document pairs might not represent the overall distribution of Wikipedia in general.

Furthermore, there are limitations of the corpus with regards to the short length of the document pairs (fewer than 1,000 words). This limitation was carried out in order to avoid annotator fatigue when annotating all document pairs. Again, this has introduced a bias because the overall Wikipedia articles may be longer on average.

Due to these limitations, the similarity degrees exhibited by the evaluation corpus are not be used to represent the similarity of Wikipedia in general. Instead, its purpose is to include document pairs with a wide range of similarity, that allows the author to investigate the characteristics behind Wikipedia document pairs of different similarity degree. Another purpose of the corpus is to evaluate the different approaches against a gold-standard dataset. Both these purposes were able to be completed using the evaluation set. Future work is required to add more document pairs that are more representative of the Wikipedia natures in order to produce a more robust evaluation of the approaches, such as by applying the classifier to identify more Wikipedia articles of varying similarity degrees to improve the size of the evaluation corpus.

11.3.2 Method limitations

The experiments explored in this study have investigated the use of a number of language-independent features, both to measure the content similarity and structure similarity, to identify similarity in Wikipedia. Most of these features can be extracted without the need of any linguistic approaches. Some features require the use of Wikipedia and Wiktionary (which are available in a large number of languages) for assisting the translation process in order to identify overlapping information across languages. In this section, the author identified some limitations relating to the methods, such as the required resoures and their applicability to languages and other sites.

Wikipedia interlanguage links

Some of these features rely on the use of interlanguage links in Wikipedia. The author did not investigate whether there was a minimum number of interlanguage links for the approaches to work. In this study, LV-EN contains the fewest number of interlanguage-linked articles (in this case, 21,302). Since the link overlap methods were not tested for languages with fewer than 21,302 links, the author will assume that this is the smallest

Table 11.1 Comparison of interlanguage-linked articles in 2010 and 2018

Language Pair	**Total Interlanguage-linked Articles**	
	2010	**2018**
DE-EN	637,382	1,278,776 (⇑ 101%, ⇑ 641,394)
EL-EN	36,752	153,951 (⇑ 319%, ⇑ 117,199)
ET-EN	42,008	126,144 (⇑ 200%, ⇑ 84,136)
HR-EN	51,432	135,614 (⇑ 164%, ⇑ 84,182)
LT-EN	57,954	128,975 (⇑ 123%, ⇑ 71,021)
LV-EN	21,302	85,148 (⇑ 300%, ⇑ 63,846)
RO-EN	97,815	351,183 (⇑ 259%, ⇑ 253,368)
SL-EN	51,332	157,288 (⇑ 206%, ⇑ 105,956)

number required for the method to work.

By September 2018, 312 Wikipedia language versions contain at least one interlanguage-linked article to English. The numbers of interlanguage links, however, vary drastically between languages; Min=1, Max=1,537,198, Mean=88,375.61 (SD=219,938.95). A third of these language pairs (107 language pairs) contain over 21,302 links. Table 11.1 shows a comparison of the number of interlanguage links between these language pairs in 2010 and 2018. The numbers show that the links between these language pairs have increased significantly. E.g., the number of LV-EN articles, although is still the fewest, has now quadrupled to over 85K (from 21K). The rest of the language pairs now contain between 2-4 times as many links as the 2010 links. These numbers are also expected to increase; therefore, they allow the links overlap to be applied to more languages in the future.

Wiktionary

For section heading similarity, the use of Wiktionary is also required. At the time of writing,[1] Wiktionary contains 174 language versions. The English Wiktionary contains 5,942,802 entries, although the translation available for the rest of the languages may differ significantly.

Alternatively, other general dictionaries such as Wikidata can be used instead. By September 2018, the English Wikidata contained descriptions for over 50M items. Simi-

[1] `https://meta.wikimedia.org/wiki/Wiktionary/Table`, accessed on 11 February 2019

larly, each item may also contain its translations in multiple languages. However, Wikidata (similar to Wikipedia) contains mostly concepts, while Wiktionary also contains verbs to assist with the translation process. The use of other dictionaries in this study should be investigated as future work.

Nature of languages

When the two above resources (Wikipedia interlanguage links and Wiktionary) are not available, the use of the classification or regression approach may work by relying on the remaining language-independent features, such as word length ratio and char-n-gram overlap. These require the languages to be written in Latin characters. Alternatively, a transliteration tool is required prior to using these methods. Char-n-gram overlap itself is expected to work better on similar languages; however, this was not investigated in detail. The word length ratio feature is expected to work effectively for languages where there is a clear separation between words (i.e., a whitespace). For languages where there are no clear separation between words, some pre-processing tasks are required to tokenise the words before the word length ratio feature can be calculated.

Application to other sites

Finally, the work presented in this study was evaluated only on Wikipedia documents. Its performance on identifying similarity in non-Wikipedia sites, e.g., news or Web articles, was not evaluated. With regards to the features, some features such as char-n-gram overlap and length ratio are expected to be transferable for other domains. However, there are other features which are specifically used in Wikipedia and may not exist in other domains, such as the availability of section titles and the links within the contents. As a result, the use of these features for non-Wikipedia sites may not be suitable.

11.4 Future work

There are a few avenues that can be pursued for future work in this area. Firstly, recent work in this area have found promising results with the use of more advanced features such as word embeddings and topic modelling. Although most of these approaches are not language-dependent, more approaches in the recent few years have explored the possibility of training them without the use of language-dependent resources. Incorporating these features into the classification approach will be investigated as future work.

Secondly, the author plans to apply these methods for more languages to evaluate how well they perform. This will also require the evaluation corpus to be expanded in order to evaluate the performance of these methods.

Thirdly, the author plans to further investigate the usability of this method when applied to tasks that utilise Wikipedia. In particular, the author is interested in investigating whether increasing the similarity in Wikipedia corpus can further increase the accuracy of subsequent approaches, such as CL-ESA, multilingual word embeddings, etc., which utilised Wikipedia as a corpus.

Finally, the findings in this study have provided some insights on the similarity and dissimilarity in Wikipedia. Future work will be required to further understand the degree of similarity in Wikipedia, if this differs between different domains. More understanding on the reasons behind similarity of Wikipedia articles (e.g., different point of views, political/cultural background) will also be very useful to understand.

References

Adafre, S. F., & de Rijke, M. (2006). Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources.*

Adar, E., Skinner, M., & Weld, D. S. (2009). Information arbitrage across multi-lingual Wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 94–103). New York, NY, USA: ACM.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19–27). Stroudsburg, PA, USA: Association for Computational Linguistics.

Agirre, E., Diab, M., Cer, D., & Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (pp. 385–393).

Aker, A., Kanoulas, E., & Gaizauskas, R. J. (2012). A light way to collect comparable corpora from the Web. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (pp. 15–20).

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop* (Vol. 1998, pp. 194–218).

Argaw, A. A., & Asker, L. (2005). Web mining for an Amharic-English bilingual corpus. In *Proceedings of the 1st International Conference on Web Information Systems and Technologies* (pp. 239–246). INSTICC Press.

Artetxe, M., Labaka, G., & Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 451–462).

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). ACM press New York.

Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., & Gergle, D. (2012). Omnipedia: Bridging the Wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1075–1084). New York, NY, USA: ACM.

Bär, D., Biemann, C., Gurevych, I., & Zesch, T. (2012). UKP: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (pp. 435–440).

Barker, E., & Gaizauskas, R. (2012). Assessing the comparability of news texts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (p. 3996-4003).

Barrón-Cedeño, A., Paramita, M. L., Clough, P., & Rosso, P. (2014). A comparison of approaches for measuring cross-lingual similarity of Wikipedia articles. In *Proceedings of the 36th European Conference on IR Research* (pp. 424–429). Springer.

Barrón-Cedeño, A., Rosso, P., Agirre, E., & Labaka, G. (2010). Plagiarism detection across distant language pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 37–45). Stroudsburg, PA, USA: Association for Computational Linguistics.

Barzilay, R., & Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 25–32). Stroudsburg, PA, USA: Association for Computational Linguistics.

Benedetti, F., Beneventano, D., Bergamaschi, S., & Simonini, G. (2018). Computing inter-document similarity with context semantic analysis. *Information Systems.*

Bharadwaj, R. G., & Varma, V. (2011a). Language-independent context aware query translation using Wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web* (pp. 145–150).

Bharadwaj, R. G., & Varma, V. (2011b). Language independent identification of parallel sentences using Wikipedia. In *Proceedings of the 20th International Conference Companion on World Wide Web* (pp. 11–12). New York, NY, USA: ACM.

Bigi, B. (2003). Using Kullback-Leibler Distance for text categorization. In F. Sebastiani (Ed.), *Advances in information retrieval* (Vol. 2633, p. 305-319). Springer Berlin Heidelberg.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Bollegala, D., Kontonatsios, G., & Ananiadou, S. (2015). A cross-lingual similarity measure for detecting biomedical term translations. *PloS one*, *10*(6), e0126196.

Brants, T., & Stolle, R. (2002). Finding similar documents in document collections. In *Using Semantics for Information Retrieval and Filtering: State of the Art and Future Research Workshop at LREC-2002.*

Braschler, M., & Schäuble, P. (1998). Multilingual information retrieval based on document alignment techniques. In *Research and Advanced Technology for Digital Libraries* (pp. 183–197). Springer.

Brin, S., Davis, J., & Garcia-Molina, H. (1995). Copy detection mechanisms for digital documents. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data* (Vol. 24, pp. 398–409).

Bronner, A., Negri, M., Mehdad, Y., Fahrni, A., & Monz, C. (2012). Cosyne: Synchronizing multilingual wiki content. In *Proceedings of the Eighth Annual International*

Symposium on Wikis and Open Collaboration (p. 33).

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics, 19*(2), 263–311.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., & Navigli, R. (2017). Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 15–26).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with naïve bayes. *Expert Systems with Applications, 36*(3, Part 1), 5432 - 5435.

Chu, C., Dabre, R., & Kurohashi, S. (2016). Parallel sentence extraction from comparable corpora with neural network features. In *Proceedings of 10th edition of the Language Resources and Evaluation Conference* (p. 2931-2935).

Clark, M., Ruthven, I., & Holt, P. O. (2009). The evolution of genre in Wikipedia. *Journal of Language Technology and Computational Linguistis (JLCL 2009), 24*(1), 1–22.

Clough, P., Gaizauskas, R., Piao, S. S., & Wilks, Y. (2002). Meter: Measuring text reuse. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 152–159).

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Cooper, J. W., Coden, A. R., & Brown, E. W. (2002). Detecting similar documents using salient terms. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (pp. 245–251). New York, NY, USA: ACM.

Cosma, A. E. (2015). *Semiautomatic completion of Wikipedia contents with domain-specific mt and clir* (Unpublished master's thesis). Universitat Politècnica de

Catalunya.

Creswell, J. W. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage publications.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 708–716).

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). A framework and graphical development environment for robust nlp tools and applications. In *Acl* (pp. 168–175).

Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL recognising textual entailment challenge. In J. Quiñonero-Candela, I. Dagan, B. Magnini, & F. d'Alché Buc (Eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment* (pp. 177–190). Berlin, Heidelberg: Springer Berlin Heidelberg.

Denkowski, M., & Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the 6th Workshop on Statistical Machine Translation* (pp. 85–91).

Do, Q., Roth, D., Sammons, M., Tu, Y., & Vydiswaran, V. (2009). Robust, light-weight approaches to compute lexical similarity. *Computer Science Research and Technical Reports, University of Illinois, 9*.

Dolan, W. B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the third international workshop on paraphrasing (iwp2005).*

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the 7th International Conference on Information and Knowledge Management* (pp. 148–155).

Dumais, S. T. (2007). Lsa and information retrieval: Getting back to basics. *Handbook of Latent Semantic Analysis*, 293–321.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 281–285).

Dumais, S. T., Letsche, T. A., Littman, M. L., & Landauer, T. K. (1997). Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval* (Vol. 15, p. 21).

Elsayed, T., Lin, J., & Oard, D. W. (2008). Pairwise document similarity in large collections with mapreduce. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (pp. 265–268).

Erdmann, M., Nakayama, K., Hara, T., & Nishio, S. (2008). An approach for extracting bilingual terminology from Wikipedia. In J. Haritsa, R. Kotagiri, & V. Pudi (Eds.), *Database Systems for Advanced Applications* (Vol. 4947, pp. 380–392). Springer Berlin Heidelberg.

Erdmann, M., Nakayama, K., Hara, T., & Nishio, S. (2009). Improving the extraction of bilingual terminology from Wikipedia. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *5*(4), 31:1–31:17.

Fernando, S., & Stevenson, M. (2008). A semantic similarity approach to paraphrase detection. In *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguitics* (pp. 45–52).

Filatova, E. (2009). Directions for exploiting asymmetries in multilingual Wikipedia. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies* (pp. 30–37). Stroudsburg, PA, USA: Association for Computational Linguistics.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*(1), 116–131.

Frank, E., & Hall, M. (2001). A simple approach to ordinal classification. In L. De Raedt & P. Flach (Eds.), *Machine Learning: ECML 2001* (pp. 145–156). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fung, P., & Cheung, P. (2004). Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 57–63).

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (pp. 1606–1611). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics, 19*(1), 75–102.

Gamallo, P., & Garcia, M. (2012). Extraction of bilingual cognates from Wikipedia. In *Computational Processing of the Portuguese Language* (pp. 63–72). Springer.

Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., & Huang, C. (2001). Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 96–104).

Gottschalk, S., & Demidova, E. (2016). Analysing temporal evolution of interlingual Wikipedia article pairs. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 1089–1092).

Gottschalk, S., & Demidova, E. (2017). Multiwiki: Interlingual text passage alignment in Wikipedia. *ACM Transactions on the Web (TWEB), 11*(1), 6:1–6:30.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational evaluation and policy analysis, 11*(3), 255–274.

Greene, J. C., & Hall, J. N. (2010). Dialectics and pragmatism: Being of consequence. *Handbook of mixed methods in social and behavioral research*, 119–144.

Grefenstette, G., & Tapanainen, P. (1994). *What is a word, what is a sentence?: Problems of tokenisation.* Rank Xerox Research Centre Meylan.

Gupta, P., Barrón-Cedeno, A., & Rosso, P. (2012). Cross-language high similarity search using a conceptual thesaurus. In *Information Access Evaluation, Multilinguality,* *Multimodality, and Visual Analytics* (pp. 67–75). Springer.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(Mar), 1157–1182.

Hall, M. A. (1998). *Correlation-based feature subset selection for machine learning* (Unpublished doctoral dissertation). University of Waikato, Hamilton, New Zealand.

Hassan, S., & Mihalcea, R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3* (pp. 1192–1201). Stroudsburg, PA, USA: Association for Computational Linguistics.

Hersh, W. R., Bhuptiraju, R. T., Ross, L., Johnson, P., Cohen, A. M., & Kraemer, D. F. (2004). Trec 2004 genomics track overview. In *Proceedings of the Thirteenth Text Retrieval Conference* (pp. 14–24).

Hoad, T. C., & Zobel, J. (2003). Methods for identifying versioned and plagiarized documents. *Journal of the American Society for Information Science and Technology, 54*(3), 203–215.

Holloway, T., Bozicevic, M., & Börner, K. (2007). Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity, 12*(3), 30–40.

Hu, M., Lim, E.-P., Sun, A., Lauw, H. W., & Vuong, B.-Q. (2007). Measuring article quality in Wikipedia: models and evaluation. In *Proceedings of the sixteenth ACM Conference on Information and Knowledge Management* (pp. 243–252).

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference* (pp. 49–56).

Huang, L., Milne, D., Frank, E., & Witten, I. H. (2012). Learning a concept-based document similarity measure. *Journal of the American Society for Information Science and Technology, 63*(8), 1593–1608.

Ion, R., Tufiş, D., Boroş, T., Ceauşu, A., & Ştefănescu, D. (2010). On-line compilation of comparable corpora and their evaluation. In *Proceedings of the 7th International Conference of Formal Approaches to South Slavic and Balkan Languages* (pp. 29–34).

Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD), 2*(2), 10.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New phytologist, 11*(2), 37–50.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis, 6*(5), 429–449.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics.*

Jiang, Y., Zhang, X., Tang, Y., & Nie, R. (2015). Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. *Information Processing & Management, 51*(3), 215–234.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning* (pp. 137–142).

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher, 33*(7), 14–26.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing.* Pearson London.

Kandola, J., Cristianini, N., & Shawe-Taylor, J. S. (2003). Learning semantic similarity. In *Advances in Neural Information Processing Systems* (pp. 673–680).

Kim, S., Park, S., Hale, S. A., Kim, S., Byun, J., & Oh, A. H. (2016). Understanding editing behaviors in multilingual Wikipedia. *PloS one, 11*(5), e0155305.

Koberstein, J., & Ng, Y.-K. (2006). Using word clusters to detect similar Web documents. In J. Lang, F. Lin, & J. Wang (Eds.), *Knowledge Science, Engineering and Management* (pp. 215–228). Berlin, Heidelberg: Springer Berlin Heidelberg.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Mt summit* (Vol. 5, pp. 79–86).

Koehn, P. (2009). *Statistical machine translation.* Cambridge University Press.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology.* Sage Publications.

Lakkaraju, P., Gauch, S., & Speretta, M. (2008). Document similarity based on concept tree distance. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia* (pp. 127–132). New York, NY, USA: ACM.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning* (pp. 1188–1196).

Lee, C., & Lee, G. G. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. *Information Processing & Management, 42*(1), 155–165.

Lee, M. D., Pincombe, B., & Welsh, M. (2005). An empirical evaluation of models of text document similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 27).

Lee, S., & Choeh, J. Y. (2014). Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications, 41*(6), 3041 - 3046.

Li, B., & Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 644–652).

Lin, W.-P., Snover, M., & Ji, H. (2011). Unsupervised language-independent name translation mining from Wikipedia infoboxes. In *Proceedings of the First Workshop on Unsupervised Learning in NLP* (pp. 43–52).

Lin, Y., Jiang, J., & Lee, S. (2014). A similarity measure for text classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 26*(7), 1575-1590.

Maia, B. (2003). What are comparable corpora? In *Proceedings of Pre-Conference workshop Multilingual Corpora: Linguistic Requirements and Technical Perspectives, Corpus Linguistics* (pp. 27–34).

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge University Press, Cambridge.

Maurer, H. A., Kappe, F., & Zaka, B. (2006). Plagiarism - a survey. *Journal of Universal*

Computer Science, *12*(8), 1050–1084.

McEnery, A., & Xiao, Z. (2007). Parallel and comparable corpora: What are they up to? In *Incorporating Corpora: Translation and the Linguist. Translation Europe. Chap XX.* Multilingual Matters.

McNamee, P., & Mayfield, J. (2004). Character n-gram tokenization for european language text retrieval. *Information Retrieval*, *7*(1-2), 73–97.

Mehdad, Y., Negri, M., & Federico, M. (2010). Towards cross-lingual textual entailment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 321–324).

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st Conference on Artificial Intelligence* (Vol. 6, pp. 775–780).

Mihalcea, R., & Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management* (pp. 233–242).

Mikolov, T., Le, Q. V., & Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. *CoRR*, *abs/1309.4168*. Retrieved from `http://arxiv.org/abs/1309.4168`

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Milios, E., Zhang, Y., He, B., & Dong, L. (2003). Automatic term extraction and document similarity in special text corpora. In *Proceedings of the 6th Conference of the Pacific Association for Computational Linguistics* (pp. 275–284).

Miller, G. A. (1995). Wordnet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Milne, D. (2007). Computing semantic relatedness using Wikipedia link structure. In *Proceedings of the New Zealand Computer Science Research Student Conference.*

Milne, D., & Witten, I. H. (2008). Learning to link with Wikipedia. In *Proceedings of the*

17th ACM Conference on Information and Knowledge Management (pp. 509–518).

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2* (pp. 880–889).

Mogadala, A., & Rettinger, A. (2016). Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 692–702).

Mohammadi, M., & GhasemAghaee, N. (2010). Building bilingual parallel corpora based on Wikipedia. In *Proceedings of the Second International Conference on Computer Engineering and Applications* (Vol. 2, pp. 264–268).

Muhr, M., Kern, R., Zechner, M., & Granitzer, M. (2010). External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system. In *Notebook Papers of CLEF 2010 LABs and Workshops.*

Müller, C., & Gurevych, I. (2009). Using Wikipedia and Wiktionary in domain-specific information retrieval. In *Proceedings of CLEF 2008* (pp. 219–226).

Munteanu, D. S., Fraser, A., & Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 265–272).

Munteanu, D. S., & Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics, 31*(4), 477–504.

Nakayama, K., Hara, T., & Nishio, S. (2007). Wikipedia mining for an association web thesaurus construction. In *Web Information Systems Engineering - WISE 2007* (pp. 322–334). Springer.

Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., & Giampiccolo, D. (2012). Semeval-2012 task 8: cross-lingual textual entailment for content synchronization. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2:*

Proceedings of the Sixth International Workshop on Semantic Evaluation (pp. 399–407).

Ni, X., Sun, J.-T., Hu, J., & Chen, Z. (2009). Mining multilingual topics from Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web* (pp. 1155–1156). New York, NY, USA: ACM.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, *29*(1), 19–51.

Otero, P. G., & López, I. G. (2010). Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora* (pp. 21–25).

Paepcke, A., Garcia-Molina, H., Rodriguez-Mula, G., & Cho, J. (2000). Beyond document similarity: Understanding value-based search and browsing technologies. *SIGMOD Rec.*, *29*(1), 80–92.

Paramita, M. L., Clough, P., Aker, A., & Gaizauskas, R. J. (2012). Correlation between similarity measures for inter-language linked Wikipedia articles. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (pp. 790–797).

Paramita, M. L., Clough, P., & Gaizauskas, R. (2017). Using section headings to compute cross-lingual similarity of Wikipedia articles. In *Proceedings of the 39th European Conference on Information Retrieval* (pp. 633–639).

Patry, A., & Langlais, P. (2011). Identifying parallel documents from a large bilingual collection of texts: application to parallel article extraction in Wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web* (pp. 87–95). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004* (pp. 38–41).

Peters, C., Braschler, M., & Clough, P. (2012). *Multilingual information retrieval: From research to practice.* Springer Science & Business Media.

Ponzetto, S. P., & Strube, M. (2007). Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, *30*(1), 181–212.

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation*, *45*(1), 45–62.

Potthast, M., Stein, B., & Anderka, M. (2008). A Wikipedia-based multilingual retrieval model. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. White (Eds.), *Advances in information retrieval* (Vol. 4956, p. 522-530). Springer Berlin Heidelberg.

Pouliquen, B., Steinberger, R., Ignat, C., Käsper, E., & Temnikova, I. (2004). Multilingual and cross-lingual news topic tracking. In *Proceedings of the 20th International Conference on Computational Linguistics.* Stroudsburg, PA, USA: Association for Computational Linguistics.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 1995 International Joint Conference on Artificial Intelligence.*

Resnik, P. (1999). Mining the Web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 527–534).

Resnik, P., & Smith, N. A. (2003). The Web as a parallel corpus. *Computational Linguistics*, *29*(3), 349–380.

Ruder, S., Vulić, I., & Søgaard, A. (2017). A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902.*

Saad, M., Langlois, D., & Smaïli, K. (2014). Cross-lingual semantic similarity measure for comparable articles. In *International Conference on Natural Language Processing* (pp. 105–115).

Sadat, F. (2010). Exploiting a multilingual Web-based encyclopedia for bilingual terminology extraction. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation* (pp. 519–526).

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval.* New York, NY, USA: McGraw-Hill, Inc.

Schönhofen, P., Benczúr, A., Bíró, I., & Csalogány, K. (2008). Cross-language retrieval with Wikipedia. In C. Peters et al. (Eds.), *Advances in multilingual and multimodal information retrieval* (Vol. 5152, p. 72-79). Springer Berlin Heidelberg.

Shivakumar, N., & Garcia-Molina, H. (1995). SCAM: A copy detection mechanism for digital documents. In *Proceedings of the 2nd International Conference in Theory and Practice of Digital Libraries (DL 1995).*

Simard, M., Foster, G. F., & Isabelle, P. (1993). Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research: Distributed Computing-Volume 2* (pp. 1071–1082).

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., … Pinnis, M. (2012). Collecting and using comparable corpora for statistical machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation.*

Smith, J. R., Quirk, C., & Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 403–411). Stroudsburg, PA, USA: Association for Computational Linguistics.

Søgaard, A., Agić, Ž., Alonso, H. M., Plank, B., Bohnet, B., & Johannsen, A. (2015). Inverted indexing for cross-lingual NLP. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing.*

Sorg, P., & Cimiano, P. (2008). Cross-lingual information retrieval with explicit semantic analysis. In *Working Notes for the CLEF 2008 Workshop.*

Štajner, T., & Mladenić, D. (2018). Cross-lingual document similarity estimation and dictionary generation with comparable corpora. *Knowledge and Information Systems*, 1–15.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlüter, P. (2012). DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th Interna-*

tional Conference on Language Resources and Evaluation (pp. 454–459).

Steinberger, R., Pouliquen, B., & Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 415–424). Springer.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (pp. 2142–2147).

Strube, M., & Ponzetto, S. P. (2006). Wikirelate! computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2* (pp. 1419–1424). AAAI Press.

Taieb, M. A. H., Aouicha, M. B., & Hamadou, A. B. (2014). Ontology-based approach for measuring semantic similarity. *Engineering Applications of Artificial Intelligence, 36*, 238–261.

Tomás, J., Bataller, J., Casacuberta, F., & Lloret, J. (2008). Mining Wikipedia as a parallel and comparable corpus. In *Language Forum* (Vol. 1, p. 34).

Trivison, D. (1987). Term co-occurrence in cited/citing journal articles as a measure of document similarity. *Information Processing & Management*, *23*(3), 183–194.

Tsai, C.-T., & Roth, D. (2016). Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 589–598).

Turdakov, D., & Velikhov, P. (2008). Semantic relatedness metric for Wikipedia concepts based on link analysis and its application to word sense disambiguation. In *Proceedings of the Spring Young Researcher's Colloquium On Database and Information Systems.* Citeseer.

Ture, F., Elsayed, T., & Lin, J. (2011). No free lunch: Brute force vs. locality-sensitive hashing for cross-lingual pairwise similarity. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp.

943–952). New York, NY, USA: ACM.

Tyers, F. M., & Pienaar, J. A. (2008). Extracting bilingual word pairs from Wikipedia. In *Proceedings of the SALTMIL Workshop at the Language Resources and Evaluation Conference* (pp. 19–22).

United Nations. (2006). *ODS UN parallel corpus.* `http://ods.un.org`.

Uszkoreit, J., Ponte, J. M., Popat, A. C., & Dubiner, M. (2010). Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 1101–1109). Stroudsburg, PA, USA: Association for Computational Linguistics.

Vilarino, D., Pinto, D., Tovar, M., León, S., & Castillo, E. (2012). BUAP: Lexical and semantic similarity for cross-lingual textual entailment. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (pp. 706–709).

Voorhees, E. M. (2001). Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 74–82). New York, NY, USA: ACM.

Voss, J. (2005). Measuring Wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics.*

Vulić, I., & Moens, M.-F. (2013). Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 106–116).

Vulić, I., & Moens, M.-F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 363–372).

Vulić, I., & Moens, M.-F. (2016). Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, *55*, 953–994.

Wan, X. (2007). A novel document similarity measure based on earth mover's distance. *Information Sciences, 177*(18), 3718–3730.

Wan, X., & Peng, Y. (2005). The earth mover's distance as a semantic measure for document similarity. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (pp. 301–302). New York, NY, USA: ACM.

Wäschle, K., & Fendrich, S. (2012). HDU: Cross-lingual textual entailment with SMT features. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (pp. 467–471).

Wu, Z., Zhu, H., Li, G., Cui, Z., Huang, H., Li, J., . . . Xu, G. (2017). An efficient Wikipedia semantic matching approach to text document classification. *Information Sciences, 393*, 15–28.

Xu, B., Guo, X., Ye, Y., & Cheng, J. (2012). An improved random forest classifier for text categorization. *Journal of Computers, 7*(12), 2913–2920.

Yapomo, M., Bernhard, D., & Gançarski, P. (2015). Comparison of crosslingual similarity measures for multilingual documents clustering. *12ème atelier sur la Fouille de Données Complexes (FDC) Extraction et Gestion des Connaissances (EGC 2015).*

Yasuda, K., & Sumita, E. (2008). Method for building sentence-aligned corpus from Wikipedia. In *Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy* (pp. 64–66). USA: Association for the Advancement of Artificial Intelligence.

Yeh, E., Ramage, D., Manning, C. D., Agirre, E., & Soroa, A. (2009). WikiWalk: Random walks on Wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing* (pp. 41–49). Stroudsburg, PA, USA: Association for Computational Linguistics.

Yu, K., & Tsujii, J. (2009). Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *Proceedings of Human Language Technologies: The*

2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers (pp. 121–124). Stroudsburg, PA, USA.

Zanettin, F. (1998). Bilingual comparable corpora and the training of translators. *Translators' Journal*, *43*(4), 616–630.

Zesch, T., & Gurevych, I. (2010). The more the better? Assessing the influence of Wikipedia's growth on semantic relatedness measures. In *Proceedings of the International Conference on Language Resources and Evaluation* (pp. 1374–1380).

Zesch, T., Gurevych, I., & Mühlhäuser, M. (2007). Analyzing and accessing Wikipedia as a lexical semantic resource. *Data Structures for Linguistic Resources and Applications*, 197–205.

Zhang, Y., Wu, K., Gao, J., & Vines, P. (2006). Automatic acquisition of Chinese–English parallel corpus from the Web. In *Advances in information retrieval* (pp. 420–431). Springer.