A Cross-domain and Cross-language Knowledge-based Representation of Text and its Meaning

PhD Thesis

Marc Franco Salvador

Supervised by Paolo Rosso

February, 2017

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

This research has been carried out in the framework of the European Commission project WIQ-EI IRSES (no. 269180), and the national projects DIANA-APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01), *Destilado de opiniones desde contenidos generados por usuarios* (TIN2011-14726-E), and SomEMBED: SOcial Media language understanding - EMBEDing contexts (TIN2015-71147-C2-1-P).

The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom.

— Isaac Asimov

Acknowledgments

I would like to give my most sincere thanks to some people without which would not have been possible to carry out this work.

First, to all the people who contributed to make me a better professional and researcher. My advisor, Paolo Rosso, who helped and guided me during these four years. The people who invited and advised me during my internships: Roberto Navigli (Sapienza University of Rome, Italy; 1 year), José Antorio Troyano and Fermín Cruz (University of Seville, Spain; 3 months), Yassine Benajiba and Lutz Lukas (Symanto Group, Germany; 3 months), Vasudeva Varma and Prasad Pingali (IIIT Hyderabad and Veooz, India; 2 months), Manuel Montes-y-Gómez (INAOE, Mexico; 1 month), and Vivek Singh (South Asian University, India) for inviting me to give a tutorial on knowledge graph-based natural language processing in the International Workshop on Data and Text Analytics. My research colleagues Enrique Flores, Juanma Cotelo, Adrián Pastor, Stefano Faralli and Aditya Joshi. The reviewers of this thesis: Simone Paolo Ponzetto, Alessandro Moschitti, and Rada Mihalcea; thanks for your work and kind words about my thesis. The members of the evaluation tribunal of this thesis: Simone Paolo Ponzetto, Bernardo Magnini, and Nicola Ferro. Also to all the people of the PRHLT research center and all those acknowledged in my publications.

Next, I would like to thank the closest people, the ones who heard my complains and gave me support during these years. My parents, Pepe and Rosa, who were always there. My girlfriend Julia, who always listened and encouraged me. Also to the friends who were interested in my mood and work during my PhD thesis.

Finally, thanks to everyone who I forgot to name in this section and has influenced somehow in the success of this large period of my life.

Valencia, February 2017

Abstract

Natural Language Processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human languages. One of its most challenging aspects involves enabling computers to derive meaning from human natural language. To do so, several meaning or context representations have been proposed with competitive performance. However, these representations still have room for improvement when working in a cross-domain or cross-language scenario.

In this thesis we study the use of knowledge graphs as a cross-domain and cross-language representation of text and its meaning. A knowledge graph is a graph that expands and relates the original concepts belonging to a set of words. We obtain its characteristics using a wide-coverage multilingual semantic network as knowledge base. This allows to have a language coverage of hundreds of languages and millions human-general and -specific concepts.

As starting point of our research we employ knowledge graph-based features — along with other traditional ones and meta-learning — for the NLP task of single-domain and cross-domain polarity classification. The analysis and conclusions of that work provide evidence that knowledge graphs capture meaning in a domain-independent way. The next part of our research takes advantage of the multilingual semantic network and focuses on cross-language Information Retrieval (IR) tasks. First, we propose a fully knowledge graph-based model of similarity analysis for cross-language plagiarism detection. Next, we improve that model to cover out-of-vocabulary words and verbal tenses and apply it to cross-language document retrieval, categorisation, and plagiarism detection. Finally, we study the use of knowledge

graphs for the NLP tasks of community questions answering, native language identification, and language variety identification.

The contributions of this thesis manifest the potential of knowledge graphs as a cross-domain and cross-language representation of text and its meaning for NLP and IR tasks. These contributions have been published in several international conferences and journals.

~

El Procesamiento del Lenguaje Natural (PLN) es un campo de la informática, la inteligencia artificial y la lingüística computacional centrado en las interacciones entre las máquinas y el lenguaje de los humanos. Uno de sus mayores desafíos implica capacitar a las máquinas para inferir el significado del lenguaje natural humano. Con este propósito, diversas representaciones del significado y el contexto han sido propuestas obteniendo un rendimiento competitivo. Sin embargo, estas representaciones todavía tienen un margen de mejora en escenarios transdominios y translingües.

En esta tesis estudiamos el uso de grafos de conocimiento como una representación transdominio y translingüe del texto y su significado. Un grafo de conocimiento es un grafo que expande y relaciona los conceptos originales pertenecientes a un conjunto de palabras. Sus propiedades se consiguen gracias al uso como base de conocimiento de una red semántica multilingüe de amplia cobertura. Esto permite tener una cobertura de cientos de lenguajes y millones de conceptos generales y específicos del ser humano.

Como punto de partida de nuestra investigación empleamos características basadas en grafos de conocimiento — junto con otras tradicionales y meta-aprendizaje — para la tarea de PLN de clasificación de la polaridad mono- y transdominio. El análisis y conclusiones de ese trabajo muestra evidencias de que los grafos de conocimiento capturan el significado de una forma independiente del dominio. La siguiente parte de nuestra investigación aprovecha la capacidad de la red semántica multilingüe y se centra en tareas de Recuperación de Información (RI). Primero proponemos un modelo de análisis de similitud completamente basado en grafos de conocimiento para detección de plagio translingüe. A continuación, mejoramos ese modelo para cubrir palabras fuera de vocabulario y tiempos verbales, y lo aplicamos

a las tareas translingües de recuperación de documentos, clasificación, y detección de plagio. Por último, estudiamos el uso de grafos de conocimiento para las tareas de PLN de respuesta de preguntas en comunidades, identificación del lenguaje nativo, y identificación de la variedad del lenguaje.

Las contribuciones de esta tesis ponen de manifiesto el potencial de los grafos de conocimiento como representación transdominio y translingüe del texto y su significado en tareas de PLN y RI. Estas contribuciones han sido publicadas en diversas revistas y conferencias internacionales.

El Processament del Llenguatge Natural (PLN) és un camp de la informàtica, la intel·ligència artificial i la lingüística computacional centrat en les interaccions entre les màquines i el llenguatge dels humans. Un dels seus majors reptes implica capacitar les màquines per inferir el significat del llenguatge natural humà. Amb aquest propòsit, diverses representacions del significat i el context han estat proposades obtenint un rendiment competitiu. No obstant això, aquestes representacions encara tenen un marge de millora en escenaris trans-dominis i trans-llenguatges.

En aquesta tesi estudiem l'ús de grafs de coneixement com una representació trans-domini i trans-llenguatge del text i el seu significat. Un graf de coneixement és un graf que expandeix i relaciona els conceptes originals pertanyents a un conjunt de paraules. Les seves propietats s'aconsegueixen gràcies a l'ús com a base de coneixement d'una xarxa semàntica multilingüe d'àmplia cobertura. Això permet tenir una cobertura de centenars de llenguatges i milions de conceptes generals i específics de l'ésser humà.

Com a punt de partida de la nostra investigació emprem característiques basades en grafs de coneixement — juntament amb altres tradicionals i meta-aprenentatge — per a la tasca de PLN de classificació de la polaritat mono- i trans-domini. L'anàlisi i conclusions d'aquest treball mostra evidències que els grafs de coneixement capturen el significat d'una forma independent del domini. La següent part de la nostra investigació aprofita la capacitat de la xarxa semàntica multilingüe i se centra en tasques de recuperació d'informació (RI). Primer proposem un model d'anàlisi de similitud completament basat en grafs de coneixement per a detecció de plagi trans-llenguatge.

A continuació, vam millorar aquest model per cobrir paraules fora de vocabulari i temps verbals, i ho apliquem a les tasques trans-llenguatges de recuperació de documents, classificació, i detecció de plagi. Finalment, estudiem l'ús de grafs de coneixement per a les tasques de PLN de resposta de preguntes en comunitats, identificació del llenguatge natiu, i identificació de la varietat del llenguatge.

Les contribucions d'aquesta tesi posen de manifest el potencial dels grafs de coneixement com a representació trans-domini i trans-llenguatge del text i el seu significat en tasques de PLN i RI. Aquestes contribucions han estat publicades en diverses revistes i conferències internacionals.

Contents

1 Introduction

Given the vastness of the Web and its still growing size, user and media generated contents are today a reference representation of our culture and knowledge. This has not been overlooked by the industry and the research communities. In recent years, there has been an increase in the efforts to process this information. Natural Language Processing (NLP) (Manning and Schütze, 1999) enables computers to derive meaning from contents written in human natural language. This allows to classify and analyse the information in order to exploit it with several purposes. These include advertising, searching, and education. Related to searching tasks, Information Retrieval (IR) (Baeza-Yates et al., 1999), the activity of obtaining information resources relevant to an information need from a collection of information resources, also increased popularity. Not only search engines benefit from IR but also tasks such as plagiarism detection (Barrón-Cedeño, 2012) and recommendation (Balabanović and Shoham, 1997).

Common text processing methods are domain-dependent and are consequently adapted for concrete tasks. Therefore, the problem is exacerbated when the NLP or IR task is between different domains (Blitzer et al., 2007) or languages (Potthast et al., 2011a). In order to perform in these settings, several methods and text representations have been proposed. Domain adaptation is one of the preferred methods to perform at cross-domain level. In contrast, vector-based models are typically used in the literature for representing documents both in monolingual and cross-language levels.

In this thesis we study the use of knowledge graphs[1] as a cross-domain and cross-language representation of text and its meaning. To do so, we

[1] A knowledge graph is a subset of a semantic network (also known as knowledge base) focused on the concepts belonging to a text, and the intermediate concepts and relations between them. It can also be referred to as semantic annotation or semantic interpretation. The methods based on knowledge graphs can be referred to as semantic tagging-based methods.

generate the knowledge graphs (Mihalcea and Radev, 2011) with BabelNet[2], the widest-coverage multilingual semantic network.[3] This allows to have graphs that expand and relate the original concepts belonging to a set of words. This also provides with a language coverage of hundreds of languages and millions human-general and -specific concepts. The knowledge graph representation is also closely related with the cognitive science and how our mind and its processes represent information (see Section 1.1).

Our research is structured as follows. We first study the use of knowledge graph-based features for the NLP task of single- and cross-domain polarity classification. The analysis and conclusions of that work provided evidence that knowledge graphs capture meaning in a domain-independent way. The next part of our research took advantage of the multilingual semantic network and focused on cross-language IR tasks. First, we proposed a fully knowledge graph-based model of similarity analysis for cross-language plagiarism detection. Next, we improved that model to cover out-of-vocabulary words and verbal tenses and applied it to cross-language document retrieval, categorization, and plagiarism detection. Finally, we studied the use of knowledge graphs for the NLP tasks of community questions answering, native language identification, and language variety identification.

The structure of this chapter is the following. In the next section we study the relationship of the knowledge representations with the cognitive science. We overview the reference methods for cross-domain and cross-language text representation. Next, we motivate our work and present our objectives and research questions. Finally, we present the contributions and the structure of this thesis.

1.1 Knowledge Representations and Cognitive Science

Semantic networks and knowledge graphs have a close relationship with the cognitive science and how mind and its processes represent information. The meaning of language is represented in regions of our mind known as the "semantic system" (Tyler et al., 2003). Huth et al. (2016) showed that the semantic system is organised into intricate patterns that seem to be consistent across individuals. They also provided evidences of areas of that sys-

[2] `http://babelnet.org`

[3] A multilingual semantic network is a (un)directed graph where nodes represent multilingual concepts and edges represent semantic relations between them.

Figure 1.1. Two semantic domains extracted from the semantic system of our brain. Domain (a) seems to be related to life and death. Domain (b) seems to be related to properties of space and materials.

tem selective for specific semantic domains or groups of related concepts (Caramazza and Shelton, 1998; Damasio et al., 1996; Mitchell et al., 2008). These domains were modelled[4] using the whole-brain blood-oxygen-level-dependent responses to stimuli and regression models.[5] Our analysis of the words contained on those domains manifests that there are strong similarities regarding the knowledge graph or the semantic network representation. You can see examples of semantic domains in Figure 1.1. BabelNet and WordNet[6] represent concepts by means of sets of synonyms — known as *synsets*—, which are more explicit and accurate in meaning than these domains. However, the rationale behind representing knowledge using synsets, or in this case concepts of close meaning, is the same: close ideas in context should be close in the representation as well. In addition, if we employ BabelNet to search for some of the words in the domains, we can see direct semantic relationships between many of them, e.g. “pregnant”, “child”, “birth”. It is interesting as well to analyse the words contained in those do-

[4]Online explorer of the semantic system of our brain: `http://gallantlab.org/huth2016/`

[5]Article in The Guardian about this discovery: `https://www.theguardian.com/science/2016/apr/27/brain-atlas-showing-how-words-are-organised-neuroscience`

[6]`https://wordnet.princeton.edu/`

mains, and also in the synsets of the semantic networks, from the perspective of the distributional semantics.

Distributional semantics studies the meaning of words and how their combination gives meaning to texts (Bruni et al., 2012, 2014).[7] The distributional hypothesis originated in linguistics (Yarlett and Ramscar, 2008), has a relationship with cognitive science, especially regarding the abstract representations and the meaning of words (Baroni and Lenci, 2009; McDonald and Ramscar, 2001). In order to generate this abstract representations, some distributional semantic models employ projections to provide with a low-dimensional space (see Section 1.2.2). If we employ one of those models to generate distributed representations of words, the distance between the words contained in the aforementioned domains of our brain is, in general, very short. The same occurs when we measure the distance between the words contained in the synsets. This highlights more the relationship between all these types of representations for modelling the meaning of texts.

In order to illustrate the relationship between semantic networks, knowledge graphs, distributed representations, and the semantic domains inferred by Huth et al. (2016) from the semantic system of the human mind, we show a real example in Figure 1.2. This example contains a two-dimensional projection of the distributed representations of the words of the domains (a) and (b) (cf. Figure 1.1). In addition, we included arrows between the words were exists a semantic relation in BabelNet. As we can see, the words of the domains have been perfectly clustered using distributed representations. This highlights the potential of these type of representations to measure relatedness between words. In addition, these relations between the words have been also captured by BabelNet. The domain (a) has a very strong connection. In contrast, the relatedness of the domain (b) is not obvious when using BabelNet. This manifests that the quality of the inferred domain (b) is not as good as (a), and also highlights the potential of distributed representations to measure relatedness between any pair of words. This type of graph representation, that combines the concepts of a multilingual semantic network and distributed representations to measure relatedness, will be explored in Chapter 3 of this thesis. Although the core of this thesis is focused on the

[7] Multimodal distributional semantics learns the meaning of words not only with text but also with visual words extracted using computer vision techniques.

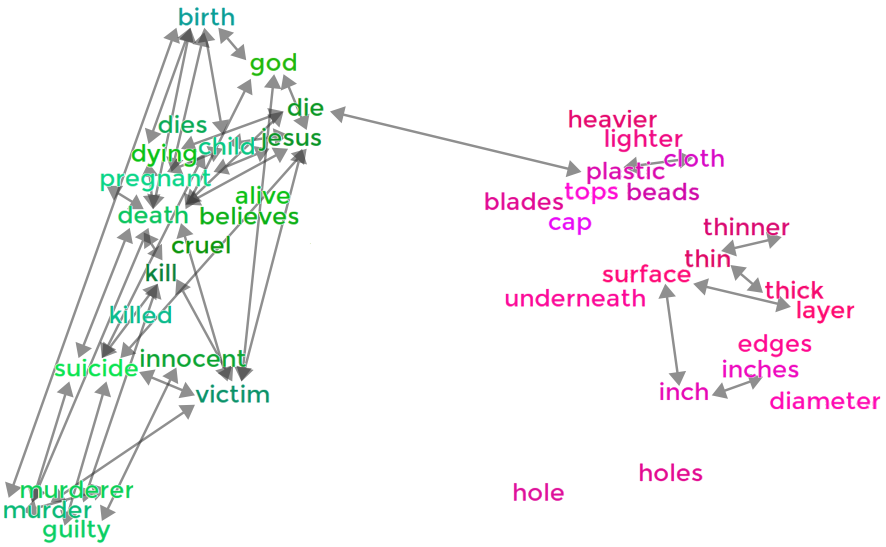

Figure 1.2. Two-dimensional PCA (Jolliffe, 1986) projection of 200-dimensional vectors estimated with the continuous Skip-gram model (Mikolov et al., 2013b). The vectors belong to the words of the semantic domains (a) and (b) extracted from the semantic system of our brain. Arrows represent semantic relations between any synset containing that word in the BabelNet multilingual semantic network.

use of knowledge graphs, we also employ distributed representation-based models for comparison in order to solve some tasks in Chapter 5.

1.2 Cross-domain and Cross-language Text Representations

In this section we first describe the reference techniques of domain adaptation employed for representing text in cross-domain tasks. Next, we describe the state-of-the-art cross-language representations of text. We note that our aim is not to design domain adaptation algorithms, but to design general and effective features. The latter being general have more probability to work across different domains. Indeed, our results in Chapters 2 and 5 show that our models (designed without any specific approach for domain adaptation) outperform other specific domain adaptation models. However, most of the state-of-the-art models in cross-domain classification use domain adaptation. For that reason we start reviewing this kind of methods and representations.

1.2.1 DOMAIN ADAPTATION TEXT REPRESENTATIONS

The most popular cross-domain methods employ standard features — such as unigrams and bigrams — combined with domain adaptation techniques (Ben-David et al., 2010, 2007; Blitzer et al., 2008) to represent texts.[8] These techniques exploit relevant features from the source domains which are also important in the target domain. Using these features is possible to determine the co-occurrence with the target domain features in order to classify its texts.

Domain adaptation has been widely used in cross-domain classification. The Structural Correspondence Learning (SCL) (Blitzer et al., 2006) model selects pivot features frequently appearing in both source and target domains. Then it learns to predict those pivot features using unlabeled data from both domains. Finally, the text representation is obtained by performing a singular value decomposition to reduce the dimensionality. The SCL variant exploiting Mutual Information (SCL-MI) (Blitzer et al., 2007) is employed in tasks such as cross-domain sentiment classification. These tasks require pivot features also to be good predictors of the source label. The mutual information is employed to select the pivots with highest mutual information to those labels. As well in the task of cross-domain sentiment classification, Spectral Feature Alignment (SFA) (Pan et al., 2010) exploits mutual information to differentiate domain-specific and domain-independent features. Next, SFA generates clusters using a spectral clustering over a bipartite graph dividing both types of features. Finally, these clusters are employed to create a feature alignment mapping function that provides with the text representation.

Domain adaptation not only has been successfully employed for classification tasks. Daumé III (2007) proposed a simple and effective supervised domain adaptation[9] model for named-entity recognition, shallow parsing, and part-of-speech tagging. This model uses linear kernels to differentiate domain-specific and domain-independent features from source and target domains. The final representation is in an augmented feature space, which allows learning algorithms to learn the domain adaptation by themselves. As well in part-of-speech tagging and also in classification, Jiang and Zhai (2007) proposed a flexible instance weighting method for domain adaptation

[8] In this thesis we always refer to the semi-supervised version of domain adaptation if not otherwise stated. This version requires annotated data from the source domain, unannotated data from the target domain, and also a "small" set of annotated target domain data.

[9] In the supervised domain adaptation all the considered data are supposed to be labeled.

from a distributional view. This method has the advantage of supporting many different strategies for the adaptation and can be easily extended.

In this thesis we do not employ any domain adaptation or unlabeled data from the target domain. Our approach is focused on proposing new knowledge graph-based features extracted from the source domains that are able to be directly applied to the target domain.

1.2.2 CROSS-LANGUAGE TEXT REPRESENTATIONS

The mainstream representation of texts for monolingual and cross-lingual NLP and IR is vector-based (Manning et al., 2008). The historical representation that set the stage was the Vector Space Model (VSM) (Salton and McGill, 1986). It quantifies the relevance of the terms of a text with a dimension representing each of them.

There are several cross-language representations following the VSM trend. Cross-Language Character n-Grams (CL-CNG) (Mcnamee and Mayfield, 2004), represent documents as vectors of character n-grams. It has proven to obtain good results in cross-language similarity tasks (Potthast et al., 2011a) between languages with lexical and syntactic similarities. In contrast, the Cross-Language Alignment-based Similarity Analysis (CL-ASA) (Barrón-Cedeño et al., 2008; Pinto et al., 2009) model does not need these languages similarities. It employs a statistical dictionary to represent texts as a vector of translated terms.

Note that, because of the variety of terms used in a document collection, a document vector is usually highly dimensional. As a consequence, the resulting computing time may be considerable. To address this issue, several approaches to the reduction of dimensionality of document vectors have been proposed in the literature. A popular class of methods is based on projections, which provide a low-dimensional abstract space referred to as embedding or latent space. This resulting representations, commonly referred as embeddings, distributed or continuous representations, have gained popularity in the NLP and IR community. There are broadly two categories of approaches: (i) generative topic models, and (ii) projection based models. Generative topic models, like Latent Dirichlet Allocation (LDA) (Blei et al., 2003), represent the high dimensional term vectors in a low-dimensional latent space of hidden topics. The projection based methods, like Latent Semantic Indexing

(LSI) (Dumais et al., 1995),[10] or the continuous Skip-gram model (Mikolov et al., 2013b), learn a projection operator to map high-dimensional term vectors to a low-dimensional latent space. There also exist cross-lingual variants of these models which try to learn embeddings of text in a cross-language space.

Inside the generative topic models, we can see adaptations of LDA to perform in a multilingual scenario with models such as Polylingual Topic Models (Mimno et al., 2009), Joint Probabilistic LSA and Coupled Probabilistic LSA (Platt et al., 2010). With respect to projection based models, Cross-Language Latent Semantic Indexing (CL-LSI) (Dumais et al., 1997a) is a cross-lingual extension of LSI. Oriented Principle Component Analysis (OPCA) tries to learn a translingual projection matrix by solving a generalised eigen value problem (Platt et al., 2010). Similarly, Siamese neural network based S2Net learns the same projection matrix through backpropagation error of distance between parallel sentence pairs (Yih et al., 2011). There also exist non-linear deep neural network based solutions to learn such cross-lingual embeddings through deep autoencoders (Gupta et al., 2014; Lauly et al., 2014a,b) and composition neural networks (Gupta et al., 2015). Finally, there also exist knowledge-based projection methods such as Cross-Language Explicit Semantic Analysis (CL-ESA) (Cimiano et al., 2009; Potthast et al., 2011a, 2008). CL-ESA adapts ESA to be used in a cross-language scenario by exploiting the comparable documents across languages from Wikipedia. It represents each document written in a language L by a vector with its similarities with a document collection in the same language L. Using a multilingual document collection with comparable documents across languages, the resulting vectors from different languages can be compared directly.

Note that most of the projection approaches need a high number of training texts to achieve state-of-the-art performance (Platt et al., 2010; Yih et al., 2011). In contrast, other approaches have low performance in cases of similarity with paraphrasing (Franco-Salvador et al., 2016c). In this thesis, we propose a cross-language knowledge graph representation for text which is obtained from a large multilingual semantic network, without using any training information. Our knowledge graph representation explicitly models the semantics and meaning of the text expanding and relating the concepts be-

[10] LSI is Latent Semantic Analysis (LSA) (Deerwester et al., 1990) in the IR context.

longing to its original words. These concepts inherit the multilinguality of the semantic network and are, consequently, able to be compared directly — even if they are created from texts in different languages.

1.3 Motivation and Objectives

The use of distributed representations, and the regularities they capture (Mikolov et al., 2013c), enable to accurately model text meaning and produced significant improvements in NLP tasks (Le and Mikolov, 2014; Mikolov et al., 2013b). However, there is a room of improvement in the cross-language scenario. Most of the approaches need high amounts of data in order to train representative models. In addition, the computational complexity and the amount of training data is proportional to the number of languages employed (Gupta et al., 2014; Platt et al., 2010). On the other hand, the heuristic-, instance-, or knowledge-based cross-language similarity models, which do not employ training data, are designed to only detect verbatim or soft-modified cases of similarity (Franco-Salvador et al., 2016c, 2014b).

The use of knowledge graphs provided with state-of-the-art results in the task of mono- and cross-language Word Sense Disambiguation (WSD) (Navigli and Ponzetto, 2012a). The starting point of this work is the observation that, if these graphs provided with the correct disambiguations of a text, even at cross-language level, they are adequate as representation of the meaning of that text. In addition, we believe that the multilingual semantic network employed to generate the graphs — BabelNet —, with a language coverage of hundreds of languages and millions human-general and -specific concepts, makes this representation domain- and language-independent. Therefore, its complexity is also independent of the number of languages employed. In consequence, knowledge graphs are adequate for cross-domain and cross-language NLP and IR tasks. Moreover, knowledge graphs have several implicit characteristics — WSD, vocabulary expansion, and language independence — that have different impact on their performance in NLP similarity analysis tasks. Finally, we consider that this representation may be useful for other non-cross-language or non-cross-domain NLP tasks such as community questions answering (Nakov et al., 2016), native language identification (Koppel et al., 2005), and language variety identification (Franco-Salvador et al., 2015c).

Considering what aforementioned statements said, this research has the following objectives:

- To study the potential of knowledge graph-based features for cross-domain NLP tasks.
- To develop a cross-language similarity analysis model for NLP and IR tasks.
- To study the knowledge graph characteristics for cross-language similarity analysis tasks.
- To evaluate the performance of the developed approaches and compare them with the state-of-the-art models.
- To employ knowledge graphs for other NLP tasks.

1.4 Research Questions

The aforementioned objectives can be divided in three groups according to the scenario where we employ knowledge graphs: cross-domain, cross-language, and the scenario where we evaluate other NLP tasks. With respect to these groups, the research questions we aim to answer in this thesis are:

Questions about the cross-domain scenario

- *What is the contribution of the knowledge graph-based features for cross-domain NLP tasks?* We are interested in studying the impact of this type of features on cross-domain domain NLP tasks and compare them to traditional ones such as bag of words or n-grams. For this purpose, we have selected the task of cross-domain polarity classification.

Questions about the cross-language scenario

- *What is the contribution of the knowledge graph characteristics in cross-language similarity?* In this thesis we independently study these characteristics to analyse their impact on the task of cross-language plagiarism detection.

- *Could knowledge graphs be employed to successfully solve cross-language similarity tasks?* Aiming to study their potential for this type of tasks, we compare their performance with the state of the art. We evaluate the different models in the cross-language document retrieval, categorization, and plagiarism detection tasks.

Questions about the use of knowledge graphs in other NLP tasks

- *What is the performance of knowledge graphs in other NLP tasks?* We aim to investigate the robustness of knowledge graphs as a general representation of text and its meaning for NLP tasks. To do so, we employ them for the tasks of community questions answering, native language identification, and language variety identification.

1.5 Contributions of this Thesis

Next we summarise the main contributions of this thesis.

From the representation viewpoint, we proved that knowledge graphs can be employed as a cross-domain and cross-language representation of text and its meaning. We employed several reference datasets to show diverse results and comparisons with the state of the art and to justify the validity and potential of this representation. We supported all our conclusions with standard tests of statistical significance of results. In addition, we studied from a theoretical and practical perspective, the main characteristics that contribute to the knowledge graphs performance.

With respect to the tasks, we showed how to obtain state-of-the-art performance with knowledge graphs in several single- and cross-domain NLP and IR tasks: single- and cross-domain polarity classification (Franco-Salvador et al., 2015b), cross-language plagiarism detection (Franco-Salvador et al., 2015a, 2012, 2013a,b, 2014a, 2016a,c), document retrieval and categorization (Franco-Salvador et al., 2014b), and community questions answering (Franco-Salvador et al., 2016b). In addition, we showed the potentiality of knowledge graphs for native language identification (Franco-Salvador et al., 2017) and language variety identification (Franco-Salvador et al., 2015c,d; Rangel et al., 2016).

From the modelling viewpoint, we employed knowledge graphs to obtain state-of-the-art performance in two different ways: (i) as a source of feature extraction for classification and regression, and (ii) as a representation, as part of the proposed cross-language similarity analysis models. With respect to these two models, we proposed one that employs knowledge graphs as representation of the text and its meaning, and we proposed another one that complements that representation with a vector-based representation in order to cover the graph shortcomings. In addition, we proposed a new embedding-based weighting scheme for the semantic relations between the knowledge graph concepts. This scheme proved to outperform the classical one employed in the BabelNet multilingual semantic network.

Finally, some contributions only partially related to knowledge graphs were achieved during this research. First, we proposed the continuous word alignment-based similarity analysis model that notably improved the performance of distributed representations of words in cross-language plagiarism detection. Next, we proved the relationship between the native language and the language variety identification tasks by solving both with the same approach without any task-specific adaptation. The string kernels approach obtained state-of-the-art performance in several datasets of the two tasks. Finally, following our hypothesis of the relationship between knowledge graphs and distributed representations (see Section 1.1), we studied with interesting results how both complement each other for several NLP and IR tasks.

1.6 Structure of this Thesis

This PhD thesis is presented as a compendium of research articles which were published during the study phase of this PhD. We include two international journal articles and an international conference paper as chapters of this work. Next we briefly overview the content of the remaining chapters and appendices:

- **Chapter 2** Cross-domain polarity classification using a knowledge-enhanced meta-classifier

  In this chapter we present our work published in the *Knowledge-Based Systems (KNOSYS)* journal. In that work we employed knowledge

graph-based features, such as WSD- and vocabulary expansion-based ones — along with other traditional ones (bag of words and n-grams) —, for single- and cross-domain polarity classification of product reviews. Experimental results, compared to the state-of-the-art models that employ domain adaptation, show that these types of features capture information in a domain independent way. Moreover, we proved that the type of information obtained from disambiguated concepts is different, and consequently complementary, to the one obtained with the traditional bag-of-words or n-gram features.

- **Chapter 3** A systematic study of knowledge graph analysis for cross-language plagiarism detection

  This chapter is composed by our work published in the *Information Processing & Management (IPM)* journal. This is the reference article of our cross-language similarity model for cross-language knowledge graph analysis. That method employs knowledge graphs as a cross-language representation of the text and its meaning. We also study the implicit and most relevant characteristics — WSD, vocabulary expansion, and language independence — of the knowledge graphs at cross-language level. The comparison with the state-of-the-art models, in the Spanish-English and German-English settings, shows that this type of representation captures the meaning in a more accurate way, and therefore improves the performance, even in cases where paraphrasing occurred.

- **Chapter 4** A knowledge-based representation for cross-language document retrieval and categorization

  In this chapter we present our work published in the conference of the *European Chapter of the Association for Computational Linguistics (EACL)*. This publication presents a modified version of our cross-language knowledge graph analysis model. This also includes a vector component to cover shortcomings such as out-of-vocabulary words and verbal tenses. Experimental results between several pairs of languages, in the tasks of cross-language document retrieval and categorization, show its potential for these tasks, and cross-language similarity tasks in general.

- **Chapter 5** Discussion of the results

  In this chapter we discuss the results that have been previously obtained. Moreover, we complement our study with some further experiment in order to complete the picture at task level, and analyse the obtained results from a cross-domain and cross-language perspective. In addition, we present our experiments and results with knowledge graphs in other NLP tasks such as community questions answering, native language identification, and language variety identification.

- **Chapter 6** Conclusions

  In this chapter we draw the main conclusions of this thesis and answer the research questions made in the introduction. In addition, we detail our scientific contributions disseminated in the form of publications. Finally, we comment the open research lines for possible future works.

2 *Cross-domain Polarity Classification using a Knowledge-enhanced Meta-classifier*

Published in:

- **Franco-Salvador, M.**, Cruz, F. L., Troyano, J. A., and Rosso, P. (2015b). Cross-domain polarity classification using a knowledge-enhanced meta-classifier. *Knowledge-Based Systems*, 86:46–56. **(Impact Factor: 2.92)**

This chapter of the thesis studies the contribution of knowledge graph-based features for single- and cross-domain polarity classification. For this purpose, we measure the performance and amount of information of these features, and compare them to traditional ones such as bag of words and word n-grams. In addition, we employ meta-learning to put all the evaluated features together and to compare the resulting model with the state of the art.

Abstract

Current approaches to single and cross-domain polarity classification usually use bag of words, n-grams or lexical resource-based classifiers. In this paper, we propose the use of meta-learning to combine and enrich those approaches by adding also other knowledge-based features. In addition to the aforementioned classical approaches, our system uses the BabelNet multilingual semantic network to generate features derived from word sense disambiguation and vocabulary expansion. Experimental results show state-of-the-art performance on single and cross-domain polarity classification. Contrary to other approaches, ours is generic. These results were obtained without any domain adaptation technique. Moreover, the use of meta-learning allows our approach to obtain the most stable results across domains. Finally, our empirical analysis provides interesting insights on the use of semantic network-based features.

Keywords: **—** Sentiment analysis, Cross-domain polarity classification, Meta-learning, Word sense disambiguation, Semantic network

2.1 Introduction

Text classification (also known as text categorization) is the task of assigning a category or categories to a text document from a set of predefined categories. Although at first this topic was approached from a knowledge engineering perspective (manually defining a set of rules encoding expert knowledge), in the 90's machine learning became the main approach, and so it stands today. A good survey on machine learning approaches to text classification can be found in (Sebastiani, 2002).

The nature of the predefined categories in text classification can be very heterogeneous. The most common task is that of topic-based classification, attemping to classify documents according to their subject matter (e.g. Sports vs. Politics vs. Economics). More recently, in the context of the Web 2.0 and social media, it emerged the task of deciding whether a subjective text (typically, a textual review of some product or a cultural or political issue) is positive or negative, depending on the overall sentiment detected. This particular task is known as polarity classification or sentiment classification (Pang et al., 2002; Turney, 2002). Although it can be defined in terms

of text classification (being positive and negative the predefined categories) and tackled with similar approaches, polarity classification has been proved to be a more difficult task (Pang et al., 2002): while topics are often identifiable by keywords alone, sentiment can be expressed in a more subtle manner, and even more when for instance irony is employed (Reyes and Rosso, 2013). Therefore, solutions blue based only on bag-of-words representations of documents may not be enough.

In this work we are interested in single and cross-domain polarity classification. Since we are applying machine learning techniques, we start with a training set of documents to build some classifiers. In this context, single-domain classification is the aforementioned common text classification; it refers to training and testing classifiers on the same domain (e.g. movie reviews). Meanwhile, cross-domain classification refers to testing on a different domain (target domain) from that or those used in training (source domains), e.g. training on movie reviews and testing on books reviews. Because manually labeled documents are needed for training, the latter allows to work with domains where no labeled documents are available. The problem of cross-domain text classification was first tackled by Dai et al. (2007), and the first results on cross-domain polarity classification were reported by Blitzer et al. (2007).

In order to combine different approaches from the research literature and recent knowledge-based approaches, and also to measure the contributions of each one, we propose the use of a meta-learning scheme called Stacked Generalization (Wolpert, 1992). The set of base classifiers to be combined using that scheme include solutions used in the past as a TF-IDF bag-of-words classifier, a TF-IDF word n-gram classifier, and a lexical resource for opinion mining-based classifier; but also two new proposals, a word sense disambiguation-based classifier and a vocabulary expansion-based classifier. The latter two classifiers are trained on the basis of knowledge graphs, a subset of a semantic network, i.e., BabelNet (Navigli and Ponzetto, 2012a), focused on the concepts belonging to the text being classified.

The rest of the paper is structured as follows. In Section 2.2 we describe the related work on single and cross-domain polarity classification. In Section 2.3 we introduce our new knowledge-enhanced meta-classifier. In Section 2.4 we evaluate our approach in the tasks of single and cross-domain polarity classification, and compare it with other state-of-the-art approaches.

In that section we evaluate also the performance of our different base classifiers. Finally, in Section 2.5 we draw the conclusions and mention directions for future work.

2.2 Related Work

The first experiments on single-domain polarity classification using machine learning techniques were performed by Pang et al. (2002). They used a movie review dataset extracted from IMDb.[1] They concluded that polarity classification achieves worse results than other text classification tasks when applying the standard machine learning techniques. Another interesting conclusion was that using unigram presence instead of unigram frequency leads to better results, contrary to observations in other works on text classification (McCallum and Nigam, 1998)

Recent works on polarity classification use the Multi-Domain Sentiment Dataset (Blitzer et al., 2007) for evaluation. In its last version, the resource is composed by Amazon product reviews of 25 product types, though most works report results on only the four domains used by Blitzer et al. (2007): Books, Electronics, DVDs and Kitchen appliances. Focused on single-domain polarity classification, Dredze et al. (2008) presented a new online learning method named confidence-weigthed learning. The method is based on measuring the confidence of each parameter of the classifier; less confident parameters are updated more aggressively than more confident ones. They performed experiments on standard datasets related to different text classification tasks, reporting very good results for the Multi-Domain Sentiment Dataset. Another approach, proposed by Li and Zong (2008), use n-grams combined with Binormal Separation (Forman, 2008), an alternative to TF-IDF to select the optimal set of features. They reported interesting results in single domain classification.

Cross-domain polarity classification has gained popularity thanks to the advances in domain adaptation (Ben-David et al., 2010; Blitzer et al., 2008; Daumé III, 2007). These techniques make use of labeled data from a source domain, and unlabeled data from source and target domains to train their classifiers. Using the different domains available in the Multi-Domain Sentiment Dataset, Blitzer et al. (2007) was also the first to report results on cross-

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data/

domain classification proposing two algorithms: structural correspondence learning (SCL), and its variant using mutual informaton (SCL-MI). The SCL model selects pivot (unigram and bigram) features frequently appearing in both source and target domains. Then it learns to predict those pivot features in the unlabeled data from both domains. Later, a singular value descomposition is performed to reduce dimensions, and a binary classifer is trained to determine the polarity. Similarly, interesting results on cross-domain polarity classification have been reported by spectral feature alignment (SFA) (Pan et al., 2010). Using unigram and bigram features, the model exploits the mutual information between each feature and the domain label to differentiate domain-specific and domain-independent features. Next, a bipartite graph is constructed by dividing both types of features. An edge connects features from different types if there exists co-ocurrence. Finally, a spectral clustering is performed to generate feature clusters and a binary classifer is built for the polarity classification. More recently, Bollegala et al. (2011, 2013) used a cross-domain lexicon creation to generate a sentiment-sensitive thesaurus (SST) that groups different words expressing the same sentiment, using also unigram and bigram features as representation. This approach also obtained competitive results in single-domain polarity classification.

Note that all cross-domain approaches use domain adaptation techniques extracting relevant features from the source domains, in order to obtain important features to classify the target domain. In contrast, we do not use unlabeled data from the target domain. Our approach is focused on proposing new knowledge-based features which allows for training models using the source domains that are able to be directly applied to the target domain. In Section 2.4.4 we compare our approach in the task of single-domain polarity classification against SST and the state-of-the-art approaches proposed by Dredze et al. (2008) and Li and Zong (2008). Next, in Section 2.4.5 we compare our approach in the task of cross-domain polarity classification against SCL-MI, SFA and SST models.

2.3 Knowledge-enhanced Meta-classifier

We propose the use of a meta-learning scheme for combining different classical approaches, i.e., bag of words, n-grams or lexical resource-based classifiers. Key to our approach is adding also other knowledge-based classifiers. By using a semantic network, we perform word sense disambiguation and

generate new independent classifiers for the main part-of-speech tags: disambiguated adjectives, nouns, verbs and adverbs. Using the disambiguated terms, the semantic network allows us to obtain a vocabulary expansion-based classifier. In Section 2.3.1 we present the semantic network, and the word sense disambiguation and vocabulary expansion methods. Then, in Section 2.3.2 we describe the base classifiers that compose our system. Finally, in Section 2.3.3 we define the Stacked Generalization that we use to combine those classifiers.

2.3.1 Word Sense Disambiguation and Vocabulary Expansion via a Semantic Network

A semantic network (Sowa, 2006) is a (un)directed graph consisting of vertices, which represent concepts, and edges, which represent semantic relations between them. Concepts are usually organized into a taxonomic hierarchy. Figure 2.1 shows a simple example of semantic network.

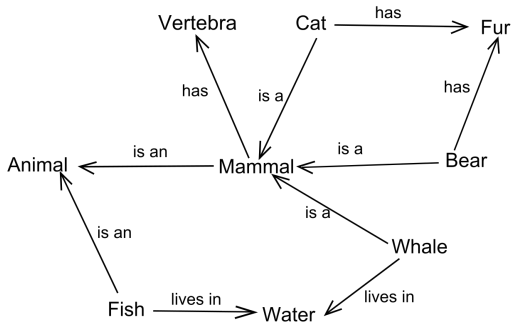

Figure 2.1. Semantic network example focused on the animal world.

In this work we use the semantic network graph to: (i) perform word sense disambiguation, and (ii) perform a vocabulary expansion using the disambiguated words. Despite having the WordNet Semantic Network (Fellbaum, 1998), which is an historical resource including 117,000 synsets[2] in English, in this work we are interested in employing a larger size wide-coverage lexical knowledge resource. Among those, we can find knowledge bases extracted automatically from Wikipedia such as DBPedia (Bizer et al., 2009) or YAGO (Hoffart et al., 2013). However, due to its WordNet-based

[2]Set of word synonyms.

internal structure combined with Wikipedia, the high amount of synsets included, and the lexicalizations of its concepts available in multiple languages,[3] we chose the BabelNet Multilingual Semantic Network.

2.3.1.1 BabelNet

BabelNet[4] 2.5 (Navigli and Ponzetto, 2012a) is a multilingual semantic network whose concepts and relations are obtained from the automatic mapping onto Wordnet of Wikipedia,[5] OmegaWiki,[6] Wiktionary,[7] Wikidata,[8] and Open Multilingual WordNet.[9] BabelNet is therefore a multilingual "encyclopedic dictionary" that combines lexicographic information with wide-coverage encyclopedic knowledge. Concepts in BabelNet are represented similarly to WordNet, i.e., by grouping sets of synonyms in the different languages into multilingual synsets. Multilingual synsets contain lexicalizations from WordNet and Open Multilingual WordNet synsets, the corresponding Wikipedia pages, the OmegaWiki, Wiktionary and Wikidata entries, and additional translations by a statistical machine translation system. The relations between synsets are collected from WordNet, Open Multilingual WordNet, and from Wikipedia's hyperlinks between pages. The current version of BabelNet includes 9,348,287 synsets, covers 50 languages, and has a WordNet-Wikipedia mapping correctness of 91% (Navigli et al., 2013).

2.3.1.2 Word Sense Disambiguation

Word sense disambiguation (WSD) (Navigli, 2009) is the process of identifying which sense (i.e., meaning) of a word is used in a sentence, when the word is polysemic. In general, the approaches for WSD can be classified into three types: (i) supervised, with a considerable effort for new languages and domains due to the huge amount of annotated data required (Pilehvar and Navigli, 2014; Shen et al., 2013); (ii) unsupervised approaches, which have to deal with data sparsity and an intrinsic difficulty with their evalua-

[3] While this work is exclusively evaluated on English, this multilinguality allows us to perform at multilingual level.

[4] http://babelnet.org

[5] http://wikipedia.org

[6] http://omegawiki.org

[7] http://wiktionary.org

[8] http://wikidata.org

[9] http://compling.hss.ntu.edu.sg/omw/

tion (Agirre et al., 2006; Di Marco and Navigli, 2013); (iii) knowledge-based approaches, which exploit the knowledge available in structured knowledge bases (Agirre et al., 2014; Moro et al., 2014; Navigli and Lapata, 2010; Ponzetto and Navigli, 2010). Vocabulary expansion benefits from the WSD performed using a knowledge base by exploiting the relations in its network.

BabelNet has been used for WSD in several works, including some of the aforementioned publications and also as part of the Multilingual Word Sense Disambiguation Task of the SemEval Workshop (Navigli et al., 2013). Similarly to Navigli and Ponzetto (2012a) and Franco-Salvador et al. (2013a, 2014b), we followed Navigli and Lapata (2010) to create knowledge graphs[10] in order to perform the WSD and the vocabulary expansion. The five-step method we used to perform the WSD is the following:

(i) Part-of-speech tagging and lemmatization Initially we process a document d with tokenization, multi-word extraction, part-of-speech (POS) tagging and lemmatization[11] to obtain the list of tuples (lemma,tag) T. We are interested only in the POS tags available on BabelNet (adjectives, nouns, verbs and adverbs).

(ii) Populating the graph with initial concepts Next, we create an initially-empty knowledge graph $G = (V, E)$, i.e., such that $V = E = \emptyset$. We populate the vertex set V with the set S_K of all the synsets in BabelNet which contain any tuple (lemma,tag) in T in the document language L, that is:

$$S_K = \bigcup_{t \in T} \text{Synsets}_L(t), \tag{2.1}$$

where $\text{Synsets}_L(t)$ is the set of synsets which contains a tuple (lemma,tag) t in the language of interest L.

(iii) Creating the knowledge graph We create the knowledge graph by searching on BabelNet to obtain the set of paths P connecting pairs of synsets

[10] A knowledge graph is a subset of the original semantic network focused on the concepts belonging to a text, and in the intermediate concepts and relations between them.

[11] For this purpose we used the Stanford Log-linear Part-Of-Speech Tagger: `http://nlp.stanford.edu/software/tagger.shtml`. For the multi-word extraction we implemented our own tool based on the matching of typical patterns.

in V. Formally, for each pair $\{v, v'\} \in V$ such that v and v' do not share any lexicalization[12] in T, for each path in BabelNet $v \rightarrow v_1 \rightarrow \cdots \rightarrow v_n \rightarrow v'$, we set: $V := V \cup \{v_1, \ldots, v_n\}$ and $E := E \cup \{(v, v_1), \ldots, (v_n, v')\}$. That is, we add all the path vertices and edges to G. Following Navigli and Ponzetto (2012a), the path length is limited to maximum length of 3, in order to avoid an excessive semantic drift.

As a result of populating the graph with intermediate edges and vertices, we obtain a knowledge graph which models the semantic context of document d.

(iv) Knowledge graph weighting The next step consists of weighting all the concepts and semantic relations of the knowledge graph G. For weighting relations we use the original weights from BabelNet, which provide the degree of relatedness between the synset end points of each edge.[13] For weighting concepts different methods, including the PageRank (Page et al., 1998) algorithm, have been tested in the past. In this work, we score each concept using its own outdegree, which has proved to obtain the best results.(Navigli and Ponzetto, 2012a)

(v) Selecting the corresponding disambiguations Finally, for each tuple (lemma,tag) $t \in T$, we collect from BabelNet the set of synsets S_t containing t, and we select as proper disambiguation t_{WSD} the synset with the highest score:

$$t_{WSD} = \arg\max_{s \in S_t} \text{score}(s), \tag{2.2}$$

2.3.1.3 Vocabulary Expansion

Once we have disambiguated the words of a document d, to enrich and increase the available context, we perform an automatic vocabulary expansion (Ehrlich and Rapaport, 1997; Ehrlich, 1995) using the BabelNet graph topology. A simple vocabulary expansion can be done using directly any connected concept to a disambiguated one, up to a certain distance in the graph.

[12] This prevents different senses of the same term from being connected via a path in the resulting knowledge graph.

[13] At this point, we removed the edges below a certain threshold that represents a low semantic relationship.

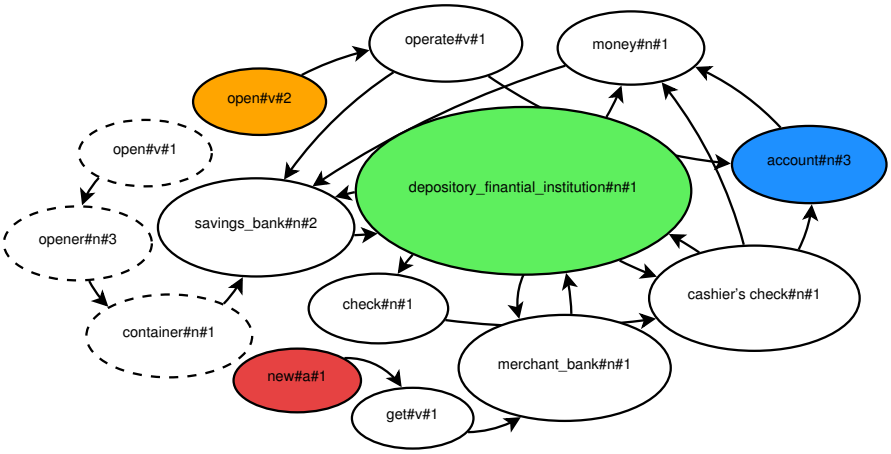

Figure 2.2. Simplified knowledge graph created from the sentence "I opened a new bank account". Colored nodes are the resulting disambiguations while white nodes are expanded concepts. Dashed noded will not be included in the vocabulary expansion set.

However, to preserve as much context as possible and to avoid introducing noise, we include only intermediate concepts between pairs of disambiguated words. Formally, using the knowledge graph G created in Section 2.3.1.2, we obtain a vocabulary expansion as follows:

(i) Collecting the disambiguation senses We first use the process described in the previous section to obtain the set S_{WSD}. This set is composed by the disambiguation synsets of the original words of document d.

(ii) Removing alternative senses We create a path set P' by removing from the path set P all the paths between synsets which are not in S_{WSD}. This step removes noise by creating a knowledge graph focused on the disambiguated concepts.

(iii) Obtaining the expanded concepts We obtain the vocabulary expansion by creating a set S_{exp} including the intermediate concepts in the paths of P'. We remove the source and target concepts from paths to evaluate the

performance of the vocabulary expansion without the original words (see Section 2.4.3).[14]

Figure 2.2 provides an example[15] of disambiguation and vocabulary expansion using knowledge graphs.

2.3.2 BASE CLASSIFIERS

We can now define the base classifiers that compose our system. We first include a TF-IDF bag-of-words classifier, a TF-IDF word n-gram classifier and a lexical resource for the opinion mining-based classifier. The choice of these components has been motivated by the good results that they achieved in the past. In addition, in this work we want to investigate the impact of knowledge-based classifiers; therefore we include an independent classifier to study the contribution of WSD for each POS tag employed (adjectives, nouns, verbs and adverbs). Finally, under the assumption that semantically-related concepts have a common near relative, we want to exploit this possible relateness between concepts including a vocabulary expansion-based classifier. Next we explain in more detail our eight base classifiers:

(i) Bag-of-words classifier This approach transforms a document d into a traditional vector representation. Following the literature, we selected the most widely used representation for real-valued feature vectors, commonly used as baseline: the Term Frequency-Inverse Document Frequency (TF-IDF) weighting (Salton et al., 1983; Salton and McGill, 1986).

$$\text{tf-idf}(w) = \text{tf}(w)N/n(w). \tag{2.3}$$

where $tf(w)$ is the number of times a term w occurs in document d, N is the total number of documents in the collection and $n(w)$ is the number of documents that contain w. We removed stopwords from documents for all the base classifiers.

[14] This last part is optional, although it helps to focus on the vocabulary expanded concepts.

[15] Weights and nodes representing alternative senses or intermediate concepts are removed for simplicity.

As classifier, we selected Support Vector Machines (SVM) (Chang and Lin, 2011), with a linear kernel function,[16] given its good performance for text classification (Joachims, 1998) using TF-IDF weighting.

(ii) Word n-gram classifier The use of word n-grams has been proposed several times (Cavnar, 1995; Li and Zong, 2008; Mayfield and McNamee, 1999) as a better alternative to single word vector representation due to the additional information that it provides. Using n-grams is a plus for a complex classification task like polarity classification: while topics are often identifiable by keywords alone, sentiment can be expressed in a more subtle manner (Pang et al., 2002). For example, the keyword *like* may be correlated with positive sentences (e.g. "I *like* this paper a lot.") or with negative sentences (e.g. "I do *not like* this paper at all."). Using n-grams also allows us to learn frequent, opinion-bearing multiword expressions (e.g. "*you will love* (this story)").

This n-gram representation is processed with a TF-IDF weighting and an SVM classifier. Since larger n-grams will not be frequent, we included only a combination (Li and Zong, 2008) of 1, 2, and 3-grams.

(iii) Lexical resource-based classifier The use of lexical resources for opinion mining was strongly popularized by the release of SentiWordNet (Baccianella et al., 2010; Esuli and Sebastiani, 2006). This resource assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. It has been sucessfully applied to polarity classification in the past (Hamouda and Rohaim, 2011; Ohana and Tierney, 2009).

We selected as lexical resource ML-SentiCon (Cruz et al., 2014), which proved to make several improvements with respect to the original SentiWordnet 3.0, with a significative better positivity, negativity and objectivity estimation, reflecting those results on their evaluation.

For this base classifier, we decided to use the tree-based C4.5 (Quinlan, 1996) model, which infers a hierarchy of rules as a function of different

[16] We use the linear kernel function for all the SVM base classifiers.

[17] As we can see, we take advantage of WSD to remove noise (unrelared synsets).

[18] We refer to the disambiguations of the original words of the document.

[19] Since the format of ML-SentiCon is the same as SentiWordNet, and BabelNet has a synset for each WordNet synset, we can map directly our disambiguated words to that lexical resource.

Model features
Number of words in document d.
Number of disambiguated synsets[17] in the knowledge graph G (see Section 2.3.1.2).
Number of directly connected disambiguated synsets in G[18].
Number of adjetives in d.
Number of nouns in d.
Number of verbs in d.
Number of adverbs in d.
Average positivity of the disambiguated words of d[19].
Average negativity of the disambiguated words of d.
Average objetivity of the disambiguated words of d

Table 2.1. List of features selected for the lexical resource-based classifier.

feature values to determine the final class, and provides good performance for polarity classification (Jia et al., 2009). Its use is also motivated by the different types of features that we selected for this classifier (see Table 2.1): some of them are discrete and unbounded. In addition, considering that there are only 10 features, using SVM did not pose any additional advantage with regard to a simpler C4.5 tree-based classifier.

(iv-vii) Word sense disambiguation-based classifiers As we stated at the beginning of this section, to study the impact of WSD on polarity classification, we generate an independent classifier for each POS tag available on BabelNet (adjectives, nouns, verbs and adverbs) on the basis of the method explained in Section 2.3.1.2.

During the prototyping process, we realized that due to the use of independent classifiers for each POS tag, and the error introduced by wrong disambiguations, the TF-IDF weighting provided an imprecise representation of documents, and worse results than using only binary TF (presence or not of the word w in the document). Since the use of this technique has been studied in the past with good results (Pang et al., 2002), for the WSD-based models we decided to use binary TF as weighting and SVM as classifier.

(viii) Vocabulary expansion-based classifier The last base classifier uses the vocabulary expansion explained in Section 2.3.1.3 to represent each document d as a binary TF of synsets, which are related to the original disambiguated ones of d. The classification is performed using SVM. Since we

Base classifier ID	Description	Weighting	Classifier	Avg. # feat.
BOW	Bag-of-words representation	TF-IDF	SVM	19,976
(1+2+3)-grams	Combine $\{1, 2, 3\}$-grams to represent documents	TF-IDF	SVM	58,636
ML-SentiCon	Use a lexical resource to extract different polarity-related features	-	C4.5	10
Noun WSD	Represent documents by its set of disambiguated nouns	Binary TF	SVM	13,139
Adjective WSD	Represent documents by its set of disambiguated adjectives	Binary TF	SVM	3,241
Verb WSD	Represent documents by its set of disambiguated verbs	Binary TF	SVM	2,138
Adverb WSD	Represent documents by its set of disambiguated adverbs	Binary TF	SVM	689
Vocab. Exp.	Use a vocabulary expansion to represent the documents	Binary TF	SVM	59,372

Table 2.2. Summary of base classifiers.

are removing the original concepts of the documents from the vocabulary expansion, a document containing the concepts "Michael Jordan" and "NBA" will be represented by concepts as "Basketball" and "Sport", but not by the original concepts. As previously stated, the original concept removal was performed because we are interested in evaluating the performance of the vocabulary expansion without the original words.

Table 2.2 provides a summary[20] of all the base classifiers.

2.3.3 STACKED GENERALIZATION

We combine the base classifiers with one of the most popular combination methods in meta-learning: stacking. It has been used successfully in Natural Language processing (NLP) tasks (Enríquez et al., 2013; Van Halteren et al., 1998) in the past. This method follows the original Stacked Generalization method (Wolpert, 1992) to project documents onto a new dimensional space, which is composed by the annotations of a first-level base classifiers set. This combination is able to exploit additional information from a corpus by processing it with different classifiers. A second-level classifier uses all of the annotations of the first level to obtain a final decision, with the advantage of recognizing and classifying correctly patterns in which the correct class tag is in inferiority. In this work, instead of representing the results of the first level as a vector of class tags, we represent them as a vector of class probabilities, which proved to obtain better results using SVM (Martín-Valdivia et al., 2013).

[20]Column "Avg. # feat." shows the average number of potential features of the classifier across domains before applying their respective thresholds (see Section 2.4.2).

Algorithm 2.1 Stacking Generalization algorithm.

Input: a tagged training corpus T and a untagged test corpus t.
Output: a tagged test corpus t''.

1: Split T into K parts to obtain $T_{1,\ldots,K}$ partitions.
2: Tag $T_{1,\ldots,K}$ using cross-validation with the $C_{1,\ldots,N}$ base classifiers to obtain $T'_{1,\ldots,K}$ partitions containing the transformed samples of T.
3: Using $T_{1,\ldots,K}$ for training, classify t with $C_{1,\ldots,N}$ to obtain the transformed corpus t'.
4: Use $T'_{1,\ldots,K}$ as a single partition to train the second-level classifier $C_{comb.}$.
5: Classify t' with $C_{comb.}$ to obtain the tagged test corpus t''.

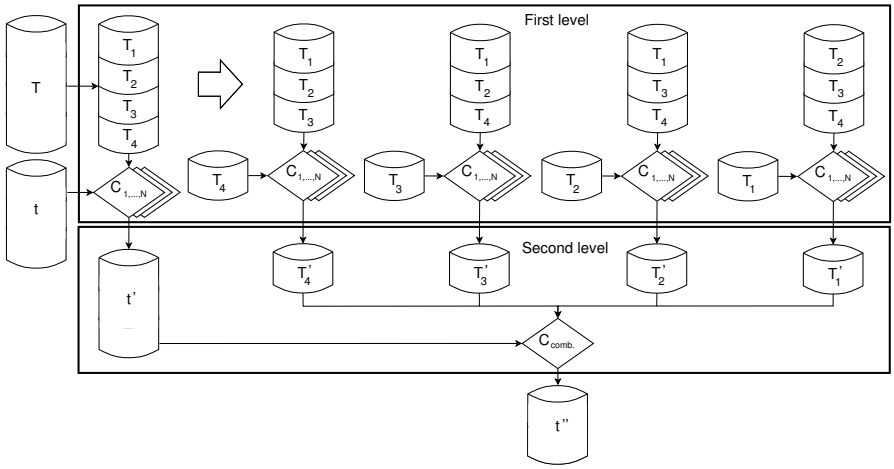

Figure 2.3. Stacked Generalization scheme. Training and test partitions are projected into a new dimensional space which is composed by the first-level classifier class probabilities. The second-level classifier uses those probabilities to obtain the final decision.

We can see the Stacked Generalization method detailed in Algorithm 2.1. Lines 1–3 correspond to the first level of the classifier, which makes the transformation of the training corpus. The second level of the classfier is explained in Lines 4–5, which obtains the final classification of the test corpus. A complete scheme of the model is shown in Figure 2.3.

2.4 Evaluation

In this section we evaluate the base classifiers of our Knowledge-enhanced Meta-classifier (KE-Meta), and compare our approach with state-of-the-art models on single and cross-domain polarity classification.

2.4.1 DATASET

To evaluate our system we chose a classical state-of-the-art dataset, the Multi-Domain Sentiment Dataset (version 2.0)[21] (Blitzer et al., 2007), which has been used for the evaluation of several research works on sentiment analysis (Blitzer et al., 2007; Bollegala et al., 2013; Dredze et al., 2008; Li and Zong, 2008). The dataset is composed by Amazon product reviews of 25 product types. Each review contains metadata including a rating of 0-5 stars, the reviewer name and location, the product name, the review date and title, and the review text. In addition, for research purposes, a subset of the reviews with rating < 3 were originally labeled as negative, and with rating > 3 as positive. Following the literature, in this work we use the Books, Electronics, DVDs, and Kitchen appliances reviews, with 1,000 positive and 1,000 negative documents per domain, having a total of 8,000 reviews. With this setup, we can compare our results on single and cross-domain polarity classification directly with the state of the art.

2.4.2 METHODOLOGY

The evaluation of our approach in single-domain polarity classification is performed using a stratified 10-fold cross-validation setup for each domain. In cross-domain, we followed the same 10-fold cross-validation setup,[22] in this case, training always with all domains available and excluding the target domain to classify, e.g. we train with Books, Electronics, and DVDs, and we classify Kitchen reviews. We selected as the evaluation metric the accuracy of the classifiers, which is the proportion of correctly classified reviews among the test dataset. We detail the models compared with our approach on its respective evaluation sections. Note that the number of dimensions of all our base classifiers is limited to a maximum number of 20,000. However,

[21] `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/`

[22] The cross-validation here is used only to train our KE-Meta classifier, which needs a splitting of the data to obtain training and testing partitions to generate the final second-level classifier.

similar results were obtained with sizes ranging between 15,000 and 25,000 during the prototyping step.

2.4.3 EVALUATION OF BASE CLASSIFIERS

To evaluate the eight base classifiers that compose our approach (cfr Section 2.3.2) summarized in Table 2.2, we first employ a traditional measure of information theory (Hall and Smith, 1998): the information gain ratio (IGR) (Quinlan, 1986; Raileanu and Stoffel, 2004). Once analyzed the IGR, we will continue with the study of the accuracy of classification of each base classifier.

Having a training set T and its set of attributes $Attr$, the IGR measure provides a normalized estimation (between 0 and 1) of the amount of information that an attribute $a \in Attr$ provides to determine the class attribute.[23] The IGR of an attribute a is calculated as the ratio between the information gain (IG) and the intrinsic value (IV):

$$\mathrm{IGR}(T,a) = \frac{\mathrm{IG}(T,a)}{\mathrm{IV}(T,a)} \tag{2.4}$$

$$\mathrm{IG}(T,a) = \mathrm{H}(T) - \sum_{v \in \mathrm{values}(a)} \left(\frac{|\{x \in T | \, \mathrm{value}(x,a) = v\}|}{|T|} \cdot H(\{x \in T | \, \mathrm{value}(x,a) = v\}) \right) \tag{2.5}$$

where we substract to the total entropy H of the train set T the sum of the relative entropies of the different values of a in T. For each of the attributes, if a unique classification can be made for the result attribute, the information gain is equal to the total entropy of a. The IV is a normalization factor estimated as a function of the substracted entropies of $\mathrm{H}(T)$ in IG.

$$\mathrm{IV}(T,a) = -\sum_{v \in \mathrm{values}(a)} \frac{|\{x \in T | \, \mathrm{value}(x,a) = v\}|}{|T|} \cdot \log_2 \left(\frac{|\{x \in T | \, \mathrm{value}(x,a) = v\}|}{|T|} \right) \tag{2.6}$$

[23] Note that each attribute $a \in Attr$ corresponds to a base classifier in our approach.

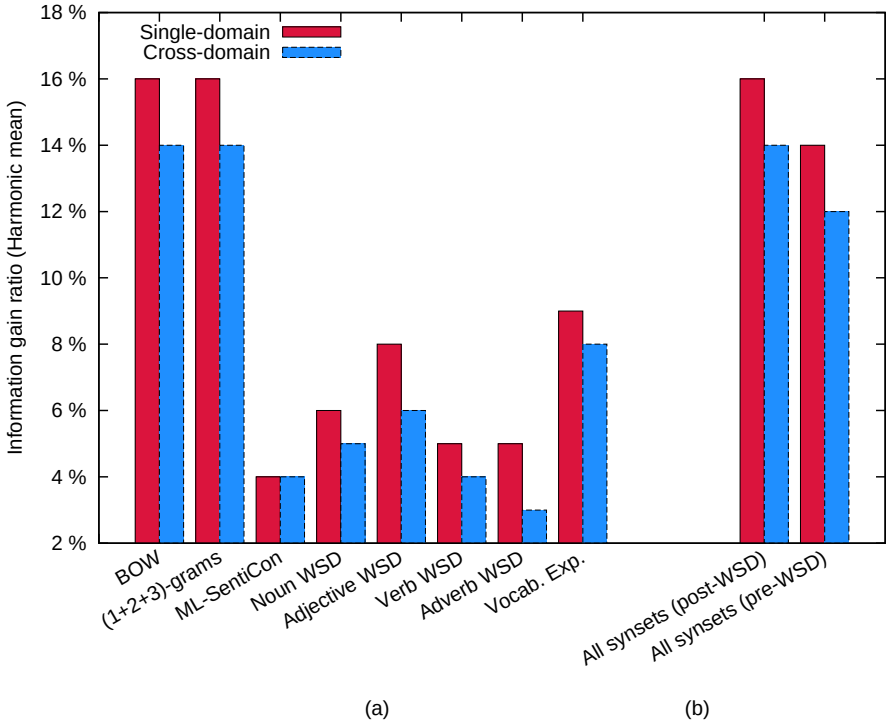

Figure 2.4. Information gain ratio of the eight base classifiers in single and cross-domain polarity classification. We show the harmonic mean of the IGR of each feature among the different tested domains. (a) Base classifiers; (b) other classifiers.

To obtain the IGR of our base classifiers, we estimated the IGR on each tested domain and we calculated the harmonic mean[24] of those results. This test was performed on single and cross-domain polarity classification. We show the results in Figure 2.4. As expected, the IGR in cross-domain is lower than working on single domain for almost all of the base classifiers. This is not the case of the model using ML-SentiCon, which, despite getting a low IGR, is able to preserve all its gain when performing at cross-domain level. These results put forward the advantage of knowledge bases to model the information in a domain-independent way. We can see that BOW and (1+2+3)-grams classifiers obtained the highest information gain ratios, with almost identical values. The results prove that these models are a good choice

[24]The harmonic mean is the most adequate measure to average percentages of different domains.

as base classifiers to be complemented with other classifiers. The vocabulary expansion, which does not include the original words of the documents, is able to obtain comparable results. Models disambiguating different POS tags obtained considerably low IGR. Adjective WSD was the most informative classifier. This is unsurprising if we consider that often, the polarity of a text could be given by adjectives. This is followed by the classifier for nouns, verbs, and finally adverbs. These last two with identical results on single-domain. Since WSD has been divided into four models, it is difficult to evaluate its contribution. For this reason, we included also the results of two additional classifiers: All synsets (Post-WSD) and All synsets (Pre-WSD). They represent the IGR of a binary TF[25] classifier trained using SVM with: (i) all the disambiguated words together (All synsets (Post-WSD) classifier), and (ii) all the possible senses of the words together before disambiguation (All synsets (Pre-WSD) classifer). As we can see, the performance of All synsets (Post-WSD) significantly outperforms the Pre-WSD model, and obtained similar result to BOW and *n*-grams based approaches. This highlights the capability of WSD to remove noisy senses, leaving only the appropiate one.

Base classifiers	**Books**	**Electronics**	**DVDs**	**Kitchen**
BOW	0.788	0.803	0.804	0.821
(1+2+3)-grams	0.805	0.817	0.803	0.819
ML-SentiCon	0.612	0.644	0.644	0.651
Noun WSD	0.684	0.655	0.679	0.677
Adjective WSD	0.683	0.695	0.729	0.712
Verb WSD	0.669	0.670	0.633	0.675
Adverb WSD	0.651	0.638	0626	0.649
Vocab. Exp.	0.718	0.700	0.709	0.704
Other classifiers				
All synsets (Post-WSD)	0.775	0.782	0.785	0.806
All synsets (Pre-WSD)	0.758	0.765	0.784	0.800

Table 2.3. Base classifiers accuracy per domain in single-domain polarity classification.

[25] Similarly to the other WSD-based classifiers, binary TF is preferred to TF-IDF to smooth the error in case of a wrong disambiguation.

Base classifiers	**Books**	**Electronics**	**DVDs**	**Kitchen**
BOW	0.756	0.804	0.791	0.809
(1+2+3)-grams	0.744	0.798	0.771	0.769
ML-SentiCon	0.643	0.652	0.639	0.673
Noun WSD	0.626	0.625	0.644	0.649
Adjective WSD	0.665	0.687	0.699	0.686
Verb WSD	0.584	0.619	0.590	0.605
Adverb WSD	0.617	0.661	0.622	0.646
Vocab. Exp.	0.666	0.695	0.694	0.695
Other classifiers				
All synsets (Post-WSD)	0.745	0.765	0.776	0.775
All synsets (Pre-WSD)	0.726	0.757	0.765	0.769

Table 2.4. Base classifiers accuracy per domain in cross-domain polarity classification.

Once evaluated the IGR of the base classifiers, the next step is to evaluate them separately in the polarity classification task. Following the setup of Section 2.4.2, we can see the results on single-domain in Table 2.3. The results are in line with those obtained for IGR: (1+2+3)-grams obtained the highest results, followed by BOW. The vocabulary expansion achieved averaged results followed by Adjective WSD and the rest of WSD-based classifiers. Finally, ML-SentiCon was the model with the lowest accuracy. Looking at the results on cross-domain in Table 2.4, we can see a similar trend. Despite there is a general decrease in the results, as we stated while analyzing its IGR, ML-SentiCon has even improved its results on cross-domain, taking advantage of all the other domains to train a domain independent model which is able to outperform the noun, verb and adverb WSD-based approaches. Note that, as we can see in both tables, All synsets (Post-WSD) classifier outperforms the Pre-WSD model, and gets similar results to the best base classifiers.

Looking at all the previous results, due to the different type of classifers selected, each one of them should provide different information when combined in a meta-classifier. The next experiment studies the improvement in the accuracy when adding base classifiers one by one to our KE-Meta ap-

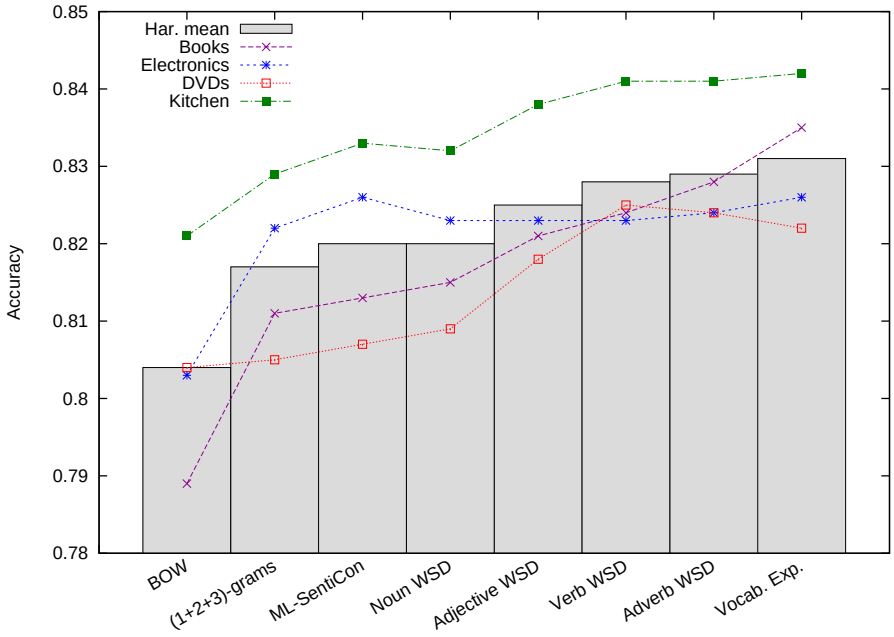

Figure 2.5. KE-Meta classifier improvement across domains when incrementally adding new base classifiers to single-domain polarity classification. Each column represents the accuracy of the model when we combine that base classifier with the classifiers at its left.

proach. We can see the single-domain results in Figure 2.5. As expected, considering the harmonic mean, there is an improvement when each new base classifier is added. As one classifier might provide information included by others, the improvements were shown to be greater at the beginning. The results on cross-domain are shown in Figure 2.6. Also in this case there is a clear improvement compared to the first base classifier included, being BOW, ML-SentiCon and Adjective WSD, the models with higher contribution. However, the vocabulary expansion seems to have a negative contribution in this cross-domain combination. We assume that expanding vocabulary from different domains and combining all the documents together, contributes to obtaining a noisy base classifier with several clusters of vocabulary of concepts related to each training domain. In the next cross-domain experiments we will show also the results without the vocabulary expansion base classifier.

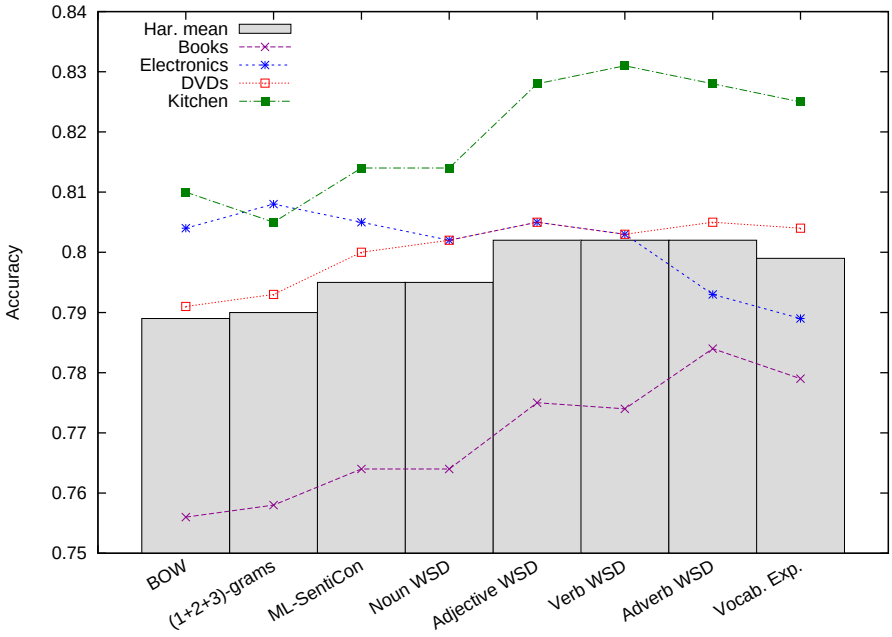

Figure 2.6. KE-Meta classifier improvement across domains when incrementally adding new base classifiers to cross-domain polarity classification. Each column represents the accuracy of the model when we combine that base classifier with the classifiers at its left.

2.4.4 SINGLE-DOMAIN POLARITY CLASSIFICATION

We compared our knowledge-enhanced meta classifier against the state-of-the-art SST model, and those proposed by Dredze et al. (2008) and Li and Zong (2008)[26] (cfr Section 2.2). In addition we included the results of our BOW and (1+2+3)-grams classifiers as baselines.

As we can see from Table 2.5,[27] thanks to the additional information included when combining groups of words as single feature, (1+2+3)-grams obtained better results than BOW. However, all of the compared models outperformed these baselines. Dredze et al.'s approach obtained interesting results, specially classifying electronics. This model benefited from confidence-weighted classification to create very precise linear frontiers among

[26]Results of compared approaches are taken from their original works: Bollegala et al. (2013), Dredze et al. (2008) and Li and Zong (2008).

[27]In this work, statistically significant results according to a χ^2 test are highlighted in bold.

	Method	Books	Electronics	DVDs	Kitchen
(a)	(Dredze et al., 2008)	**0.826**	**0.859**	0.809	0.857
	SST	0.804	0.844	0.824	**0.877**
	(Li and Zong, 2008)	0.790	**0.850**	**0.845**	0.845
(b)	(1+2+3)-grams	0.805	0.817	0.803	0.819
	BOW	0.788	0.803	0.804	0.821
(c)	KE-Meta	**0.835**	0.826	0.823	0.842

Table 2.5. Accuracy results in single-domain polarity classification. (a) State-of-the-art approaches; (b) baselines; (c) proposed approach.

classes. The SST model, using its sentiment sensitive thesaurus, took advantage of the type of reviews used in kitchen domain and obtained the best results, with good accuracy in the other domains. Li and Zong's approach, based on a optimized n-gram selection criteria, obtained the best results on DVD reviews. Our approach obtained the best results on Books domain and considerably high results on the rest. We hypothesize that when reviewers analyze books summarizing parts from the story of the book, our meta-classifier is able to distinguish this pattern by contrasting the probabilities of the base classifiers, and the polarity of the book summary has less influence in the final review classification. Note that our approach is the most stable, with no less than 82.3% of accuracy in all the tests. Using meta-classification, KE-Meta is able to determine which base classifier is better on each domain, maximizing its contribution in the combination. We highlight also that each state-of-the-art approach obtained specially low (or high) results in some domain. This may be produced by the writing style employed by reviewers when commenting on those products. At the end of Section 2.4.5, in Table 2.7 we analyze the vocabulary of domains to investigate these differences further.

2.4.5 CROSS-DOMAIN POLARITY CLASSIFICATION

In this task we compared our KE-Meta approach against the state-of-the-art SFA, SCL-MI and SST approaches.[28] As we mentioned in Section 2.4.3, we included also the results of our approach without the vocabulary expansion-

[28]The results of the approaches compared are taken from Bollegala et al. (2013)

	Method	Books	Electronics	DVDs	Kitchen
(a)	SST	0.763	**0.839**	0.783	**0.852**
	SFA	**0.777**	0.753	0.763	0.815
	SCL-MI	0.746	0.789	0.763	0.820
(b)	BOW	0.756	0.804	0.791	0.809
	(1+2+3)-grams	0.744	0.798	0.771	0.769
(c)	KE-Meta$_B$	**0.784**	0.793	**0.805**	0.828
	KE-Meta	**0.779**	0.789	**0.804**	0.825

Table 2.6. Accuracy results in cross-domain polarity classification. (a) State-of-the-art approaches; (b) baselines; (c) proposed approaches.

based base classifier: KE-Meta$_B$. The BOW and (1+2+3)-grams models are included as baselines.

Table 2.6 shows the cross-domain polarity classification accuracy. The (1+2+3)-grams baseline achieved the lowest results. Training a cross-domain n-gram-based classifier using only three domains does not seem to be sufficient to obtain a good domain-independent n-gram inventory. Evidence of this observation are the close results obtained by SCL-MI and SFA, other two n-gram based approaches. SCL-MI excelled especially in the kitchen domain. We hypothesize that the singular value descomposition used to reduce dimensions worked better with the reduced size of the vocabulary in kitchen domain. The second domain with less vocabulary, electronics, excelled too. The bipartite graph constructed to differentiate domain-specific and independent n-grams helped SFA to obtain significant results on books domain. Precisely despite obtaining the lowest results in that domain, the BOW baseline outperformed SFA and SCL-MI on average. In contrast to n-gram-based approaches, the training data provided was sufficient to infer a vocabulary, which made this classifier more stable. The SST model proved to be a good option in cross-domain, with significative results on electronics and kitchen reviews. Bollegala et al. (2013) justified the low results on books because of the low number of unlabeled data available on that domain, which is necessary to create its sentiment sentitive thesaurus. Finally, our KE-Meta approach obtained the best results on books and DVD reviews, being again the most stable approach across domains, thanks to the combination of different base classifiers. KE-Meta$_B$, the classifier that does not consider the vocabulary expansion, obtained not significative better results in all domains.

Since the use of this base classifer improved the results in single-domain, future work is needed in order to understand how to improve its performance also in cross-domain.

Statistics	Books	Electronics	DVDs	Kitchen
Average document length	175	113	190	96
# different lemmas domain	26,108	13,947	28,757	11,095
Average # different lemmas per document	53.4	33.6	57.8	28.3
% nouns domain	66.5%	64.7%	67.4%	61.3%
% adjectives domain	16.7%	15.8%	15.5%	17.4%
% verbs domain	10.0%	11.9%	9.5%	14.1%
% adverbs domain	3.4%	3.7%	3.2%	4%
# different senses domain	17,523	8,809	18,487	8,416
Average # different senses per document	51.2	31.8	54.3	27.9
# different lemmas domain / # different senses domain	0.671	0.632	0.643	0.759
KE-Meta results (single-domain)	**0.835**	0.826	0.823	0.842
KE-Meta results (cross-domain)	**0.784**	0.793	**0.805**	0.828

Table 2.7. Corpus statistics per domain. Bold results indicate statistical significance.

Experimental results of Tables 2.5 and 2.6 show that review polarity classification of evaluated approaches differ across domains. These differences could be due to the different language employed by reviewers when commenting on products of different domains. In Table 2.7 we can see some statistics of the corpus divided by domain. While kitchen appliance and electronic reviewers evaluate using short comments, reviews of book and DVD domains are longer, e.g. some of them include a summary of the story. Interesting also the reduced percentage of nouns in kitchen compared to the rest. It seems that kitchen appliance reviewers do not cite so often other products, and use more qualifying adjectives. This makes this domain the easiest to classify, probably also explained by its shorter length. In general, single-domain n-gram-based approaches obtained better results with the two domains with shorter reviews. However, the same trend is not clearly appreciated for the BOW classifier.

We include in the table also statistics of the disambiguated senses. Note that the ratio between the number of different lemmas per domain and the different senses per domain is a measure of the polysemy employed[29] by reviewers. As we can see, the results of our KE-Meta approach are better when

[29] A value of 1.0 here highlights 0% of polysemy in the corpus.

the percentage of polysemy is lower and, consequently, less WSD effort is required.

2.5 Conclusions

In this work we introduced a knowledge-enhanced meta-classifier for single and cross-domain polarity classification. The main contributions of this work are: (i) KE-Meta, a new generic approach that combines different types of classifiers to categorize documents according to their polarity; and (ii) the study of the impact of WSD and vocabulary expansion-based features as document representation.

In single and cross-domain polarity classification, KE-Meta has proven to perform at par or better than state-of-the-art when classifying Amazon product reviews. Thanks to the combination of different classifiers, our approach obtained the most stable results across domains, and was able to excel in domains such as books and DVDs, which often combine a review and a summary of the product together. In contrast to the state-of-the-art, our meta-classifier does not perform any domain adaptation, which renders our approach generic. Moreover, the study of the information gain of our base classifiers concluded that WSD and vocabulary expansion-based features provide additional information not included in other BOW or n-gram-based classifiers.

Future work will investigate how it affects the inclusion of new base classifiers in KE-Meta. The use of other state-of-the-art approaches combined with our approach should provide better results. In addition, we will improve the current base classifiers, specially the vocabulary expansion-based one, to perform better both at single and cross-domain level. We will study also the performance of our classifier in other popular datasets like the well-known movie review dataset. Moreover, we will evaluate our polarity classification approach in other languages.[30] Finally, we will investigate how to apply multilingual semantic networks and knowledge graphs in other NLP tasks, from both, monolingual and multilingual perspectives.

[30] As we stated in Section 2.3.1, our approach is multilingual. This is due to the use of BabelNet, which performs WSD, vocabulary expansion, and mapping of SentiWordNet with the disambiguated words.

Acknowledgements

This research has been carried out in the framework of the European Commission WIQ-EI IRSES (no. 269180) and DIANA-APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) projects. This research is partially funded by the national project ACOGEUS (TIN2012-38536-C03-02) and the regional project AORESCU (P11-TIC-7684 MO). We thank Juan M. Cotelo and Luis A. Leiva for their support and comments.

3 A Systematic Study of Knowledge Graph Analysis for Cross-language Plagiarism Detection

Published in:

- **Franco-Salvador, M.**, Rosso, P., and Montes-y-Gómez, M. (2016c). A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management*, 52(4): 550–570. **(Impact Factor: 1.26)**

This chapter of the thesis presents the cross-language knowledge graph analysis model for cross-language similarity. We show its performance in Spanish-English and German-English plagiarism detection and compare it with the state of the art. We also perform this evaluation with paraphrase cases of plagiarism. In addition, we study the most relevant characteristics of the knowledge graphs for this type of similarity tasks.

Abstract

Cross-language plagiarism detection aims to detect plagiarised fragments of text among documents in different languages. In this paper, we perform a systematic examination of Cross-language Knowledge Graph Analysis; an approach that represents text fragments using knowledge graphs as a language independent content model. We analyse the contributions to cross-language plagiarism detection of the different aspects covered by knowledge graphs: word sense disambiguation, vocabulary expansion, and representation by similarities with a collection of concepts. In addition, we study both the relevance of concepts and their relations when detecting plagiarism. Finally, as a key component of the knowledge graph construction, we present a new weighting scheme of relations between concepts based on distributed representations of concepts. Experimental results in Spanish-English and German-English plagiarism detection show state-of-the-art performance and provide interesting insights on the use of knowledge graphs.

***Keywords:* —** Cross-language, Plagiarism detection, Knowledge graphs, Multilingual Semantic Network, Distributed representations, Evaluation

3.1 Introduction

Given the vastness of the Web, plagiarism, or the deliberate use of someone else's original material without acknowledging its source, has become a serious problem in areas such as Literature, Education, and Science. The ease of access to copyrighted contents has become matter of concern also for researchers. The problem is exacerbated when the source of plagiarism comes from another language, which is known as cross-language (CL) plagiarism. It is not only the additional difficulty of manually detecting the translation performed, but also the people's lack of knowledge about the ethical issues derived from plagiarism. A recent survey about scholar practices and attitudes (Barrón-Cedeño, 2012), reveals that only 36.25% of students believe that translating text fragments and including them in their work is plagiarism.

Although the CL plagiarism detection task could be potentially performed manually, the amount of data, languages, and time required make it impossible to perform in practice. Current approaches to CL plagiarism detection exploit syntactic and lexical properties of the writing, statistical dic-

tionaries or similarities with a multilingual collection of documents. However, most of these techniques are designed for verbatim copies and performance is reduced when dealing with light and specially heavy cases of plagiarism (Clough and Stevenson, 2011), which include paraphrasing.

In a previous work, we proposed Cross-Language Knowledge Graph Analysis (CL-KGA) (Franco-Salvador et al., 2013a), an approach for CL plagiarism detection aiming at representing context, which employs knowledge graphs both to expand and relate the concepts in a text. Knowledge graphs are generated using BabelNet (Navigli and Ponzetto, 2012a), the most large multilingual semantic network. Thanks to the multilingual representation of concepts available, BabelNet allows for a straightforward comparison of the knowledge graphs obtained in different languages.

In this work, we perform a systematic study of our CL-KGA model. We analyse the impact of the implicit aspects of knowledge graphs on CL plagiarism detection. The research questions we aim to answer are:

- *What is the contribution of the word sense disambiguation (WSD) performed by the knowledge graphs?* These graphs have been explored in the past to perform WSD; our current representation includes disambiguated concepts, which are combined with their intermediate concepts and other disambiguation candidates. We are interested in analysing the performance when the representation is exclusively composed by disambiguated words. This leads us to our next research question.

- *What is the contribution of the vocabulary expansion performed during graph creation?* In our previous work we assumed that the new intermediate concepts that relate the original ones could be a key component in order to obtain a common intersection between related texts. In this work we study this aspect in order to determine if the vocabulary expansion is needed as part of the representation or just as a component during the WSD process itself.

- *What is the relationship between CL-KGA and Cross-Language Explicit Semantic Analysis (CL-ESA)?* These two models represent text by exploiting a collection of multilingual concepts, for instance employing Wikipedia. We are interested in studying the similarities and

the differences between the two models. We aim to clarify the particularities that make the two models perform completely different.

In this paper, we also address key aspects such as the language independence of the knowledge graphs. In addition, we study the relevance of the concepts (nodes) and relations (edges) of the knowledge graphs, and the most suitable threshold to consider that their weighted relations are semantically related. Finally, we compare our model with the state of the art according to different scenarios and criteria: (i) we evaluate CL plagiarism detection using a dataset composed by automatic and manually generated paraphrasing cases of plagiarism; (ii) we study the performance of detection using only paraphrasing cases; and (iii) we compare the computational efficiency of the models and the size of the graphs.

The classical weighting scheme used for the relations between the concepts of the knowledge graphs is based on bag of words generated from short concept definitions as representation of WordNet's concepts. Because it is exclusively based on the original wording of the definition, this type of representation is very explicit. In addition to the detailed study of our previous model, in this work we follow the recent and popular trend in the use of distributed representations of words (Mikolov et al., 2013a; Pennington et al., 2014), and present a new weighting scheme for relations between concepts which generates distributed representations of concepts. Our distributed concepts are generated using the continuous Skip-gram model to obtain vector representations of definitions of concepts. In contrast to the classical weighting, our proposed representation measures semantic relatedness modelling not only of the original words in a definition, but also their context. This allows our scheme to successfully measure similarity between definitions which do not share the same words but have the same meaning.

Experimental results show that the vocabulary expansion is more useful when it is only employed to perform the WSD, which is the essential component of our model. The differences between CL-KGA and CL-ESA are proved favouring the first model, which offers a higher performance thanks to the high coverage of BabelNet and the concept relatedness. Our new weighting scheme using distributed representations of concepts achieves state-of-the-art performance compared to the classical weighting and several alternative CL plagiarism detectors. The study with CL paraphrasing cases proved also CL-KGA superiority on this type of plagiarism. Finally, a comparison

of the computational efficiency of the models demonstrated that our model is more adequate for systems that only require a fast document similarity and perform the indexing in a preprocessing stage.

The rest of the paper is organised as follows. In Section 3.2 we provide an overview of the state of the art in CL plagiarism detection and distributed representations of concepts. In Section 3.3 we describe the knowledge graphs, their weighting schemes, including our new approach, and their main characteristics. In Section 3.4 we describe the CL-KGA model for CL plagiarism detection. Finally, in Section 3.5 we evaluate our approach for Spanish-English and German-English plagiarism detection, comparing our results with several state-of-the-art models. We compare also our new weighting scheme based on distributed representations of concepts with the classical weighting. As part of our analysis, we show the results when detecting only paraphrasing cases and evaluate the computational efficiency of the models.

3.2 Related Work

In this section we first review the approaches of CL similarity analysis that have been used for CL plagiarism detection. Next, we summarise the last advances in the use of distributed representations for conceptual semantic relatedness.

3.2.1 Cross-language Plagiarism Detection

Similarly to some monolingual models for plagiarism (Clough et al., 2003; Maurer et al., 2006), an effective approach for languages with lexical and syntactic similarities, such as Romance and Germanic languages, is the Cross-Language Character N-Gram (CL-CNG) model (Mcnamee and Mayfield, 2004). This model employs vectors of character n-grams to model texts, and uses a weighting scheme and a measure of similarity between vectors such as the cosine similarity.

Several approaches have been proposed to measure CL similarity between any language pair. Cross-Language Explicit Semantic Analysis (CL-ESA) (Potthast et al., 2008) extends the classical ESA (Gabrilovich and Markovitch, 2007) to work in a cross-language scenario. This model represents each text by its similarities with a document collection D i.e., the

topic of a document is qualified using the reference collection D. Despite the fact that the indexing with D is performed at monolingual level, using a multilingual document collection with comparable documents across languages (e.g. Wikipedia), the resulting vectors from different languages can be compared directly. As we discuss in Section 3.3.4.4, our CL-KGA model is slightly related with CL-ESA, i.e., using Wikipedia and representing text using a collection of multilingual concepts. However, our model exploits also vocabulary expansion and relatedness between concepts, and has a variable concept inventory with regard to the text words.

The use of parallel corpora has been explored too. For example, the Cross-Language Alignment-based Similarity Analysis (CL-ASA) model (Barrón-Cedeño, 2012; Barrón-Cedeño et al., 2008; Pinto et al., 2009) is based on statistical machine translation. This model uses a statistical bilingual dictionary — generated with parallel corpora — to translate words and perform text alignment. The alignment takes into account the translation probabilities and the differences in length of equivalent texts in different languages.

An approach exploiting concepts like this paper is the MLPlag (Ceska et al., 2008) model. It uses the EuroWordNet semantic network[1] (Vossen, 2004) to address synonymy and to obtain language independent identifiers of words which can be directly compared. Similarly, the Cross-Language Conceptual Thesaurus based Similarity (CL-CTS) model (Gupta et al., 2012) aims at measuring the similarity between the texts in terms of shared concepts and named entities, using the Eurovoc conceptual thesaurus.[2] It offered an average performance compared to CL-ASA and CL-CNG specially excelling in Spanish-English. In contrast to CL-KGA, these last two models do not employ concept relatedness or vocabulary expansion or WSD, i.e., the assignment of concepts to words is direct and may produce ambiguity. The Cross-Language Knowledge Graph Analysis (CL-KGA) model (Franco-Salvador et al., 2013a,b) uses a multilingual semantic network to create knowledge graphs that model the context of documents. The model achieved interesting results for CL plagiarism detection, also in cases of paraphrasing (Franco-Salvador et al., 2014a). However, it left unanswered questions — relationship with CL-ESA, contributions of WSD, vocabulary expansion,

[1] http://www.illc.uva.nl/EuroWordNet/
[2] http://eurovoc.europa.eu/

etc. — and room for improvement — weighting scheme and parameter tuning —, that we address in this paper.

Other CL similarity analysis approaches such as the Cross-Language Latent Semantic Indexing (CL-LSI) (Dumais et al., 1997b) or Similarity Learning via Siamese Neural Network (S2Net) (Yih et al., 2011) linear projection models, could be employed as well for plagiarism detection. In this work we focus on comparing our model with those models that have been evaluated in the past on CL plagiarism detection.

In recent years, plagiarism detection has been actively addressed in the Evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN)[3] at the Conference and Labs of the Evaluation Forum (CLEF). The plagiarism detection shared task (Potthast et al., 2014) encourages participants to submit detectors and compete to identify plagiarism cases in the provided corpus. The 2010 and 2011 editions (Potthast et al., 2010a, 2011b) contained also cross-language partitions in German-English and Spanish-English, which we used for our evaluation. In 2015 the task invited for the first time to submit datasets (Franco-Salvador et al., 2015a; Potthast et al., 2015), increasing participation and including new languages such as Urdu, Persian and Chinese. Similarly to Corezola Pereira et al. (2010), the most popular technique to handle CL plagiarism detection at PAN involved machine translation systems, translating all the documents to the language of comparison beforehand. However, this introduces a heavy dependence on the availability of Machine Translation (MT) systems and their quality. In addition, we consider that those methods are not pure CL detectors, but excellent monolingual plagiarism detection systems with a MT preprocessing. Hence, we compare our proposed model to CL plagiarism detection systems that do not depend on fully-fledged MT systems.[4] In Barrón-Cedeño et al. (2013) we can find a comparison of CL-ASA and CL-CNG using the Spanish-English partition of PAN'11 competition, where the models have been also compared with a system (T+MA) employing MT to analyse the similarities at monolingual level. The paper concluded that T+MA is superior in short cases of plagiarism but very close to CL-ASA, which achieved a higher precision in all experiments and better performance for long cases of plagiarism.

[3] http://pan.webis.de/

[4] CL-ASA employs a statistical dictionary but includes a complex language alignment model.

A comparison of the CL-CNG, CL-ESA, and CL-ASA models for CL plagiarism detection has been provided in Potthast et al. (2011a). Different performances were observed depending on the task, languages, and dataset employed. For instance, CL-ESA and CL-CNG were more stable across datasets, obtaining a higher performance on the Wikipedia comparable dataset. In contrast, CL-ASA obtained better results on the JRC-Acquis parallel dataset. Finally, CL-CNG achieved lower quality for language pairs without lexical and syntactic similarities. Therefore, in this work we decided to compare CL-KGA with all these models.

3.2.2 Distributed Representations for Conceptual Semantic Relatedness

We introduce a new weighting scheme, based on the use of distributed representations of concepts, to measure the semantic relatedness between concepts belonging to a knowledge graph. In recent years, the use of log-linear models has been proposed as an efficient way to generate distributed representations of words (Mikolov et al., 2013a), since they reduce the complexity of the neural network hidden layer thereby improving efficiency. These representations have proved to be an excellent alternative for computing semantic relatedness with models such as the continuous Skip-gram model[5] (Mikolov et al., 2013a,b) or GloVe[6] (Pennington et al., 2014). Recent works have explored also the possibility of modelling words senses (i.e., synsets) for semantic relatedness using distributed representations. Faruqui et al. (2015) refine vector space representations using relational information from semantic resources such as WordNet or FrameNet (Baker et al., 1998). Aletras and Stevenson (2015) provide representations of synonym words derived from WordNet and exploit its hierarchy to generate synset vectors. There has been also interest in representing BabelNet synsets using distributed representations. SensEmbed (Iacobacci et al., 2015) uses Babelfy (Moro et al., 2014) to disambiguate the complete Wikipedia to the BabelNet synset inventory. Then, the continuous Bag of Words model (CBOW) (Mikolov et al., 2013a) is used on top of Wikipedia's disambiguated text to generate the distributed representation of synsets. Finally, further refinements (including properties of the BabelNet topology) are employed to measure semantic relatedness.

[5]The continuous Skip-gram model is available in the word2vec toolkit: `https://code.google.com/p/word2vec/`

[6]`http://nlp.stanford.edu/projects/glove/`

Since we aim at weighting the ∼262 million of relations of BabelNet, we have to employ a fast and efficient model. As disadvantages SensEmbed has the computational complexity required to disambiguate the ∼5 million of pages contained in the English Wikipedia, the possible errors that WSD may introduce (despite the excellent ∼70% of F_1 score with Babelfy for English), the unbounded range of weights that SensEmbed provides, and the low performance of CBOW compared to the continuous Skip-gram model when measuring semantic relatedness (Mikolov et al., 2013a). In Section 3.3.3.2 we opted for an efficient solution which exploits the high-quality definitions provided for the BabelNet's synsets (i.e., glosses) and the Skip-gram model.

3.3 Knowledge Graphs

A knowledge graph is a weighted and directed graph that expands and relates the concepts[7] belonging to a text. We may consider a knowledge graph as a subset of an original knowledge base focused on the concepts pertaining to a text. Knowledge graphs have been used for Natural Language Processing (NLP) tasks such as network text analysis (Popping, 2003), semantic relatedness (Navigli and Ponzetto, 2012b), WSD (Navigli and Ponzetto, 2012a), semantic parsing (Heck et al., 2013), sentiment analysis (Franco-Salvador et al., 2015b) — also from a WSD perspective —, or in cross-language scenarios: CL plagiarism detection (Franco-Salvador et al., 2013a), and CL document retrieval and categorization (Franco-Salvador et al., 2014b). In Figure 3.1 we show an example of a knowledge graph.

In order to generate knowledge graphs that allow for a direct comparison across languages, we need a knowledge base with a multilingual dimension of the concepts. We could use EuroWordNet or Wikipedia,[8] although in this work we employ the BabelNet multilingual semantic network, since it offers the larger set of concepts and languages to date.

3.3.1 BABENET

BabelNet[9] 2.5 (Navigli and Ponzetto, 2012a) is a multilingual semantic network whose concepts and relations are obtained from the automatic mapping

[7] Each word has a number of senses. We define "concept" as any of those senses, which may be represented via synsets (see Section 3.3.1).

[8] https://en.wikipedia.org/

[9] http://babelnet.org

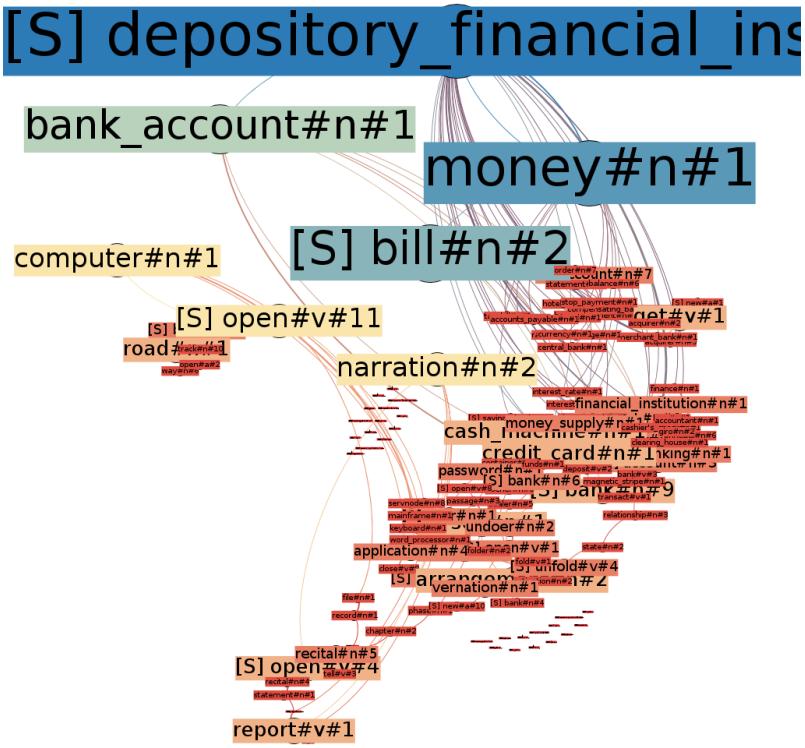

Figure 3.1. Knowledge graph built from the sentence "I opened a new bank account" (source words: ("open#v, new#a, bank#n, account#n")). Larger boxes represent concepts with higher connectivity.

onto WordNet of Wikipedia, OmegaWiki,[10] Wiktionary,[11] Wikidata,[12], and Open Multilingual WordNet.[13] Therefore, BabelNet is a multilingual "encyclopedic dictionary" that combines lexicographic information with wide-coverage encyclopedic knowledge. Concepts in BabelNet are represented similarly to WordNet, i.e., by grouping sets of synonyms in the different languages into multilingual synsets. The syntactic categories are exactly the same offered by WordNet: noun, verb, adjective, and adverb. Multi-

[10] http://omegawiki.org
[11] http://wiktionary.org
[12] http://wikidata.org
[13] http://compling.hss.ntu.edu.sg/omw/

lingual synsets contain lexicalizations from WordNet and Open Multilingual WordNet synsets, the corresponding Wikipedia pages, the OmegaWiki, Wiktionary, and Wikidata entries, and additional translations by a statistical machine translation system. The relations between synsets are collected from WordNet, Open Multilingual WordNet, and from Wikipedia's hyperlinks between pages. The version 2.5 of BabelNet includes 9,348,287 synsets, covers 50 languages,[14] and has a WordNet-Wikipedia mapping correctness of 91% (Navigli et al., 2013).

3.3.2 CREATION OF THE KNOWLEDGE GRAPHS

Similarly to the aforementioned works, we followed the approach described by Navigli and Lapata (2010) to create our knowledge graphs, which is a four step-approach described as follows:

(i) Part-of-speech tagging and lemmatization Initially we process a text fragment d with tokenization, multi-word extraction, part-of-speech (POS) tagging, and lemmatization[15] to obtain the list of tuples (lemma,tag) T. We discard POS tags not available in BabelNet.

(ii) Populating the graph with initial concepts Next, we create an initially-empty knowledge graph $G = (V, E)$, i.e., such that $V = E = \emptyset$. We populate the vertex set V with the set S_K of all the synsets in BabelNet which contain any <lemma,tag> tuple in T in the text fragment language L, that is:

$$S_K = \bigcup_{t \in T} \text{Synsets}_L(t), \tag{3.1}$$

where $\text{Synsets}_L(t)$ is the set of synsets which contains a <lemma,tag> tuple t in the language of interest L.

[14] Although in this work we employed BabelNet 2.5, the more recent BabelNet 3.0 offers 13,789,332 synsets and 271 languages via a RESTful API. We selected the previous version in order to avoid depending on the API and work offline which allows for a faster creation of knowledge graphs.

[15] Due to our multilingual focus we used TreeTagger: `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`. For the multi-word extraction we implemented our own tool based on pattern matching.

(iii) Creating the knowledge graph We create the knowledge graph by searching in BabelNet the set of paths P connecting pairs of synsets in V. Formally, for each pair $\{v, v'\} \in V$ such that v and v' do not share any lexicalization[16] in T, for each path in BabelNet $v \rightarrow v_1 \rightarrow \cdots \rightarrow v_n \rightarrow v'$, we set: $V := V \cup \{v_1, \ldots, v_n\}$ and $E := E \cup \{(v, v_1), \ldots, (v_n, v')\}$. That is, we add all the path vertices and edges to G. Following the approach of Navigli and Ponzetto (2012a), the path length is limited to maximum length of 3, in order to avoid an excessive semantic drift.[17]

As a result of populating the graph with intermediate edges and vertices, we obtain a knowledge graph which models the semantic context of text fragment d.

(iv) Knowledge graph weighting The next step consists in weighting all the concepts and semantic relations of the knowledge graph G. For weighting concepts, different methods have been tested in the past, including the PageRank (Page et al., 1998) algorithm. In this work, we score each concept using its own outdegree, which has proved to obtain the best results (Navigli and Ponzetto, 2012a). For weighting relations we will describe in detail the two methods that we evaluated in this work. We normalise weights as a function of the total sum of the outgoing relations.

3.3.3 WEIGHTING OF THE SEMANTIC RELATIONS

Relations in BabelNet are weighted to quantify the strength of the association between synsets. Knowledge graphs use these weights in order to weight their relations. In this section we describe the original approach which was employed by Navigli and Ponzetto (2012a) in order to measure this degree of association between synsets. Next, in Section 3.3.3.2 we present our new method based on distributed representations of concepts for weighting their relations.

[16] This prevents different senses of the same term from being connected via a path in the resulting knowledge graph.

[17] At this point, we removed the edges below a certain threshold that represents a low semantic relationship (see Section 3.5.3).

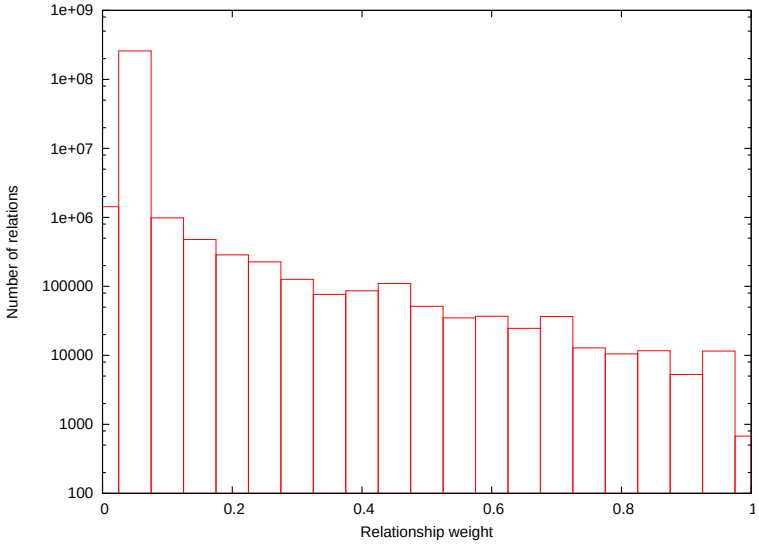

Figure 3.2. Distribution of relation weights in BabelNet using the Dice's coefficient-based weighting.

3.3.3.1 Dice's Coefficient-based Measure of Semantic Relatedness

The weights between relations provided in the original BabelNet 1.0 were computed using methods based on Dice's coefficient (Jackson et al., 1989). Two different strategies were employed to leverage the high-quality definitions from WordNet, and the large amounts of hyperlinked text from Wikipedia. Similarly to the Extended Gloss Overlap measure (Banerjee and Pedersen, 2003), for computing the semantic relatedness between two WordNet synsets s and s', they first are independently represented using a bag-of-words (BOW) representation including all the synonyms of the synsets and the lemmatised words of their glosses.[18] Stopwords are removed. The list of directly linked synsets is also included for s and s'. Next, they employ the Dice's coefficient over s and s' to measure the relationship between the two WordNet synsets:

$$\text{Semantic Relatedness}(s, s') = \frac{2|s \cap s'|}{|s| + |s'|} \tag{3.2}$$

[18] A gloss is a short definition of the sense represented within that synset.

The relationship between two synsets corresponding to Wikipedia pages is computed using a co-occurrence based method (Ito et al., 2008; Ye et al., 2009), which exploits the large amount of hyperlinked text available in Wikipedia. Given two Wikipedia page synsets w and w', the frequency of occurrence of each individual page (f_w and $f_{w'}$) is computed as the number of hyperlinks found in Wikipedia which point to it. The co-occurrence frequency of w and w' ($f_{w,w'}$) is computed as the number of times these links occur together within a context.[19] The relationship between w and w' applies the Dice's coefficient to these frequencies:

$$\text{Semantic Relatedness}(w, w') = \frac{2f_{w,w'}}{f_w + f_{w'}} \tag{3.3}$$

Using this weighting scheme, we depict in Figure 3.2 a histogram of the distribution of BabelNet's relation weights. We observe that only $\sim$15 million relations are weighted. In our evaluation we refer always to the CL-KGA model with this weighting scheme unless otherwise stated (see section 3.3.3.2).

3.3.3.2 Distributed Representations of Concepts for Computing Semantic Relatedness

The weighting described in Section 3.3.3.1 is based on an accurate and explicit representation of concepts, i.e., a concept fingerprint uses the information of its short and clear definition — in the case of WordNet —, or information of samples of text explicitly mentioning that concept — in the case of Wikipedia. However, those definitions and samples of text do not cover all of the possible contexts in which a concept may appear, and the weighting scheme is not able to infer more contexts. In contrast, the use of distributed representations has proved that the context is modelled in a more abstract[20] but precise manner, e.g. citing the words of Mikolov et al. (2013a), *"it was shown for example that vector("King") - vector("Man") + vector("Woman") results in a vector that is closest to the vector representation of the word Queen"*. This property, allowed their authors to use these representations in scenarios in which the word was never seen before, but its

[19] Navigli and Ponzetto (2012a) employed a sliding window of 40 words as context.

[20] The distributed representations, also known as continuous representations or embeddings, represent information (e.g. words or concepts) using vectors of floating numbers.

context is the most adequate, e.g. tasks of sentence completion. In this work we aim at measuring the strength of association between concepts modelling their representing context using distributed representations. We introduce a new weighting scheme based on the generation of distributed representations of concepts. In order to generate our distributed representations of concepts, we exploit the high-quality definitions provided by the BabelNet's synsets (i.e., glosses[21]) and the Skip-gram model.

Preamble and definitions The continuous Skip-gram model (Mikolov et al., 2013a,b) is an iterative algorithm which attempts to maximise the classification of the context surrounding a word. Formally, given a word w_t and its surrounding words w_{t-c}, w_{t-c+1}, ..., w_{t+c} inside a window of size $2c+1$, the goal is to maximise the average of the log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log \mathrm{p}(w_{t+j}|w_t) \tag{3.4}$$

Although $\mathrm{p}(w_{t+j}|w_t)$ can be estimated using the softmax function (Barto, 1998), its normalisation depends on the vocabulary size W which makes its usage impractical for high values of W. For this reason, more computationally efficient alternatives are used instead. In this work we used the negative sampling (Mikolov et al., 2013b), a simplified version of the Noise Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2012; Mnih and Teh, 2012), which basically uses logistic regression to distinguish the target word from a noise distribution, having k negative samples for each word. Experimental results in Mikolov et al. (2013b) showed that the negative sampling offers better results at semantic level compared to NCE and Hierarchical softmax (Morin and Bengio, 2005). Sentence vectors (SenVec) (Le and Mikolov, 2014) follow Skip-gram model to train a special vector $\vec{v}$ representing a complete sentence. Basically, the model uses all words in the sentence as context to train the vector representing its content. In contrast, the original Skip-gram model employs a fixed size window to determine the context (surrounding words) of the iterated words of a sentence. Next we detail the

[21] Although the approach described in Section 3.3.3.1 only uses the glosses provided in BabelNet for WordNet synsets, our weighting scheme is based on the most recent versions of the semantic network, which include also glosses for Wikipedia, OmegaWiki, Wiktionary, and Wikidata-derived synsets.

four-step method we used for weighting the BabelNet semantic relations using the continuous Skip-gram and SenVec models:

(i) Getting high-confidence word vectors The first step consists in obtaining a collection of vectors of words $\vec{V}_W$ from encyclopedic knowledge using the Skip-gram model.[22] $\vec{V}_W$ will provide a precise and accurate representation of the type of context we are interested in modelling, i.e., sense definitions. For this purpose we used the complete Wikipedia dump[23] of January 2015 and extracted vectors for $\sim$15 million of words.

(ii) Generating distributed representations of glosses Next, for all English glosses[24] available in BabelNet, we employ SenVec to generate their distributed representations $\vec{V}_G$. The $\vec{V}_W$ collection is used as input word vectors in order to provide the glosses with enough context to generate representative vectors. The $\vec{V}_G$ collection contains 3,857,795 gloss vectors.

(iii) Generating distributed representations of concepts (synsets) BabelNet provides a gloss for each available source (WordNet, Wikipedia, OmegaWiki, etc.) and it is very frequent to have more than one gloss per synset. We take advantage of this observation by generating vectors for all glosses, independently of their source. We get the final representation $\vec{v}_s$ of a synset s by averaging all its available gloss vectors: $\vec{v}_s = n^{-1} \sum_{i=1}^{n} \vec{v}_g(s)_i$, where $(\vec{v}_g(s)_1, \vec{v}_g(s)_2, ..., \vec{v}_g(s)_n) \in \vec{V}_G$ are all gloss vectors available for the synset s. This averaging of distributed vectors has been successfully applied in the past for classification tasks (Franco-Salvador et al., 2015c,d; Le and Mikolov, 2014).

[22] We used 300-dimensional vectors, context windows of size 8, and 25 negative words for each sample. We preprocessed the text with lowercased word, tokenisation, and removing the words of unit length. We used the same configuration for the SenVec vectors.

[23] `https://en.wikipedia.org/wiki/Wikipedia:Database_download`

[24] The multilingualism of BabelNet synsets allows to obtain multilingual vector representations using only English glosses.

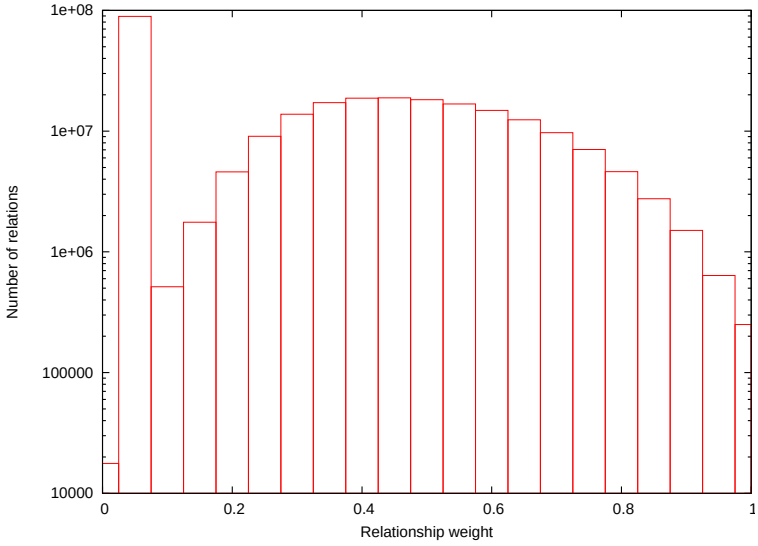

Figure 3.3. Distribution of relation weights in BabelNet using distributed concept weighting.

(iv) Weighting BabelNet's semantic relations Finally, in order to compute the strength of each pair of synsets (s, s') with a semantic relation in BabelNet, we use the cosine distance between the synset vectors $\vec{v}_s$ and $\vec{v}_{s'}$:

$$\text{Semantic Relatedness}(s, s') = \frac{\vec{v}_s \cdot \vec{v}_{s'}}{||\vec{v}_s|| ||\vec{v}_{s'}||} \tag{3.5}$$

In Figure 3.3 we can see a histogram with the distribution of the weights of the relations of BabelNet using our new weighting scheme. Note that we weighted ∼172 million of semantic relations compared to the ∼15 million of relations originally weighted with the method described in Section 3.3.3.1. In addition, if we observe Figure 3.2, we can appreciate differences in the weight distributions. Ours is more similar to a Gaussian distribution, whereas the former seems to fit a decreasing logarithmic scale. In our evaluation, we refer to the CL-KGA model that employs the proposed weighting scheme using the "Distributed Concept Weighting" (DCW) tag.

3.3.4 CHARACTERISTICS OF THE KNOWLEDGE GRAPHS

Knowledge graphs have several implicit characteristics that make them adequate for NLP tasks related to similarity analysis such as CL plagiarism detection. These characteristics have been used by the CL-KGA model in the past, but they have never been analysed independently for a CL plagiarism detection perspective. In this work we aim at studying the most relevant ones: WSD, vocabulary expansion, language independence, and representation of text using a multilingual collection of concepts.

3.3.4.1 Word Sense Disambiguation

Knowledge graphs have been successfully used in the past to perform WSD (Navigli and Ponzetto, 2012a). As we stated, the graphs created in Section 3.3.2 contain a set of S_K synsets for each <lemma,tag> tuple extracted from an original text fragment d. However, only one of these synsets corresponds to the disambiguation of the tuple. That means that we are introducing paths between synsets which are not real senses of the meaning of d. The original CL-KGA model kept all candidate synsets of the tuples and the intermediate paths in order to counterbalance possible errors that may be produced if we keep only the disambiguation synsets. We assumed that if there is enough context in d, the knowledge graph G will contain a considerably higher concept mass surrounding the real concepts representing the text d and the error will be reduced. In order to validate our theory, we introduce three additional graph variations:

(i) Knowledge graphs restricted to disambiguation source synsets These graphs use Equation 3.6 to select the disambiguation s_{WSD} among the S_K synsets of each tuple, where $\text{score}(s)$ is the outdegree of the synset s in the graph G. Then we filter the path set P which created the graph G, and keep only those paths which contain a disambiguation synset as starting and ending point. As a result we obtain the filtered graph G_f where we will remove the noise provided for the concepts which are not related to the original text d. We use the "WSD path filter" tag to refer to this model in the evaluation.

$$s_{\text{WSD}} = \underset{s \in S_K}{\arg\max} \quad \text{score}(s) \tag{3.6}$$

(ii) Knowledge graphs for extracting weighted disambiguations Using the knowledge graph G_f, this representation removes the intermediate concepts between source synsets, i.e., we use the knowledge graphs only to disambiguate d and discard the vocabulary expansion. However, we keep the original weights of the concepts of the graph G_f, which are generated using the vocabulary expansion. We use the "WSD concepts" tag to refer to this model in the evaluation.

(iii) Knowledge graphs for extracting bag-of-words of disambiguations Similarly to the previous model, we extract the disambiguations by keeping only the source synsets of the knowledge graph G_f. In contrast, in order to analyse if the weighting produced when keeping only disambiguations is noisy, we include these disambiguation concepts in a bag-of-words without weights. We use the "WSD concepts w/o weighting" tag to refer to this model in the evaluation.

3.3.4.2 Vocabulary Expansion

The vocabulary expansion of the knowledge graphs is an interesting characteristic to study in CL plagiarism detection. When plagiarising, the text is often obfuscated via paraphrasing. The use of knowledge graphs allows to relate the original concepts of a text, including also intermediate concepts between them. If the text has been modified, it is quite likely having an intersection between the expanded concepts of the original text and the plagiarised one. This vocabulary expansion has proved to be useful in tasks such as sentiment analysis (Franco-Salvador et al., 2015b). In the evaluation we will compare the performance using vocabulary expansion for CL plagiarism detection using the models introduced in Section 3.3.4.1.

3.3.4.3 Language Independence

As we mentioned at the beginning of Section 3.3, using BabelNet to generate knowledge graphs allows to compare them directly despite being generated from texts in different languages. This is possible because the multilingual dimension of the BabelNet's concepts. To illustrate this, let us describe an example. When we query BabelNet with the English word "plagiarism", the first two sense ID's we obtain are plagiarism#n#1 — "A piece of writing that has been copied from someone else and is presented as being your own

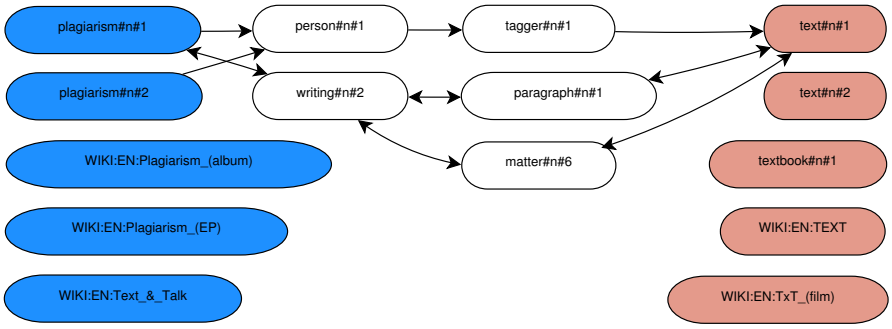

Figure 3.4. Knowledge graph built from the English sentence "text with plagiarism" (source words: ("text#n","plagiarism#n")). The coloured nodes are the different senses of the original words.

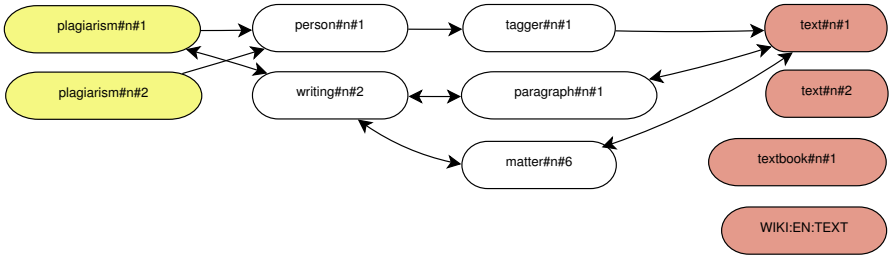

Figure 3.5. Knowledge graph built from the Spanish sentence "texto con plagio" (source words: ("texto#n","plagio#n")).

work" —, and plagiarism#n#2 — "The act of plagiarizing; taking someone's words or ideas as if they were your own". If we query now BabelNet with the Spanish word "plagio" (plagiarism), we get exactly the same two sense ID's on top of the results. If we observe the words contained inside the senses, we can see that BabelNet employed lexicalizations of the senses in different languages to match our query. In Figures 3.4 and 3.5 we can see the knowledge graphs obtained for the English sentence "text with plagiarism" and its translation into Spanish. As can be seen, both graphs share the same core concepts and can be compared directly with some graph similarity algorithm.

3.3.4.4 Representation of Text using a Multilingual Collection of Concepts

We are interested in analysing the analogies of our knowledge graph-based model with CL-ESA.[25] Both represent text using a collection of multilingual concepts. In addition, the concept inventory and the multilingual dimension is extracted (not completely in our case) using Wikipedia.[26] Finally, in the worst case, if our model has not enough context to generate a representative knowledge graph, we will have a non-related (and possibly dense) collection of multilingual concepts. In that case, it is possible that our model would produce a similar "wrong" collection of concepts for both languages and would exploit the similarities between them to counterbalance the conceptual and relational errors, i.e., in a similar way to the nature of CL-ESA. However, the differences do not go beyond. We employ a multilingual semantic network to extract the concepts of a text and, in order to model its context, we use knowledge graphs to expand and relate these concepts. In contrast, CL-ESA employs a collection of Wikipedia pages as concepts, and computes the similarities directly with the original text. This method allows to model the context but it is not computing relatedness between concepts and nor expanding the vocabulary or performing WSD. Finally, the fixed collection of pages that CL-ESA employs (several thousands compared to the millions of BabelNet) is restricting the concept inventory and the possibility of modelling the context exploiting the analogies with concepts. In Section 3.5 we compare our model with CL-ESA to show the differences in performance at detecting CL plagiarism.

3.4 Cross-language Knowledge Graph Analysis (CL-KGA)

In this section we describe more in detail the CL-KGA model for CL plagiarism detection. We discuss the original description of Franco-Salvador et al. (2013a) and the algorithm for the detailed analysis and postprocessing of similarities between text fragments. Given a source document d_L in a language L and a suspicious document $d'_{L'}$ in a language L', we compare documents in a four-step process:

(i) Segmentation into text fragments In order to detect plagiarised sections of text between the documents d_L and $d'_{L'}$, we first segment them to

[25] Most of our statements are valid also for ESA.

[26] We assume a classical CL-ESA model based on Wikipedia.

obtain the sets of fragments F_L and $F'_{L'}$. We use a 5-sentence sliding window with a 2-sentence step to make the segmentation into fragments.

(ii) Creation of knowledge graphs We next use the method described in Section 3.3.2 to create the graph collections GC and GC′ of the text fragments F_L and $F'_{L'}$. At this point the language tag has been removed due to the graph multilingualism.

(iii) Comparison of knowledge graphs For each pair of graphs (G, G'), $G \in$ GC and $G' \in$ GC′, we adapt the algorithm of Montes y Gómez et al. (2001) to compare their similarity and to obtain the set of similarities SG between graph pairs. We calculate the similarity between the concepts in the two graphs using Dice's coefficient:

$$\mathrm{S}_c(G, G') = \frac{2 \cdot \sum_{c \in \mathrm{V}(G) \cap \mathrm{V}(G')} \mathrm{w}(c)}{\sum_{c \in \mathrm{V}(G)} \mathrm{w}(c) + \sum_{c \in \mathrm{V}(G')} \mathrm{w}(c)}, \tag{3.7}$$

where $\mathrm{w}(c)$ is the weight of a concept c (see Section 3.3.2). Likewise, we calculate the similarity between the relations as:

$$\mathrm{S}_r(G, G') = \frac{2 \cdot \sum_{r \in \mathrm{E}(G) \cap \mathrm{E}(G')} \mathrm{w}(r)}{\sum_{r \in \mathrm{E}(G)} \mathrm{w}(r) + \sum_{r \in \mathrm{E}(G')} \mathrm{w}(r)}, \tag{3.8}$$

where $\mathrm{w}(r)$ is the weight of a semantic relation r (see Section 3.3.3). We interpolate[27] the two above measures of conceptual (S_c) and relational (S_r) similarity to obtain an integrated measure $\mathrm{S}_g(G, G')$ between knowledge graphs:

$$\mathrm{S}_g(G, G') = a \cdot \mathrm{S}_c(G, G') + b \cdot \mathrm{S}_r(G, G'), \tag{3.9}$$

where a and b, $a + b = 1$, are the parameters of the relevance of concepts and relations respectively. In Figure 3.6 we can see the differences among

[27] The original CL-KGA combined S_c and S_r with $\mathrm{S}_g(G, G') = \mathrm{S}_c(G, G')(a + b \cdot \mathrm{S}_r(G, G'))$. However, we observed that the current equation allows to ease the tuning of relevance of concepts and relations without affecting the performance.

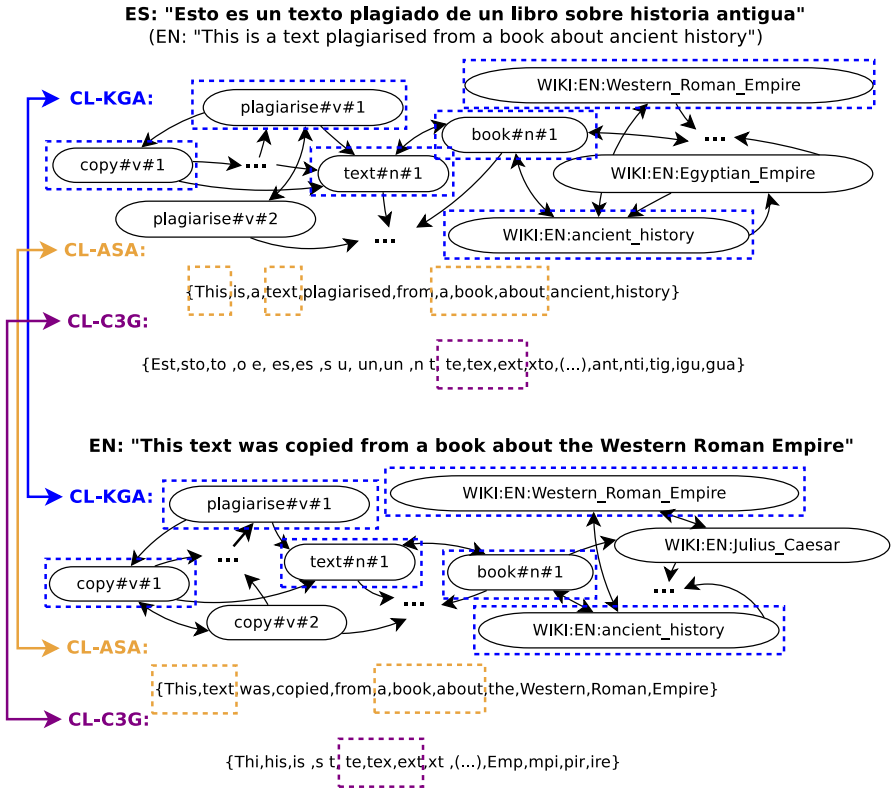

Figure 3.6. Toy example to illustrate the capability of detection of the CL-KGA model compared to the CL-ASA and the CL-C3G models. Higher intersection of same-coloured boxes between languages represents a higher potential plagiarism case retrieval.

CL-KGA, CL-C3G, and CL-ASA when detecting CL plagiarism. Thanks to the aforementioned characteristics (see Section 3.3.4), the use of knowledge graphs allows to detect similarity even when the paraphrasing is employed and the languages are not syntactically and semantically related. Note that the procedure described so far is the basic model of the candidate retrieval task (Barrón-Cedeño et al., 2013; Potthast et al., 2011a), which needs a detailed analysis component to detect plagiarism cases.

(iv) Detailed analysis and postprocessing of similarities Once we obtain the set SG with the similarities between the text fragments of the documents d_L and $d'_{L'}$, we employ the method introduced in Barrón-Cedeño

Algorithm 3.1 Detailed analysis and postprocessing.

Input: the set of similarities $\mathrm{SG} = \{\mathrm{S}_g(G, G')\}$ between all the pairs of graphs (G, G'), $G \in \mathrm{GC}$ and $G' \in \mathrm{GC}'$

Output: PlagCases, a set containing the offsets of all the identified cases of plagiarism

1: PlagCases $\leftarrow \{\}$
2: **for each** $G \in \mathrm{GC}$ **do** # Detailed analysis
3: $P_G \leftarrow \mathrm{argmax}^5_{G' \in \mathrm{GC}'} \mathrm{S}_g(G, G')$
4: **repeat** # Postprocessing
5: **for each combination of pairs** $p \in P_G$ **do**
6: **if** $\delta(p_i, p_j) < \mathrm{thres}_1$ **then**
7: merge_fragments(p_i, p_j)
8: **until no change**
9: PlagCases $=$ PlagCases $\cup \{\mathrm{offsets}(p \in P_G \mid |p| > \mathrm{thres}_2)\}$
10: **return** PlagCases

(2012) and Barrón-Cedeño et al. (2013) to analyse the values and determine which fragments of text are cases of plagiarism. This method was originally designed to process the similarity scores of CL-ASA and CL-CNG and it is described in Algorithm 3.1. Basically, for each text fragment of d_L we obtain P_G, i.e., the top 5 most similar fragments of document $d'_{L'}$ (line 3). Then, we start an iterative process until convergence that merges the fragments of P_G with a distance δ lower than a threshold thres_1 (lines 6-7). Finally, we select as plagiarism the cases which combine more than thres_2[28] text fragments (line 9). The function offsets$(\cdot)$ provides with the beginning and end offsets of the plagiarism case. This algorithm has been used for evaluating all the models compared in the evaluation section.

3.5 Evaluation

In this section we compare the different variants of our CL-KGA model with several state-of-the-art approaches in the task of CL plagiarism detection. Given a suspicious document d_L in a language L and a collection of source

[28] In this work we used the original thresholds employed in Barrón-Cedeño (2012) and Barrón-Cedeño et al. (2013): $\mathrm{thres}_1 = 1,500$ and $\mathrm{thres}_2 = 2$.

documents $D'_{L'}$ in a language L', the task is to identity all the plagiarised fragments of d_L from the document collection $D'_{L'}$.

3.5.1 DATASETS

To evaluate our model we selected the datasets employed for the CL plagiarism detection competition of PAN at CLEF.[29] The two available datasets, PAN-PC-10[30] and PAN-PC-11,[31] contain the used Spanish-English (ES-EN) and German-English (DE-EN) partitions. Both datasets contain plagiarism cases generated using machine translation with Google translate.[32] In addition, PAN-PC-11 contains also cases of plagiarism with manual correction after automatic translation. These cases are CL paraphrasing cases of plagiarism. We selected the complete PAN-PC-10 dataset to perform the comparison of the CL-KGA weighting schemes and the tuning of our parameters. Then, we used the PAN-PC-11 dataset to perform the evaluation of the CL-KGA model and the comparison with the state-of-the-art. In Table 3.1 we can see the statistics of the datasets.

3.5.2 METHODOLOGY

As evaluation metric we selected the measures employed at the PAN shared task: precision, recall, granularity, and plagdet (Potthast et al., 2010b). Let S denote the set of plagiarism cases in the suspicious documents, and let R denote the set of plagiarism detections the detector reports for these documents. A plagiarism case $s \in S$ represents a reference to the characters that form that case. Likewise, a plagiarism detection $r \in R$ is represented as r. Based on these representations, the precision and the recall at character level of R under S are measured as follows:

$$\text{precision}(S,R) = \frac{1}{|R|} \sum_{r \in R} \frac{\left| \bigcup_{s \in S} (s \sqcap r) \right|}{|r|}; \tag{3.10}$$

[29] http://www.clef-initiative.eu/

[30] http://www.uni-weimar.de/en/media/chairs/webis/corpora/corpus-pan-pc-10/

[31] http://www.uni-weimar.de/en/media/chairs/webis/corpora/corpus-pan-pc-11/

[32] https://translate.google.com/

PAN-PC-10			
ES-EN documents		**DE-EN documents**	
Suspicious	277	Suspicious	280
Source	187	Source	414
Plagiarism cases {ES,DE}-EN			
Automatic translation			9,598
PAN-PC-11			
ES-EN documents		**DE-EN documents**	
Suspicious	304	Suspicious	251
Source	202	Source	348
Plagiarism cases {ES,DE}-EN			
Automatic translation			5,142
Automatic translation + Manual correction			433

Table 3.1. Statistics of PAN-PC-10 and PAN-PC-11 cross-language plagiarism detection partitions.

$$\text{recall}(S,R) = \frac{1}{|S|}\sum_{s\in S}\frac{|\bigcup_{r\in R}(s \sqcap r)|}{|s|}, \tag{3.11}$$

where $s \sqcap r = s \cap r$ if r detects s and $\emptyset$ otherwise. Note that precision and recall do not account for the fact that plagiarism detectors sometimes report overlapping or multiple detections for a single plagiarism case. To address this issue, we also measured the detector's granularity:

$$\text{granularity}(S,R) = \frac{1}{|S_R|}\sum_{s\in S_R}|R_s|, \tag{3.12}$$

where $S_R \subseteq S$ are cases detected by detectors in R, and $R_s \subseteq R$ are detections of S, i.e., $S_R = \{s | s \in S \wedge \exists r \in R : r \text{ detects } s\}$ and $R_s = \{r | r \in R \wedge r \text{ detects } s\}$. The three previous measures were integrated together in order to obtain an overall score for plagiarism detection (plagdet):

$$\text{plagdet}(S,R) = \frac{F_1(S,R)}{\log_2(1+\text{granularity}(S,R))} \tag{3.13}$$

	System	Description
(a)	CL-KGA (BabelNet 1.0)	Results of cross-language knowledge graph analysis using BabelNet 1.0 and the classical weighting.
	CL-ASA	Cross-language alignment based similarity analysis.
	CL-ESA	Cross-language explicit semantic analysis.
	CL-C3G	Cross-language character n-gram.
(b)	statDict	Translate all words with a statistical dictionary and apply Dice's coefficient to compare.
	POS + statDict	statDict with a POS tagging and lemmatization preprocessing.
	POS + statDict + MFS	Same as previous but disambiguating words using the most frequent sense baseline.
(c)	CL-KGA	CL-KGA using classical weighting (See Section 3.3.3.1).
	CL-KGA (DCW)	CL-KGA using the distributed concept weighting (see Section 3.3.3.2).
	CL-KGA (WSD path filter)	CL-KGA keeping only paths related to WSD concepts (see Section 3.3.4.1).
	CL-KGA (WSD concepts)	CL-KGA keeping only weighted WSD concepts (see Section 3.3.4.1).
	CL-KGA (WSD concepts w/o weighting)	CL-KGA keeping only a BOW of WSD concepts (see Section 3.3.4.1).
	CL-KGA (DCW) (WSD concepts w/o weighting)	Same as previous using the distributed concept weighting.

Table 3.2. Models compared in the evaluation: (a) state-of-the-art approaches; (b) baselines; (c) proposed CL-KGA model and variants (using BabelNet 2.5).

We compared our CL-KGA model with the state-of-the-art CL-ESA,[33] CL-ASA[34] and CL-C3G models.[35] We included also the results obtained previously by the original CL-KGA (Franco-Salvador et al., 2013a) — CL-KGA (BabelNet 1.0) from here —, and those obtained by the CL-KGA variations introduced in Section 3.3.4.1: CL-KGA (WSD path filter), CL-KGA (WSD concepts), and CL-KGA (WSD concepts w/o weighting). We showed the results of our model using the distributed concept weighting for the CL-KGA model and also for its better performing variant when employing the classic weighting. We introduced also three baselines: (i) *statDict*, which used a statistical dictionary — the same used by CL-ASA — to obtain all

[33] We used 10,000 Spanish-German-English comparable Wikipedia pages as document collection. All pages contain more than 10,000 characters and were represented using the term frequency-inverse document frequency (TF-IDF) weighting. The similarities are computed using the cosine similarity and the IDF of the words of the documents to index is calculated from Wikipedia.

[34] We used a statistical dictionary trained using the word-alignment model IBM M1 (Och and Ney, 2003) on the JRC-Acquis (Steinberger et al., 2006) corpus. Similar performance for Spanish-English is obtained using BabelNet as statistical dictionary (Franco-Salvador et al., 2012), but not for German-English.

[35] CL-C3G is CL-CNG using character 3-grams, as recommended in Potthast et al. (2011a).

possible translations of each word. A BOW representation was obtained for each text fragment.[36] Text fragments were compared using the Dice's coefficient; (ii) *POS + statDict*, same as statDict but using TreeTagger to POS tag and lemmatize words before translation; and (iii) *POS + statDict + MFS*, which additionally used the Most Frequent Sense (MFS) baseline[37] to disambiguate the words before generating the BOW. In Table 3.2 we can find a summary of all the models included in the evaluation.

The experiments were divided into three subsections: (i) in Section 3.5.3 we used the PAN-PC-10 dataset to perform the comparison and tuning of the CL-KGA weighting schemes of semantic relations; (ii) in Section 3.5.4 we compared the different variants of CL-KGA and studied the characteristics of the model using the PAN-PC-11 dataset; and (iii) in Section 3.5.5 we compared our model with the state of the art, evaluating the performance when detecting the CL plagiarism cases of the PAN-PC-11 dataset. In this last section we also studied the performance on exclusively the CL cases with paraphrasing, and compared the computational efficiency of the models.

3.5.3 Evaluation of CL-KGA Weighting Schemes for Semantic Relations

In this section we compared the classical graph weighting for semantic relations based on Dice's coefficient (cf. Section 3.3.3.1) and the new method using distributed representations of concepts (cf. Section 3.3.3.2). We used these experiments to optimize also the parameters of the CL-KGA model.[38] For these experiments we used the Spanish-English and German-English partitions of PAN-PC-10 and measured the overall score of plagiarism detection, i.e., plagdet.

First, for each weighting scheme, we determined the threshold to consider that the concepts of the knowledge graphs are semantically related (cf. Section 3.3.2). Next, we selected the values of relevance for concepts and relations used with CL-KGA (cf. Section 3.4) for both weightings.

[36] By generating a BOW with all possible translations, we attempted to counterbalance possible errors introduced when using a statistical dictionary for translating.

[37] Basically, for each word it provides the first sense suggested by WordNet, which represents the most frequent use of that word.

[38] Since all the CL-KGA variants share the same basic structure and graphs, we used the same parameters for all of them.

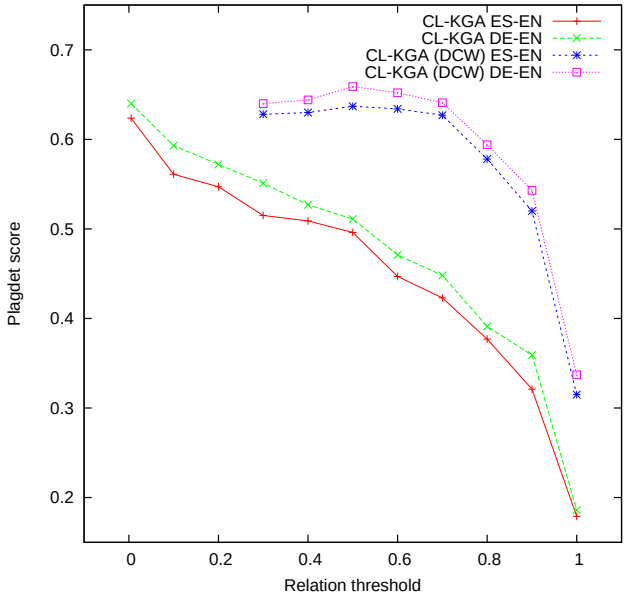

Figure 3.7. Plagdet score in PAN-PC-10 dataset in function of the threshold between relations.

To determine the threshold of the semantic relations, we tested values between 0.001 and 1.[39] In Figure 3.7 we can see the results of the experiments. For the model using the classical weighting, we obtained the best results with the minimum threshold: 0.001. Similar results were obtained using 0.005 as in previous works. In this case, because of the low number of weighted edges, augmenting the threshold considerably reduced the connectivity of the graphs and, consequently, the plagdet. In contrast, the CL-KGA model using DCW had 0.5 as optimal value in both language pairs, with close results using values between 0.3 and 0.7. The DCW scheme was less sensitive to the threshold value, probably because the higher number of relations contained in the graphs, and remained stable with a strong decreasing for high thresholds. We assume that the key concepts of the graphs were present and connected until those values were higher than 0.8. In contrast

[39] We start at 0.001 because a value of zero would suppose using all the relations of BabelNet and would generate too much dense and noisy graphs. For the DCW weighting we started using 0.3 as threshold because lower values were computationally very expensive. In this experiment, we set the values of relevance for concepts and relations to 50-50%.

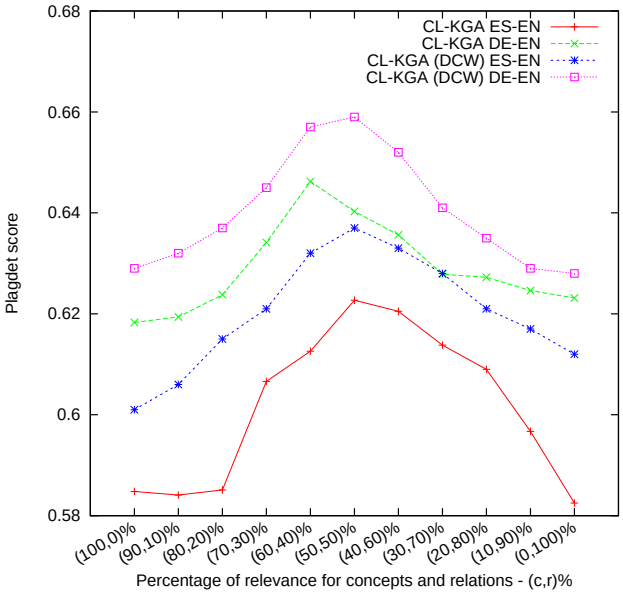

Figure 3.8. Plagdet score in the PAN-PC-10 dataset as function of percentage of relevance of concepts and relations.

to the results shown in the next section using PAN-PC-11, the PAN-PC-10 dataset provided better results on the German-English partition.

To select the values of relevance of concepts and relations, we modified the percentage of relevance between 0% and 100% for both parameters. Figure 3.8 shows the results of these experiments. We observed a similar trend using both weighting schemes. The best values were obtained for equal relevance for concepts and relations, with similar values for the close percentages, excluding German-English with the classical weighting, which obtained the best values using a 60-40% distribution. These results show that CL-KGA benefits both from the weight of the concepts and the relations to detect CL plagiarism. Note that our DCW scheme obtained better performance on each language pair in all the tested configurations. The use of distributed representations to model concepts benefited our model with a more accurate and human interpretable[40] semantic relation weights.

[40]By "human interpretable" we refer to the values of the weights, that have in 50% the optimal value to consider that a relation is semantically related.

System	Plagdet	Recall	Precision	Granularity
CL-KGA (BabelNet 1.0)	0.594	0.518	0.705	1.008
CL-KGA	0.619	0.558	0.699	1.000
CL-KGA (DCW)	**0.651**	0.574	0.752	1.000
CL-KGA (WSD path filter)	0.598	0.521	0.707	1.005
CL-KGA (WSD concepts)	0.464	0.408	0.655	1.119
CL-KGA (WSD concepts w/o weighting)	**0.646**	0.571	0.744	1.000
CL-KGA (DCW) (WSD concepts w/o weighting)	**0.663**	0.588	0.761	1.000

Table 3.3. Results of PAN-PC-11 Spanish-English partition using the CL-KGA variants.

Finally, we highlight also the difference in size (number of concepts) of the knowledge graphs using the classical or the DCW schemes. Using the optimal parameters determined in this section, a graph using the first weighting had on average 1,384 concepts. In contrast, using DCW graphs were much dense, containing on average 17,495 concepts. This was produced for the high number of weighted edges available when using DCW and may be reduced using a higher relation threshold if the computational speed is a priority.

We used also PAN-PC-10 to tune the threshold employed by CL-ESA to make zero the low similarity scores of a text with a Wikipedia page. The best results were obtained with 0.01. In the next section, we used the best values obtained here for each language pair and model.

3.5.4 Evaluation of the CL-KGA Variants and Characteristics

In this section we used the Spanish-English and German-English partitions of the PAN-PC-11 to compare the proposed variants (cf. Section 3.3.3.1, 3.3.3.2 and 3.3.4.1) of the CL-KGA model and study the characteristics of our approach.

In Table 3.3 we show the results for Spanish-English. The new experiments with the CL-KGA variants achieved interesting results. Despite using the same weighting, CL-KGA improved the results obtained using BabelNet 1.0. This difference is due to the new relations between concepts, and the new lexicalizations for WordNet verbs, adjectives, and adverbs in Spanish inside BabelNet 2.5, which were only in English in the previous experiments (Franco-Salvador et al., 2013a). Similarly to the results with PAN-

System	Plagdet	Recall	Precision	Granularity
CL-KGA (BabelNet 1.0)	0.514	0.443	0.631	1.017
CL-KGA	0.520	0.460	0.601	1.003
CL-KGA (DCW)	0.564	0.495	0.65	1.000
CL-KGA (WSD path filter)	0.508	0.434	0.644	1.028
CL-KGA (WSD concepts)	0.324	0.276	0.531	1.174
CL-KGA (WSD concepts w/o weighting)	**0.586**	0.508	0.692	1.000
CL-KGA (DCW) (WSD concepts w/o weighting)	**0.595**	0.516	0.703	1.000

Table 3.4. Results of PAN-11 German-English partition using the CL-KGA variants.

PC-10 of Section 3.5.3, CL-KGA with the new weighting scheme based on distributed representations of concepts, CL-KGA (DCW), obtained higher results with a significant difference,[41] and highlights the quality of the new relation weights for computing semantic relatedness. Despite theoretically providing with cleaner graphs, the version with WSD path filter was not able to improve the results of CL-KGA although its results were close. This difference may be due to the wrong disambiguations and intermediate concepts between them that we are keeping. Note that the use of knowledge graphs to perform WSD offers an accuracy close to 70% (Navigli and Ponzetto, 2012a). The CL-KGA (WSD concepts), which keeps the WSD concepts and removes the vocabulary expansion, reduced considerably the performance. We observed that the problem was due to the weighting of the concepts, which was estimated as a function of the outdegree of the complete graph. The current variant, exclusively weighting the WSD concepts, offered too sparse and unbounded values, which made it more difficult to be successfully compared using Dice's coefficient (cf. Section 3.3.2 and 3.4). We repeated the experiments without weights for the conceptual similarity measure. That model, CL-KGA (WSD concepts w/o weighting), obtained the best results with the two weighting schemes for knowledge graphs. It seems that the use of knowledge graphs to perform a multilingual WSD produced a specially precise representation of the text fragments. If we analyse the need of vocabulary expansion in knowledge graphs (cf. Section 3.3.4.2), we note that this WSD exploits the expanded concepts to determine the disambiguations. Therefore, although not using expanded concepts directly in the representation as CL-KGA, the vocabulary expansion is crucial for our model.

[41] In this work, statistically significant results of plagdet according to a χ^2 test ($p < 0.05$) were highlighted in bold.

	System	Plagdet	Recall	Precision	Granularity
(a)	CL-KGA (BabelNet 1.0)	0.594	0.518	0.705	1.008
	CL-ASA	0.517	0.448	0.689	1.070
	CL-ESA	0.471	0.448	0.534	1.048
	CL-C3G	0.170	0.127	0.616	1.372
(b)	statDict	0.613	0.548	0.696	1.000
	POS + statDict	0.632	0.558	0.730	1.000
	POS + statDict + MFS	0.632	0.560	0.728	1.001
(c)	CL-KGA	0.619	0.558	0.699	1.000
	CL-KGA (DCW)	**0.651**	0.574	0.752	1.000
	CL-KGA (WSD path filter)	0.598	0.521	0.707	1.005
	CL-KGA (WSD concepts)	0.464	0.408	0.655	1.119
	CL-KGA (WSD concepts w/o weighting)	**0.646**	0.571	0.744	1.000
	CL-KGA (DCW) (WSD concepts w/o weighting)	**0.663**	0.588	0.761	1.000

Table 3.5. Results of PAN-PC-11 Spanish-English partition: (a) state-of-the-art approaches; (b) baselines; (c) proposed approaches.

The results for German-English were a similar. In Table 3.4 we can observe the overall performance. Note that the best weighting scheme was the DCW, and the best results were again with the WSD concepts w/o weighting variant, which highlights the relevance of WSD in our model.

3.5.5 COMPARISON WITH THE STATE-OF-THE-ART

In this section we compare CL-KGA and its variants with several state-of-the-art approaches and baselines (see Table 3.2) using the PAN-PC-11 dataset for CL plagiarism detection.

In Table 3.5 we show the results obtained for Spanish-English. The lowest results were obtained by CL-C3G. This is unsurprising if we consider that Spanish and English do not share many lexical and syntactic similarities — indispensable requirement for a high character n-gram overlap. The second worst results were obtained by CL-ESA. The CL-ASA model obtained a similar recall but with higher precision, resulting in a superior plagdet. It seems that CL-ESA, based on similarities with a document collection, gave a higher number of false positives. In fact, ESA was originally meant for tasks of relatedness rather than plagiarism. The CL-KGA results obtained previously using BabelNet 1.0 were the next in the ranking. Because of the knowledge graphs, CL-KGA was able to model the text in a more precise manner and provided better results in all measures. Note that the best pos-

	System	Plagdet	Recall	Precision	Granularity
(a)	CL-KGA (BabelNet 1.0)	0.514	0.443	0.631	1.017
	CL-ASA	0.405	0.343	0.603	1.113
	CL-ESA	0.336	0.293	0.466	1.101
	CL-C3G	0.077	0.047	0.330	1.089
(b)	statDict	0.553	0.469	0.683	1.007
	POS + statDict	0.328	0.253	0.685	1.182
	POS + statDict + MFS	0.347	0.271	0.687	1.175
(c)	CL-KGA	0.520	0.460	0.601	1.003
	CL-KGA (DCW)	0.564	0.495	0.653	1.000
	CL-KGA (WSD path filter)	0.508	0.434	0.644	1.028
	CL-KGA (WSD concepts)	0.324	0.276	0.531	1.174
	CL-KGA (WSD concepts w/o weighting)	**0.586**	0.508	0.692	1.000
	CL-KGA (DCW) (WSD concepts w/o weighting)	**0.595**	0.516	0.703	1.000

Table 3.6. Results of PAN-PC-11 German-English partition: (a) State-of-the-art approaches; (b) baselines; (c) proposed approaches.

sible value of granularity is 1.0. However, the proposed baselines offered higher performance. Despite the simplicity of statDict, even the basic variant — with higher results if we POS tag and lemmatize —, obtained a very competitive performance. The disambiguation step using MFS improved the results although without significant differences. The use of a statistical dictionary to generate a BOW containing all the translations with equal relevance, provided a simple but solid model against wrong translations. The results with the CL-KGA variants provided significant differences and superior performance for the standard version with the proposed DCW scheme, and even higher results for the CL-KGA (WSD concepts w/o weighting) variant. We can observe notable differences — specially with German-English — compared to the other approach using WSD: POS + statDict + MFS. This highlights the quality of the disambiguations using knowledge graphs. Note also the differences in performance between the two models using a multilingual collection of concepts: CL-ESA and CL-KGA. These differences were due to the characteristics of the models, which were studied in Section 3.3.4.4: aimed at adjusting to the text words, our model has a variable concept inventory. In addition, CL-KGA uses relatedness between concepts and vocabulary expansion.

The differences between the models for German-English were similar but with an overall and small performance reduction. In Table 3.6 we can see the results. There are some interesting aspects to highlight. CL-C3G

	System	Plagdet	Recall	Precision	Granularity
(a)	CL-KGA (BabelNet 1.0)	0.099	0.197	0.066	1.000
	CL-ASA	0.061	0.150	0.038	1.000
	CL-ESA	0.038	0.159	0.021	1.000
	CL-C3G	0.028	0.058	0.019	1.000
(b)	statDict	0.085	0.179	0.050	1.000
	POS + statDict	0.135	0.236	0.732	1.000
	POS + statDict + MFS	0.121	0.207	0.086	1.000
(c)	CL-KGA	0.118	0.244	0.078	1.000
	CL-KGA (DCW)	**0.163**	0.261	0.119	1.000
	CL-KGA (WSD path filter)	0.102	0.223	0.066	1.000
	CL-KGA (WSD concepts)	0.052	0.126	0.033	1.000
	CL-KGA (WSD concepts w/o weighting)	0.149	0.258	0.104	1.000
	CL-KGA (DCW) (WSD concepts w/o weighting)	**0.167**	0.264	0.122	1.000

Table 3.7. Results of PAN-PC-11 Spanish-English partition, **evaluating only paraphrasing cases**: (a) State-of-the-art approaches; (b) baselines; (c) proposed approaches.

obtained even lower results than for Spanish-English. Although having the same linguistic roots, these two Germanic languages do not share enough lexical and syntactic similarities to model the content properly using character n-grams. On the other hand, the variants of statDict using POS tagging and lemmatization did not excelled as in Spanish-English. The use of the TreeTagger tool introduced errors, which reduced the quality of the representations. Note that the best results were with CL-KGA using our DCW scheme and the WSD concepts w/o weighting variant. This proves that CL-KGA is a competitive model for Spanish-English and German-English CL plagiarism detection.

3.5.5.1 Detecting Cross-language Plagiarism Detection with Paraphrasing

As we mentioned in Section 3.5.1, the PAN-PC-11 dataset contains cases of CL paraphrasing. This type of plagiarism is more difficult to detect because its text has been modified in order to hide the plagiarism action. We were interested in observing the differences of the models when trying to detect only those paraphrasing cases. We performed an additional experiment to consider only paraphrasing cases as instances of plagiarism in the corpus. In Tables 3.7 and 3.8 we can see the results. The differences in the performance of all the models compared to the results obtained previously using the complete dataset were substantial. We observed that most of these paraphrasing cases were very short in length, and probably the use of Algo-

	System	Plagdet	Recall	Precision	Granularity
(a)	CL-KGA (BabelNet 1.0)	0.100	0.210	0.066	1.000
	CL-ASA	0.046	0.097	0.030	1.000
	CL-ESA	0.035	0.117	0.021	1.000
	CL-C3G	0.018	0.038	0.012	1.000
(b)	statDict	0.109	0.187	0.076	1.000
	POS + statDict	0.064	0.113	0.044	1.000
	POS + statDict + MFS	0.066	0.117	0.046	1.000
(c)	CL-KGA	0.093	0.226	0.058	1.000
	CL-KGA (DCW)	**0.161**	0.259	0.117	1.000
	CL-KGA (WSD path filter)	0.100	0.201	0.067	1.000
	CL-KGA (WSD concepts)	0.041	0.113	0.025	1.000
	CL-KGA (WSD concepts w/o weighting)	**0.165**	0.264	0.120	1.000
	CL-KGA (DCW) (WSD concepts w/o weighting)	**0.171**	0.269	0.125	1.000

Table 3.8. Results of PAN-PC-11 German-English partition, **evaluating only paraphrasing cases**: (a) State-of-the-art approaches; (b) baselines; (c) proposed approaches.

System	Text indexing (texts/second)	Text similarity (texts/second)
CL-ASA	1,741	3,627
CL-ESA	282	1,826
CL-C3G	3,547	2,761
statDict	2,492	2,593
CL-KGA	11	1,259
CL-KGA (DCW)	3	281
CL-KGA (WSD concepts w/o weighting)	9	5,685
CL-KGA (DCW) (WSD concepts w/o weighting)	3	5,827

Table 3.9. Comparison of time required to index and compare texts. Results are estimated as the average for processing all the Spanish-English partition.

rithm 3.1, designed for longer cases, was the reason of this global quality reduction. However, we can still appreciate that the differences among the results of the models were similar at a smaller scale. CL-KGA obtained the higher performance using DCW for the relations of the knowledge graphs. In this experiments we did not observe such substantial differences between CL-KGA (DCW) and CL-KGA (DCW) (WSD concepts w/o weighting), although may be still appreciated for German-English.

3.5.5.2 Evaluation of the Computational Efficiency

In order to select a model for CL plagiarism detection, its computational efficiency is a key aspect. The purpose and requirements of the system may require a fast or an accurate model. In Table 3.9 we measured the number of text fragments indexed and compared per second for each evaluated model using the complete Spanish-English partition. These experiments were performed using a Intel-i5@2.8Ghz with 16 GB of RAM. As we can see, CL-KGA required considerably more time to index (or generate the graphs of) text. This is due to the use of the BabelNet multilingual semantic network. The 9,348,287 synsets and the $\sim$262 relations among them made the graph generation a computationally expensive task. In addition, the use of DCW made the graphs more dense and, consequently, they required more time to be compared in the similarity step. Text indexing is usually part of the preprocessing step, being the indexing of the new documents needed only once. The text similarity step is the most important, and the two weighting schemes using WSD concepts w/o weighting may be a solution. These were the fastest models in calculating similarity because they only contain a BOW of disambiguated words. In contrast, if the speed of indexing is crucial, statDict offered a balance between performance and efficiency. Note that in order to speed up graph indexing, parallel computing can be used, as we did for our experiments.

3.6 Conclusions

In this paper we performed a systematic study of Cross-Language Knowledge Graph Analysis, an approach that represents fragments of text using knowledge graphs as a language independent model of its content. We studied the impact of relevant aspects of the model for the task of cross-language plagiarism detection: word sense disambiguation, vocabulary expansion, language independence and representation by similarities with a collection of concepts. Experimental results showed that WSD is the essential component of the model, being only necessary the use of vocabulary expansion during the WSD processing. The differences between CL-ESA and CL-KGA — the two models that exploit Wikipedia as multilingual collection of concepts — favour the latter model, which thanks to the high coverage of BabelNet, the vocabulary expansion and the concept relatedness employed, offered a higher performance. In addition, we proposed a new weighting scheme of

relations between concepts based on the use of distributed representations of concepts. The use of this weighting provided our model with state-of-the-art performance on the Spanish-English and German-English partitions of the PAN-PC-11 dataset. The study of the model with cross-language paraphrasing cases proved also its superiority. However, a comparison of the computational efficiency of the models showed that our model is more adequate when a fast document similarity is required and the indexing is performed in a preprocessing step. In other situations, statDict — also introduced in this paper — is the recommended solution due to its fast indexing and similarity calculation, in addition to its high performance.

For future work we will continue exploring the use of knowledge graphs and multilingual semantic networks for cross-language similarity tasks. The use of semantic signatures allows to create a new type of knowledge graphs which have been successfully used for multilingual WSD (Moro et al., 2014), and will be studied in the future. The use of distributed representations will also be investigated further. The generation of distributed representations of concepts is only in its infancy, and works like SensEmbed, the study of Aletras and Stevenson (2015), or this paper, could be extended for tasks such as similarity analysis, conceptual relatedness or WSD.

Acknowledgements

This research has been carried out in the framework of the European Commission WIQ-EI IRSES (no. 269180) and DIANA-APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) projects. We would like to thank Tomas Mikolov, Martin Potthast, and Luis A. Leiva for their support and comments during this research.

4 A Knowledge-based Representation for Cross-language Document Retrieval and Categorization

Published in:

- **Franco-Salvador, M.**, Rosso, P., and Navigli, R. (2014b). A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 414–423. Association for Computational Linguistics. **(CORE A)**

This chapter of the thesis presents the knowledge-based document similarity model. It is a modified version of our CL-KGA model, that complements it with a vector-based representation in order to cover graph shortcomings such as out-of-vocabulary words and verbal tenses. We evaluate and compare it with the state of the art in the tasks of cross-language document retrieval and categorization.

Abstract

Current approaches to cross-language document retrieval and categorization are based on discriminative methods which represent documents in a low-dimensional vector space. In this paper we propose a shift from the supervised to the knowledge-based paradigm and provide a document similarity measure which draws on BabelNet, a large multilingual knowledge resource. Our experiments show state-of-the-art results in cross-lingual document retrieval and categorization.

4.1 Introduction

The huge amount of text that is available online is becoming ever increasingly multilingual, providing an additional wealth of useful information. Most of this information, however, is not easily accessible to the majority of users because of language barriers which hamper the cross-lingual search and retrieval of knowledge.

Today's search engines would benefit greatly from effective techniques for the cross-lingual retrieval of valuable information that can satisfy a user's needs by not only providing (Landauer and Littman, 1994) and translating (Munteanu and Marcu, 2005) relevant results into different languages, but also by reranking the results in a language of interest on the basis of the importance of search results in other languages.

Vector-based models are typically used in the literature for representing documents both in monolingual and cross-lingual settings (Manning et al., 2008). However, because of the large size of the vocabulary, having each term as a component of the vector makes the document representation very sparse. To address this issue several approaches to dimensionality reduction have been proposed, such as Principal Component Analysis (Jolliffe, 1986), Latent Semantic Indexing (Hull, 1994), Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and variants thereof, which project these vectors into a lower-dimensional vector space. In order to enable multilinguality, the vectors of comparable documents written in different languages are concatenated, making up the document matrix which is then reduced using linear projection (Platt et al., 2010; Yih et al., 2011). However, to do so, comparable documents are needed as training. Additionally, the lower dimensional representations are not of easy interpretation.

The availability of wide-coverage lexical knowledge resources extracted automatically from Wikipedia, such as DBPedia (Bizer et al., 2009), YAGO (Hoffart et al., 2013) and BabelNet (Navigli and Ponzetto, 2012a), has considerably boosted research in several areas, especially where multilinguality is a concern (Hovy et al., 2013). Among these latter are cross-language plagiarism detection (Franco-Salvador et al., 2013a; Potthast et al., 2011a), multilingual semantic relatedness (Nastase and Strube, 2013; Navigli and Ponzetto, 2012b) and semantic alignment (Matuschek and Gurevych, 2013; Navigli and Ponzetto, 2012a). One main advantage of knowledge-based methods is that they provide a human-readable, semantically interconnected, representation of the textual item at hand (be it a sentence or a document).

Following this trend, in this paper we provide a knowledge-based representation of documents which goes beyond the lexical surface of text, while at the same time avoiding the need for training in a cross-language setting. To achieve this we leverage a multilingual semantic network, i.e., BabelNet, to obtain language-independent representations, which contain concepts together with semantic relations between them, and also include semantic knowledge which is just implied by the input text. The integration of our multilingual graph model with a vector representation enables us to obtain state-of-the-art results in comparable document retrieval and cross-language text categorization.

4.2 Related Work

The mainstream representation of documents for monolingual and cross-lingual document retrieval is vector-based. A document vector, whose components quantify the relevance of each term in the document, is usually highly dimensional, because of the variety of terms used in a document collection. As a consequence, the resulting document matrices are very sparse. To address the data sparsity issue, several approaches to the reduction of dimensionality of document vectors have been proposed in the literature. A popular class of methods is based on linear projection, which provides a low-dimensional mapping from a high dimensional vector space. A historical approach to linear projection is Principal Component Analysis (PCA) (Jolliffe, 1986), which performs a singular value decomposition (SVD) on a document matrix D of size $n \times m$, where each row in D is the term vector representation of a document. PCA uses an orthogonal transformation to convert a

set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components, which make up the low-dimensional vector. Latent Semantic Analysis (LSA) (Deerwester et al., 1990) is very similar to PCA but performs the SVD using the correlation matrix instead of the covariance matrix, which implies a lower computational cost. LSA preserves the amount of variance in an eigenvector $\vec{v}$ by maximizing its Rayleigh ratio: $\frac{\vec{v}^T C \vec{v}}{\vec{v}^T \vec{v}}$, where $C = D^T D$ is the correlation matrix of D.

A generalization of PCA, called Oriented Principal Component Analysis (OPCA) (Diamantaras and Kung, 1996), is based on a noise covariance matrix to project the similar components of D closely. Other projection models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) are based on the extraction of generative models from documents. Another approach, named Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007), represents each document by its similarities to a document collection. Using a low domain specificity document collection such as Wikipedia, the model has proven to obtain competitive results.

Not only have these methods proven to be successful in a monolingual scenario (Deerwester et al., 1990; Hull, 1994), but they have also been adapted to perform well in tasks at a cross-language level (Platt et al., 2010; Potthast et al., 2008; Yih et al., 2011). Cross-language Latent Semantic Indexing (CL-LSI) (Dumais et al., 1997b) was the first linear projection approach used in cross-lingual tasks. CL-LSI provides a cross-lingual representation for documents by reducing the dimensionality of a matrix D whose rows are obtained by concatenating comparable documents from different languages. Similarly, PCA and OPCA can be adapted to a multilingual setting. LDA was also adapted to perform in a multilingual scenario with models such as Polylingual Topic Models (Mimno et al., 2009), Joint Probabilistic LSA and Coupled Probabilistic LSA (Platt et al., 2010), which, however, are constrained to using word counts, instead of better weighting strategies, such as log(tf)-idf, known to perform better with large vocabularies (Salton and McGill, 1986). Another variant, named Canonical Correlation Analysis (CCA) (Thompson, 2005), uses a cross-covariance matrix of the low-dimensional vectors to find the projections. Cross-language Explicit Semantic Analysis (CL-ESA) (Cimiano et al., 2009; Potthast et al., 2011a, 2008), instead, adapts ESA to be used at cross-language level by exploiting the comparable documents across languages from Wikipedia. CL-ESA represents

each document written in a language L by its similarities with a document collection in the same language L. Using a multilingual document collection with comparable documents across languages, the resulting vectors from different languages can be compared directly.

An alternative unsupervised approach, Cross-language Character n-Grams (CL-CNG) (Mcnamee and Mayfield, 2004), does not draw upon linear projections and represents documents as vectors of character n-grams. It has proven to obtain good results in cross-language document retrieval (Potthast et al., 2011a) between languages with lexical and syntactic similarities.

Recently, a novel supervised linear projection model based on Siamese Neural Networks (S2Net) (Yih et al., 2011) achieved state-of-the-art performance in comparable document retrieval. S2Net performs a linear combination of the terms of a document vector $\vec{d}$ to obtain a reduced vector $\vec{r}$, which is the output layer of a neural network. Each element in $\vec{r}$ has a weight which is a linear combination of the original weights of $\vec{d}$, and captures relationships between the original terms.

However, linear projection approaches need a high number of training documents to achieve state-of-the-art performance (Platt et al., 2010; Yih et al., 2011). Moreover, although they are good at identifying a few principal components, the representations produced are opaque, in that they cannot explicitly model the semantic content of documents with a human-interpretable representation, thereby making the data analysis difficult. In this paper, instead, we propose a language-independent knowledge graph representation for documents which is obtained from a large multilingual semantic network, without using any training information. Our knowledge graph representation explicitly models the semantics of the document in terms of the concepts and relations evoked by its co-occurring terms.

4.3 A Knowledge-based Document Representation

We propose a knowledge-based document representation aimed at expanding the terms in a document's bag of words by means of a knowledge graph which provides concepts and semantic relations between them. Key to our approach is the use of a graph representation which does not depend on any given language, but, indeed, is multilingual. To build knowledge graphs of this kind we utilize BabelNet, a multilingual semantic network that we

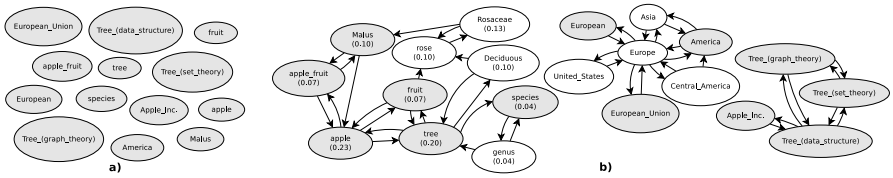

Figure 4.1. (a) initial graph from $T_K = \{$"European", "apple", "tree", "Malus", "species", "America"$\}$; (b) knowledge graph obtained by retrieving all paths from BabelNet. Gray nodes are the original concepts.

present in Section 4.3.1. Then, in Section 4.3.2, we describe the five steps needed to obtain our graph-based multilingual representation of documents. Finally, we introduce our knowledge graph similarity measure in Section 4.3.3.

4.3.1 BABELNET

BabelNet (Navigli and Ponzetto, 2012a) is a multilingual semantic network whose concepts and relations are obtained from the largest available semantic lexicon of English, WordNet (Fellbaum, 1998), and the largest wide-coverage collaboratively-edited encyclopedia, Wikipedia, by means of an automatic mapping algorithm. BabelNet is therefore a multilingual "encyclopedic dictionary" that combines lexicographic information with wide-coverage encyclopedic knowledge. Concepts in BabelNet are represented similarly to WordNet, i.e., by grouping sets of synonyms in the different languages into multilingual synsets. Multilingual synsets contain lexicalizations from WordNet synsets, the corresponding Wikipedia pages and additional translations output by a statistical machine translation system. The relations between synsets are collected from WordNet and from Wikipedia's hyperlinks between pages.

We note that, in principle, we could use any multilingual network providing a similar kind of information, e.g., EuroWordNet (Vossen, 2004). However, in our work we chose BabelNet because of its larger size, its coverage of both lexicographic and encyclopedic knowledge, and its free availability.[1] In our work we used BabelNet 1.0, which encodes knowledge for six languages, namely: Catalan, English, French, German, Italian and Spanish.

[1] `http://babelnet.org`

4.3.2 FROM DOCUMENT TO KNOWLEDGE GRAPH

We now introduce our five-step method for representing a given document d from a collection D of documents written in language L as a language-independent knowledge graph.

Building a Basic Vector Representation Initially we transform a document d into a traditional vector representation. To do this, we score each term $t_i \in d$ with a weight w_i. This weight is usually a function of term and document frequency. Following the literature, one method that works well is the log tf-idf weighting (Salton et al., 1983; Salton and McGill, 1986):

$$w_i = log_2(f_i + 1)log_2(n/n_i). \tag{4.1}$$

where f_i is the number of times term i occurs in document d, n is the total number of documents in the collection and n_i is the number of documents that contain t_i. We then create a weighted term vector $\vec{v} = (w_1, ..., w_n)$, where w_i is the weight corresponding to term t_i. We exclude stopwords from the vector.

Selecting the Relevant Document Terms We then create the set T of base forms, i.e., lemmas[2], of the terms in the document d. In order to keep only the most relevant terms, we sort the terms T according to their weight in vector $\vec{v}$ and retain a maximum number of K terms, obtaining a set of terms T_K.[3] The value of K is calculated as a function of the vector size, as follows:

$$K = (log_2(1 + |\vec{v}|))^2, \tag{4.2}$$

The rationale is that K must be high enough to ensure a good conceptual representation but not too high, so as to avoid as much noise as possible in the set T_K.

Populating the Graph with Initial Concepts Next, we create an initially-empty knowledge graph $G = (V, E)$, i.e., such that $V = E = \emptyset$.

[2]Following the setup of (Platt et al., 2010), our initial data is represented using term vectors. For this reason we lemmatize in this step.

[3]Since the vector $\vec{v}$ provides weights for all the word forms, and not only lemmas, occurring in d, we take the best weight among those word forms of the considered lemma.

We populate the vertex set V with the set S_K of all the synsets in BabelNet which contain any term in T_K in the document language L, that is:

$$S_K = \bigcup_{t \in T_K} Synsets_L(t), \tag{4.3}$$

where $Synsets_L(t)$ is the set of synsets in BabelNet which contain a term t in the language of interest L. For example, in Figure 4.1(a) we show the initial graph obtained from the set $T_K = \{$"European", "apple", "tree", "Malus", "species", "America"$\}$. Note, however, that each retrieved synset is multilingual, i.e., it contains lexicalizations for the same concept in other languages too. Therefore, the nodes of our knowledge graph provide a language-independent representation of the document's content.

Creating the Knowledge Graph Similarly to Navigli and Lapata (2010), we create the knowledge graph by searching BabelNet for paths connecting pairs of synsets in V. Formally, for each pair $v, v' \in V$ such that v and v' do not share any lexicalization[4] in T_K, for each path in BabelNet $v \rightarrow v_1 \rightarrow \cdots \rightarrow v_n \rightarrow v'$, we set: $V := V \cup \{v_1, \ldots, v_n\}$ and $E := E \cup \{(v, v_1), \ldots, (v_n, v')\}$, that is, we add all the path vertices and edges to G. After prototyping, the path length is limited to maximum length 3, so as to avoid an excessive semantic drift.

As a result of populating the graph with intermediate edges and vertices, we obtain a knowledge graph which models the semantic context of document d. We point out that our knowledge graph might have different isolated components. We view each component as a different interpretation of document d. To select the main interpretation, we keep only the largest component, i.e., the one with the highest number of vertices, which we consider as the most likely semantic representation of the document content.

Figure 4.1(b) shows the knowledge graph obtained for our example term set. Note that our approach retains, and therefore weights, only the subgraph focused on the "apple fruit" meaning.

[4]This prevents different senses of the same term from being connected via a path in the resulting knowledge graph.

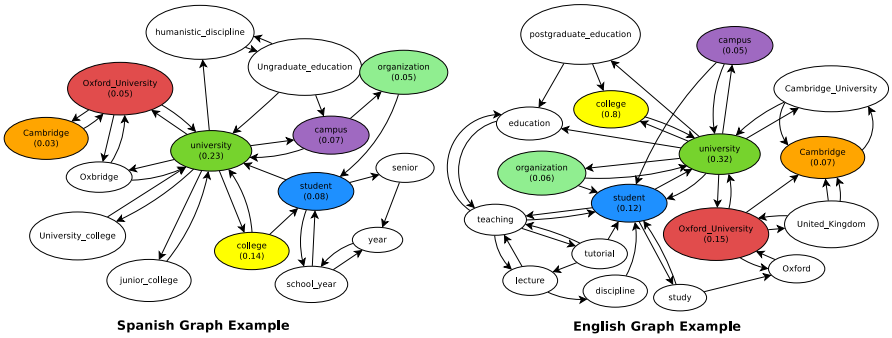

Figure 4.2. Knowledge graph examples from two comparable documents in different languages.

Knowledge Graph Weighting The final step consists of weighting all the concepts and semantic relations of the knowledge graph G. For weighting relations we use the original weights from BabelNet, which provide the degree of relatedness between the synset end points of each edge (Navigli and Ponzetto, 2012a). As for concepts, we weight them on the basis of the original weights of the terms in the vector $\vec{v}$. In order to score each concept in our knowledge graph G, we applied the topic-sensitive PageRank algorithm (Haveliwala et al., 2003) to G. While the well-known PageRank algorithm (Page et al., 1998) calculates the global importance of vertices in a graph, topic-sensitive PageRank is a variant in which the importance of vertices is biased using a set of representative "topics". Formally, the topic-sensitive PageRank vector $\vec{p}$ is calculated by means of an iterative process until convergence as follows: $\vec{p} = cM\vec{p} + (1-c)\vec{u}$, where c is the damping factor (conventionally set to 0.85), $1-c$ represents the probability of a surfer randomly jumping to any node in the graph, M is the transition probability matrix of graph G, with $M_{ji} = degree(i)^{-1}$ if an edge from i to j exists, 0 otherwise, $\vec{u}$ is the random-jumping transition probability vector, where each u_i represents the probability of jumping randomly to the node i, and $\vec{p}$ is the resulting PageRank vector which scores the nodes of G. In contrast to vanilla PageRank, the "topic-sensitive" variant gives more probability mass to some nodes in G and less to others. In our case we perturbate $\vec{u}$ by concentrating the probability mass to the vertices in S_K, which are the synsets corresponding to the document terms T_K (cf. Formula 4.3).

4.3.3 SIMILARITY BETWEEN KNOWLEDGE GRAPHS

We can now determine the similarity between two documents $d, d' \in D$ in terms of the similarity of their knowledge graph representations G and G'.

Following the literature (Montes y Gómez et al., 2001) we calculate the similarity between the vertex sets in the two graphs using Dice's coefficient (Jackson et al., 1989):

$$S_c(G, G') = \frac{2 \cdot \sum_{c \in V(G) \cap V(G')} w(c)}{\sum_{c \in V(G)} w(c) + \sum_{c \in V(G')} w(c)}, \tag{4.4}$$

where $w(c)$ is the weight of a concept c (see Section 4.3.2). Likewise, we calculate the similarity between the two edge sets as:

$$S_r(G, G') = \frac{2 \cdot \sum_{r \in E(G) \cap E(G')} w(r)}{\sum_{r \in E(G)} w(r) + \sum_{r \in E(G')} w(r)}, \tag{4.5}$$

where $w(r)$ is the weight of a semantic relation edge r.

We combine the two above measures of conceptual (S_c) and relational (S_r) similarity to obtain an integrated measure $S_g(G, G')$ between knowledge graphs:

$$S_g(G, G') = \frac{S_c(G, G') + S_r(G, G')}{2}. \tag{4.6}$$

Notably, since we are working with a language-independent representation of documents, this similarity measure can be applied to the knowledge graphs built from documents written in any language. In Figure 4.2 we show two knowledge graphs for comparable documents written in different languages (for clarity, labels are in English in both graphs). As expected, the graphs share several key concepts and relations.

Algorithm 4.1 Dictionary-based term-vector translation.

Input: a weighted document vector $\vec{v}_L = (w_1, \ldots, w_n)$, a source language L and a target language L'
Output: a translated vector $\vec{v}_{L'}$

1: $\vec{v}_{L'} \leftarrow (0, \ldots, 0)$ of length n
2: **for** $i = 1$ to n **do**
3: **if** $w_i = 0$ **continue**
4: // let t_i be the term corresponding to w_i in $\vec{v}_L$
5: $S_L \leftarrow \text{Synsets}_L(t_i)$
6: **for each** synset $s \in S_L$ **do**
7: $T \leftarrow \text{getTranslations}(s, L')$
8: **if** $T \neq \emptyset$ **then**
9: **for each** $tr \in T$ **do**
10: $w_{new} = w_i \cdot \text{confidence}(tr, t_i)$
11: // let index(tr) be the index of tr in $\vec{v}_L$
12: **if** $\exists$ index(tr) **then**
13: $v_{L'}(\text{index}(tr)) = w_{new}$
14: **return** $\vec{v}_{L'}$

4.4 A Multilingual Vector Representation

4.4.1 From Document to Multilingual Vector

Since our knowledge graphs will only cover the most central concepts of a document, we complement this core representation with a more traditional vector-based representation. However, as we are interested in the cross-language comparison of documents, we translate our monolingual vector $\vec{v}_L$ of a document d written in language L into its corresponding vector $\vec{v}_{L'}$ in language L' using BabelNet as our multilingual dictionary. We detail the document-vector translation process in Algorithm 4.1.

The translated vector $\vec{v}_{L'}$ is obtained as follows: for each term t_i with non-zero weight in v_L we obtain all the possible meanings of t_i in BabelNet (see line 5) and, for each of these, we retrieve all the translations (line 7), i.e., lexicalizations of the concept, in language L' available in the synset. We set a non-zero value in the translation vector $\vec{v}_{L'}$,[5] in correspondence with each such translation tr, proportional to the weight of t_i in the original vector

[5]To make the translation possible, while at the same time keeping the same number of dimensions in our vector representation, we use a shared vocabulary which covers both languages. See Section 4.6 for details on the experimental setup.

and the confidence of the translation (line 10), as provided by the BabelNet semantic network.[6]

In order to increase the amount of information available in the vector and counterbalance possible wrong translations, we avoid translating all vectors to one language. Instead, in the present work we create a multilingual vector representation of a document d written in language L by concatenating the corresponding vector $\vec{v}_L$ with the translated vector $\vec{v}_{L'}$ of d for language L'. As a result, we obtain a multilingual vector $\vec{v}_{LL'}$, which contains lexicalizations in both languages.

4.4.2 Similarity between Multilingual Vectors

Following common practice for document similarity in the literature (Manning et al., 2008), we use the cosine similarity as the similarity measure between multilingual vectors:

$$S_v(\vec{v}_{LL'}, \vec{v}'_{LL'}) = \frac{\vec{v}_{LL'} \cdot \vec{v}'_{LL'}}{||\vec{v}_{LL'}||\ ||\vec{v}'_{LL'}||}. \tag{4.7}$$

4.5 Knowledge-based Document Similarity

Given a source document d and a target document d', we calculate the similarities between the respective knowledge-graph and multilingual vector representations, and combine them to obtain a knowledge-based similarity as follows:

$$\text{KBSim}(d, d') = c(G)S_g(G, G') + (1 - c(G))S_v(\vec{v}_{LL'}, \vec{v}'_{LL'}), \tag{4.8}$$

where $c(G)$ is an interpolation factor calculated as the edge density of knowledge graph G:

$$c(G) = \frac{|E(G)|}{|V(G)|(|V(G)| - 1)}. \tag{4.9}$$

[6]Non-English lexicalizations in BabelNet have confidence 1 if originating from Wikipedia inter-language links and $\leq$ 1 if obtained by means of statistical machine translation (Navigli and Ponzetto, 2012a).

Note that, using the factor $c(G)$ to interpolate the two similarities in Eq. 4.8, we determine the relevance for the knowledge graphs and the multilingual vectors in a dynamic way. Indeed, $c(G)$ makes the contribution of graph similarity depend on the richness of the knowledge graph.

4.6 Evaluation

In this section we compare our knowledge-based document similarity measure, KBSim, against state-of-the-art models on two different tasks: comparable document retrieval and cross-lingual text categorization.

4.6.1 Comparable Document Retrieval

In our first experiment we determine the effectiveness of our knowledge-based approach in a comparable document retrieval task. Given a document d written in language L and a collection $D_{L'}$ of documents written in another language L', the task of comparable document retrieval consists of finding the document in $D_{L'}$ which is most similar to d, under the assumption that there exists one document $d' \in D_{L'}$ which is comparable with d.

4.6.1.1 Corpus and Task Setting

Dataset We followed the experimental setting described in (Platt et al., 2010; Yih et al., 2011) and evaluated KBSim on the Wikipedia dataset made available by the authors of those papers. The dataset is composed of Wikipedia comparable encyclopedic entries in English and Spanish. For each document in English there exists a "real" pair in Spanish which was defined as a comparable entry by the Wikipedia user community. The dataset of each language was split into three parts: 43,380 training, 8,675 development and 8,675 test documents. The documents were tokenized, without stemming, and represented as vectors using a log(tf)-idf weighting (Salton and Buckley, 1988). The vocabulary of the corpus was restricted to 20,000 terms, which were the most frequent terms in the two languages after removing the top 50 terms.

Methodology To evaluate the models we compared each English document against the Spanish dataset and vice versa. Following the original setting, the results are given as the average performance between these two experiments. For evaluation we employed the averaged top-1 accuracy and

Mean Reciprocal Rank (MRR) at finding the real comparable document in the other language. We compared KBSim against the state-of-the-art supervised models S2Net, OPCA, CCA, and CL-LSI (cf. Section 4.2). In contrast to these models, KBSim does not need a training step, so we applied it directly to the testing partition.

In addition we also included the results of CL-ESA[7], CL-C3G[8] and two simple vector-based models which translate all documents into English on a word-by-word basis and compared them using cosine similarity: the first model (CosSim_E) uses a statistical dictionary trained with Europarl using Wavelet-Domain Hidden Markov Models (He, 2007), a model similar to IBM Model 4; the second model (CosSim_{BN}) instead uses Algorithm 4.1 to translate the vectors with BabelNet.

4.6.1.2 Results

As we can see from Table 4.1,[9] the CosSim_{BN} model, which uses BabelNet to translate the document vectors, achieves better results than CCA and CL-LSI. We hypothesize that this is due to these linear projection models losing information during the projection. CosSim_E yields results similar to CosSim_{BN}, showing that BabelNet is a good alternative statistical dictionary. In contrast to CCA and CL-LSI, OPCA performs better thanks to its improved projection method using a noise covariance matrix, which enables it to obtain the main components in a low-dimensional space.

CL-C3G and CL-ESA obtain the lowest results. Considering that English and Spanish do not have many lexical similarities, the low performance of CL-C3G is justified because these languages do not share many character n-grams. The reason behind the low results of CL-ESA can be explained by the low number of intersecting concepts between Spanish and English in Wikipedia, as confirmed by Potthast et al. (2008). Despite both using Wikipedia in some way, KBSim obtains much higher performance than CL-ESA thanks to the use of our multilingual knowledge graph representation of documents, which makes it possible to expand and semantically relate its

[7] Document collections with sizes higher than 10^5 provide high performance (Potthast et al., 2008). Here we used 15k documents from the training set to index the test documents.

[8] CL-C3G is CL-CNG using character 3-grams, which has proven to be the best length (Mcnamee and Mayfield, 2004).

[9] In this work, statistically significant results according to a χ^2 test are highlighted in bold.

Model	Dimension	Accuracy	MRR
S2Net	2000	**0.7447**	0.7973
KBSim	N/A	**0.7342**	0.7750
OPCA	2000	**0.7255**	0.7734
CosSim_E	N/A	0.7033	0.7467
CosSim_{BN}	N/A	0.7029	0.7550
CCA	1500	0.6894	0.7378
CL-LSI	5000	0.5302	0.6130
CL-ESA	15000	0.2660	0.3305
CL-C3G	N/A	0.2511	0.3025

Table 4.1. Test results for comparable document retrieval in Wikipedia. S2Net, OPCA, CosSim_E, CCA and CL-LSI are from (Yih et al., 2011).

original concepts. As a result, in contrast to CL-ESA, KBSim can integrate conceptual and relational similarity functions which provide more accurate performance. Interestingly, KBSim also outperforms OPCA which, in contrast to our system, is supervised, and in terms of accuracy is only 1 point below S2Net, the supervised state-of-the-art model using neural networks.

4.6.2 CROSS-LANGUAGE TEXT CATEGORIZATION

The second task in which we tested the different models was cross-language text categorization. The task is defined as follows: given a document d_L in a language L and a corpus $D'_{L'}$ with documents in a different language L', and C possible categories, a system has to classify d_L into one of the categories C using the labeled collection $D'_{L'}$.

4.6.2.1 Corpus and Task Setting

Dataset To perform this task we used the Multilingual Reuters Collection (Amini et al., 2009), which is composed of five datasets of news from five different languages (English, French, German, Spanish and Italian) and classified into six possible categories. In addition, each dataset of news is translated into the other four languages using the Portage translation system (Sadat et al., 2005). As a result, we have five different multilingual datasets, each containing source news documents in one language and four sets of translated documents in the other languages. Each of the languages

has an independent vocabulary. Document vectors in the collection are created using TFIDF-based weighting.

Methodology To evaluate our approach we used the English and Spanish news datasets. From the English news dataset we randomly selected 13,131 news as training and 1,875 as test documents. From the Spanish news dataset we selected all 12,342 news as test documents. To classify both test sets we used the English news training set. We performed the experiment at cross-lingual level using Spanish and English languages available for both Spanish and English news datasets, therefore we classified each test set selecting the documents in English and using the Spanish documents in the training dataset, and vice versa. We followed Platt et al. (2010) and averaged the values obtained from the two comparisons for each test set to obtain the final result. To categorize the documents we applied k-NN to the ranked list of documents according to the similarity measure employed for each model. We evaluated each model by estimating its accuracy in the classification of the English and Spanish test sets.

We compared our approach against the state-of-the-art supervised models in this task: OPCA, CCA and CL-LSI (Platt et al., 2010). In addition, we include the results of the CosSim_{BN} and CosSim_{E} models that we introduced in Section 4.6.1.1, as well as the results of a full statistical machine translation system trained with Europarl and post-processed by LSA (Full MT), as reported by Platt et al. (2010).

4.6.2.2 Results

Table 4.2 shows the cross-language text categorization accuracy. CosSim_{E} obtained the lowest results. This is because there is a significant number of untranslated terms in the translation process that the statistical dictionary cannot cover. This is not the case in the CosSim_{BN} model which achieves higher results using BabelNet as a statistical dictionary, especially on the Spanish news corpus.

On the other hand, however, the linear projection methods as well as Full MT obtained the highest results on the English corpus. The differences between the linear projection methods are evident when looking at the Spanish corpus results; OPCA performed best with a considerable improvement, which indicates again that it is one of the most effective linear projection

Model	Dim.	EN News Accuracy	ES News Accuracy
KBSim	N/A	0.8189	**0.6997**
Full MT	50	**0.8483**	0.6484
CosSim$_{BN}$	N/A	0.8023	**0.6737**
OPCA	100	**0.8412**	0.5954
CCA	150	**0.8388**	0.5323
CL-LSI	5000	**0.8401**	0.5105
CosSim$_{E}$	N/A	0.8046	0.4481

Table 4.2. Test results for cross-language text categorization. Full MT, OPCA, CCA, CL-LSI and CosSim$_E$ are from (Platt et al., 2010).

methods. Finally, our approach, KBSim, obtained competitive results on the English corpus, performing best among the unsupervised systems, and the highest results on the Spanish news, surpassing all alternatives.

Since KBSim does not need any training for document comparison, and because it is based, moreover, on a multilingual lexical resource, we performed an additional experiment to demonstrate its ability to carry out the same text categorization task in many languages. To do this, we used the Multilingual Reuters Collection to create a 3,000 document test dataset and 9,000 training dataset[10] for five languages: English, German, Spanish, French and Italian. Then we calculated the classification accuracy on each test set using each training set. Results are shown in Table 4.3.

The best results for each language were obtained when working at the monolingual level, which suggests that KBSim might be a good untrained alternative in monolingual tasks, too. In general, cross-language comparisons produced similar results, demonstrating the general applicability of KBSim to arbitrary language pairs in multilingual text categorization. However, we note that German, Italian and Spanish training partitions produced low results compared to the others. After analyzing the length of the documents in the different datasets we discovered that they have different average lengths in words: 79 (EN), 76 (FR), 75 (DE), 60 (ES) and 55 (IT). German, Spanish and especially Italian documents have the lowest average length, which

[10]Note that training is needed for the k-NN classifier, but not for document comparison.

Testing datasets	**Training datasets**				
	DE	EN	ES	FR	IT
DE	0.8053	0.6872	0.5373	0.6417	0.5920
EN	0.5827	0.8463	0.5540	0.6530	0.5820
ES	0.5883	0.6153	0.8707	0.6237	0.7010
FR	0.6867	0.7103	0.6667	0.8227	0.6887
IT	0.5973	0.5487	0.6263	0.5973	0.8317

Table 4.3. KBSim accuracy in a multilingual setup.

makes it more difficult to build a representative knowledge graph of the content of each document when it is performing at cross-language level.

4.7 Conclusions

In this paper we introduced a knowledge-based approach to represent and compare documents written in different languages. The two main contributions of this work are: i) a new graph-based model for the language-independent representation of documents based on the BabelNet multilingual semantic network; ii) KBSim, a knowledge-based cross-language similarity measure between documents, which integrates our multilingual graph-based model with a traditional vector representation.

In two different cross-lingual tasks, i.e., comparable document retrieval and cross-language text categorization, KBSim has proven to perform on a par or better than the supervised state-of-the-art models which make use of linear projections to obtain the main components of the term vectors. We remark that, in contrast to the best systems in the literature, KBSim does not need any parameter tuning phase nor does it use any training information. Moreover, when scaling to many languages, supervised systems need to be trained on each pair, which can be very costly.

The gist of our approach is in the knowledge graph representation of documents, which relates the original terms using expanded concepts and relations from BabelNet. The knowledge graphs also have the nice feature of being human-interpretable, a feature that we want to exploit in future work. We will also explore the integration of linear projection models, such as OPCA

and S2Net, into our multilingual vector-based similarity measure. Also, to ensure a level playing field, following the competing models, in this work we did not use multi-word expressions as vector components. We will study their impact on KBSim in future work.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234, EC WIQ-EI IRSES (Grant No. 269180) and MICINN DIANA-Applications (TIN2012-38603-C02-01). Thanks go to Yih et al. for their support and Jim McManus for his comments.

5 Discussion of the Results

In this chapter we discuss in detail the results obtained in this thesis. We first analyse the results of the publications presented in the Chapters 2, 3, and 4 with respect to the objectives of this research. We also include some further results in order complete the picture at task level, i.e., we investigate the performance of the CL-KGA and KBSim models when evaluated with our new weighting scheme (see Section 3.3.3) in the three cross-language tasks. Finally, we present our experiments and results with knowledge graphs in the NLP tasks of community questions answering, native language identification, and language variety identification.

5.1 Single- and Cross-domain Polarity Classification Results

The experiments of Section 2.4 in the single- and cross-domain polarity classification tasks show the potential of knowledge graphs and their features at single- and specially cross-domain level.

Regarding the results of the proposed meta-learning approach, KE-Meta, combining traditional features such as BOW and word n-grams with WSD- and vocabulary expansion-based ones extracted from knowledge graphs, we have several highlights. In Tables 2.5 and 2.6 we observe how the combination of different features and classifiers make KE-Meta the most stable model across domains.[1] Thanks to the stacking generalization, the second level classifier learns how to exploit the base classifier probabilities to counterbalance wrong classifications. In addition, the results of these tables joint with the statistics of Table 2.7 manifest that our model is able to excel also in domains such as books and DVDs with larger texts, generally produced for summaries of the histories being reviewed. Finally, we note that KE-Meta

[1] Note that the use of string kernels for the same dataset and task also offers an excellent stability (Giménez-Pérez et al., 2017).

extracts knowledge from the source domains, ML-SentiCon, and the BabelNet multilingual semantic network, and performs at par or better than the state of the art without using any domain adaptation.

The advantages of the proposed model are specially relevant from the polarity classification viewpoint. However, in the framework of this thesis, we are more interested in the potential of the knowledge graph-based features in a cross-domain scenario. The Figure 2.4 shows the information gain ratio of the base classifier features at single- and cross-domain level. Key are the cross-domain values of the WSD- and vocabulary expansion-based features: "All synsets (Post-WSD)" and "Vocab. Exp.", respectively. They manifest that the amount of information of these features is similar to the ones provided by BOW and word n-grams. The isolated results of the base classifiers in Tables 2.3 and 2.4 validate more this fact. However, what is more important is not the equality of relevance of the features, but the difference in type of information, which conducts to additional improvements when these base classifiers are combined. This can be appreciated in Figures 2.5 and 2.6. Each new base classifier included in KE-Meta contributes, on average, with additional improvements of classification. The only exception is the vocabulary expansion-based one at cross-domain level, that produces a small average decrease. It seems that the vocabulary expansion of the source domains provides with too much unrelated concepts regarding the target domain, and the resulting classifier is affected by this noise. However, in Table 2.6 we show the results of KE-Meta and KE-Meta$_B$, that do not include this base classifier, and the differences in accuracy are not statistically significant.

At this point, we would like to point out that the experiments of Tables 2.5 and 2.6, published in Franco-Salvador et al. (2015b), use 10-fold cross-validation. However, this means that the number of training instances is much higher at cross-domain level (5,400 instances), where we train with all the available domains but the one to classify. Note that the state-of-the-art approaches compared in those tables use the same number of training instances at single and cross-domain level (1,800 instances).[2] For the sake of correctness, in Table 5.1 we show the results of our KE-Meta model training with 1,800 random instances taken from the original 5,400 ones. As we can see, for this dataset and task, 1,800 instances are enough to make KE-Meta a

[2] As we already pointed out, the results of the compared approaches at cross-domain level are taken from Bollegala et al. (2013).

competitive approach. We note that our second level meta-classifier has only eight features. Therefore, a few thousands of instances may provide with enough diversity of feature combinations to create a good classifier.

	Method	Books	Electronics	DVDs	Kitchen
(a)	SST	0.763	**0.839**	0.783	**0.852**
	SFA	**0.777**	0.753	0.763	0.815
	SCL-MI	0.746	0.789	0.763	0.820
(b)	BOW	0.756	0.804	0.791	0.809
	(1+2+3)-grams	0.744	0.798	0.771	0.769
	KE-Meta (#Tr=5,400)	**0.779**	0.789	**0.804**	0.825
	KE-Meta (#Tr=1,800)	**0.768**	0.767	**0.807**	0.819

Table 5.1. Accuracy results in cross-domain polarity classification. (a) State-of-the-art approaches; (b) baselines; (c) proposed approaches.

After the analysis of the features extracted from knowledge graphs, we conclude that WSD-based features are useful at cross-domain level and contribute to obtain additional improvements when combined with traditional ones. In contrast, the features of the vocabulary expansion only benefited at single-domain level and their cross-domain potential is questionable.

5.1.1 CLARIFICATION ABOUT THE POLARITY CLASSIFICATION MODELLING

In Chapter 4 we show how to perform a cross-language classification task using KBSim. However, we can appreciate that we did not choose that model for this classification task. We made this decision because of the nature of this task. As we mentioned in Section 2.2, Pang et al. (2002) concluded that polarity classification achieves worse results than other text classification tasks when approaching them in the same way. This is produced because the text polarity is based on more abstract aspects of text, words, and their meaning. The existence of sentiment analysis variants of semantic networks such as the WordNet's one — SentiWordNet —, also highlights that standard knowledge is not enough. This affects to the BabelNet multilingual semantic network too. It is also the same that we observed at the beginning of this study when we applied KBSim. Its results were notably inferior to the state of the art and even to the BOW baseline. We obtained values of accuracy

close to 60% that made us reconsider the modelling with knowledge graphs for this task. In consequence, we selected the meta-learning scheme and the features employed in Chapter 2.

5.2 Cross-language Plagiarism Detection Results

Chapter 3 covers several objectives of this work: (i) we develop the CL-KGA model for cross-language similarity analysis (cf. Section 3.4); (ii) we study the characteristics of the knowledge graphs (cf. Section 3.3.4); and (iii) we evaluate its performance in the task of cross-language plagiarism detection comparing it with the state of the art obtaining good results (cf. Section 3.5). However, in Chapter 4 we introduced KBSim, an improved version of CL-KGA that has a vector component in order to cover knowledge graph shortcomings such as out-of-vocabulary words and verbal tenses. That model has not been evaluated in the plagiarism detection task and has not been employed jointly with the new, and better (cf. Section 3.5), weighting scheme of knowledge graph relations proposed in Section 3.3.3.2. On the other hand, Chapter 4 shows the good performance that distributed representation-based models obtain in other cross-language similarity tasks. Therefore, aiming to complement the study with the knowledge graphs of Chapter 3, in this section we show the results of KBSim employing our new weighting scheme and compare our models with several cross-language distributed representation-based models. In addition, following Barrón-Cedeño et al. (2013), we perform an in-depth study of the results as function of the different types of cases of plagiarism and the ranking of similarities that the models return for each document.

This section is structured as follows. In Section 5.2.1 we review several reference methods for cross-language similarity analysis that employ distributed representations. Next, in Section 5.2.2 we evaluate these methods and compare them with the most relevant ones of Chapter 3 and the KBSim model with our new weighting scheme. Finally, in Section 5.2.3 we discuss the results of this part of the thesis.

5.2.1 Distributed Representations for Cross-language Plagiarism Detection

This section presents details of the distributed representation learning algorithms for cross-language similarity analysis. These models are usually cate-

gorised according to the objective function they optimise and the type of data they receive as input. Most of these models learn cross-lingual distributed representations using parallel or comparable corpus. For a fair comparison, all of these models are trained using the same parallel corpus. We used 250k English-Spanish and English-German parallel sentences from DGT-Translation Memory distributed by JRC[3]. For monolingual preinitialisation in XCNN (Section 5.2.1.3) we used CLEF ad-hoc retrieval corpus document titles.

5.2.1.1 Similarity Learning via Siamese Neural Network

In this section we describe more in detail the S2Net model that has been employed in Chapter 4. Following the general Siamese neural network architecture (Bromley et al., 1993), Similarity Learning via Siamese Neural Network (S2Net) trains two identical neural networks concurrently. The S2Net receives as input parallel data with binary or real-valued similarity score and updates the model parameters accordingly (Yih et al., 2011). It optimises a dynamic objective function which is directly modelled by using the cosine similarity. The projection operation can be described as follows:

$$y_d = W * x_d, \tag{5.1}$$

where, x_d is the input term vector for the document d, W is the learnt projection matrix (represented by the model parameters) and y_d is the latent representation of document d. The parameters of the S2Net are tuned accordingly to the details provided in Yih et al. (2011).

5.2.1.2 Bilingual Autoencoder

Salakhutdinov and Hinton (2009) demonstrated that semantic modelling by means of dimensionality reduction through deep autoencoders lead to superior performance compared to the conventional LSA approach. Deep autoencoders were extended to model cross-language data and are referred to as Bilingual Autoencoders (BAE) (Gupta et al., 2014; Lauly et al., 2014a,b). These networks learn cross-language associations by optimising the reconstruction error of the cross-language data.

[3] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

The building block of the autoencoder is the Restricted Boltzmann Machine (RBM). These deep networks are trained through a greedy layer-by-layer pretraining stage followed by a supervised fine-tuning. The structures of the network and the training architecture are shown in Figure 5.1. For more details, please refer to Gupta et al. (2014).

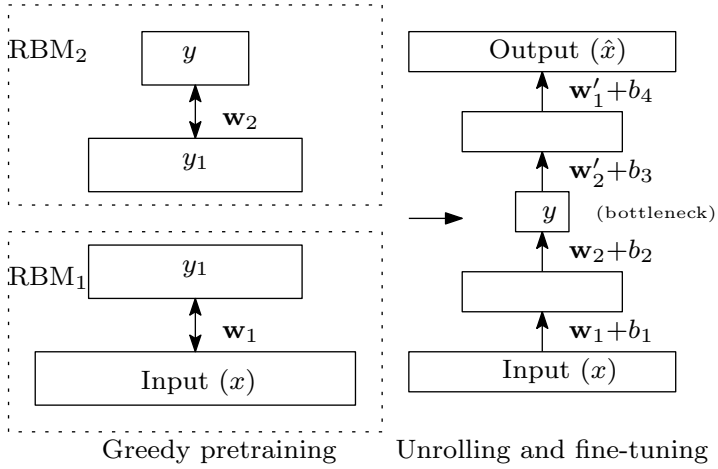

Figure 5.1. Left panel: pretraining of stacked RBMs where the upper RBM takes as input the output of the lower RBM. Right panel: After pretraining the structure is "unrolled" to create a multi-layer network which is fine-tuned by means of backpropagation to learn an identity function $\hat{x} \approx x$.

5.2.1.3 External-data Composition Neural Networks

External-data Composition Neural Network (XCNN) is based on a composition function that is implemented on top of a deep neural network that provides a distributed learning framework (Gupta et al., 2015). Different from many other models including S2Net and BAE, which solely rely on parallel/comparable data for training, XCNN exploits also monolingual data for model training purposes. Specifically, it incorporates external relevance signals such as pseudo-relevance data or clickthrough data into the learning framework. The main motivation behind this strategy is that, monolingual models can be initialised from such largely available relevance data and then, with the help of a smaller amount of parallel data, the cross-lingual model can be trained. This property helps to gain more confidence for under-represented terms in parallel data, i.e. terms with very low frequency.

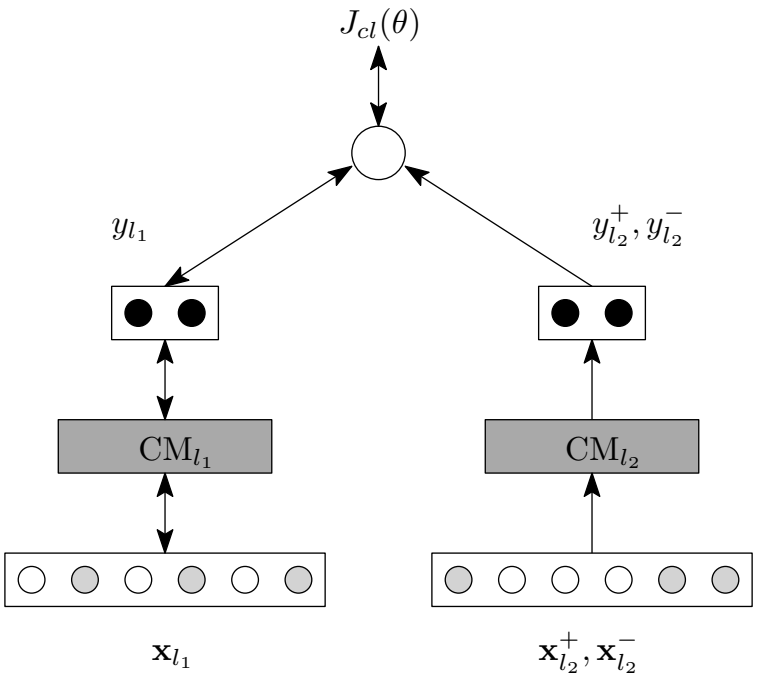

Figure 5.2. Architecture of external-data composition neural network model for cross-lingual training.

The architecture of XCNN model training is shown in Figure 5.2. XCNN learns distributed representations of words in cross-lingual setting using the objective function defined in Eq. 5.2. It maximises the cosine similarity φ for a training example for a positive sample and miminises it for a negative sample. The network parameters are updated through backpropagation as follows:

$$J_{cl}(\theta) = \varphi(y_{l_1}, y^+_{l_2}) - \varphi(y_{l_1}, y^-_{l_2}) \tag{5.2}$$

The representation of an input text is obtained through an addition composition function as described below:

$$\begin{aligned} y_i^{(l_1)} &= \mathrm{g}(W_1 * x_i + b_1) \\ y_i^{(l_j)} &= \mathrm{g}(W_j * y_i^{l_{(j-1)}} + b_j), j = 2, ..., m \\ y &= \sum_{i=1}^{n} y_i^{(l_m)} \end{aligned} \tag{5.3}$$

where $y_i^{(l^j)}$ represents the i^{th} term x_i in text in the layer j of a neural network, l_m represents the output layer. More details about XCNN can be found in Gupta et al. (2015).

5.2.1.4 Continuous Word Alignment-based Similarity Analysis

The aforementioned distributed representation models learn a real-valued high dimensional representation of texts of different length. All of them combine the word level representations by summing over all the terms present in a text as bag-of-words model. In this section, we present an alternative method to combine word level vectors by means of alignments to represent text. The Continuous Word Alignment-based Similarity Analysis (CWASA) model (Franco-Salvador et al., 2016a) modifies the text-to-text relatedness proposed by (Hassan and Mihalcea, 2011) in order to estimate the similarity between documents by efficiently aligning their distributed representations of words using directed edges, i.e., we exploit the fact that closest words between documents may have not reciprocal relationships, e.g. in the sentences “Michelle_Obama from United_States" and “Barak_Obama and the First_Lady", *United_States* could have *Barak_Obama* as closest, and this could have *Michelle_Obama*, who in turn could be the closest to *First_Lady* in both directions. Formally, the similarity $\mathrm{S}(d, d')$ between two documents d and d' is estimated as follows:

$$\mathrm{S}(d, d') = \frac{1}{|\Phi|} \sum_{c_k \in \Phi} c_k, \tag{5.4}$$

where $d = (x_1, ..., x_n)$ and $d' = (y_1, ..., y_m)$ are represented as lists of distributed representations of words, and Φ is generated from the list $\Phi' = \{c'_1, ..., c'_{n+m}\}$ that satisfies Eq. 5.5:

$$c'_k = \begin{cases} \underset{i=k, x_i \in d, y_j \in d'}{\arg\max} \varphi(x_i, y_j), & \text{if } k \leq n \\ \underset{j=k-n, x_i \in d, y_j \in d'}{\arg\max} \varphi(x_i, y_j), & \text{otherwise} \end{cases} \tag{5.5}$$

where $1 \leq i \leq n$, $1 \leq j \leq m$, $1 \leq k \leq n+m$, φ is the cosine similarity function, and being $\Phi = \{c_1, ..., c_z \mid \max(n,m) \leq z \leq n+m\}$, $\Phi \subseteq \Phi'$, the set of cosine similarities without pairing repetitions[4] that represents the strongest semantic pairing between the distributed representations of words of documents d and d'.

Basically, in Eq. 5.5 we align each word in d with the closest one in d' and vice versa using directed relationships. Next, we remove duplicated alignments, i.e., those equally aligned in both directions. Finally, we use Eq. 5.4 to estimate the similarity score between d and d' as the average of the different alignments. We note that this problem has been efficiently solved by dynamic programming. In addition, although this section is focused on a cross-lingual setting, CWASA can be directly employed with monolingual distributed representations of words (see Section 5.4.1). We compare our CWASA model with the classical bag-of-words sum representation in the next section.

5.2.2 Complementary Evaluation of Cross-language Plagiarism Detection

In this section we complement the evaluation of Chapter 3 in the task of cross-language plagiarism detection. We employ the PAN-PC-11 dataset and the PAN shared task measures described in Section 3.5: precision, recall, granularity, and plagdet. In addition, following Barrón-Cedeño et al. (2013), we also perform a deeper study of the results as function of the different types of cases of plagiarism and the ranking of similarities that models return for each document.

[4] We do not permit the same pair of words aligned twice.

Spanish-English (ES-EN) documents		**German-English (DE-EN) documents**	
Suspicious	304	Suspicious	251
Source	202	Source	348
Plagiarism cases {Spanish,German}-English			
Case length		Obfuscation	
– Long length cases	1,506	– Translated automatic obfuscation	5,142
– Medium length cases	2,118	– Translated manual obfuscation	433
– Short length cases	1,951		

Table 5.2. Statistics of PAN-PC-11 cross-language plagiarism detection partitions.

In Section 5.2.2.1, our first experiment shows the recall at character level of the models. This experiment serves to show the potential of the models detecting plagiarism cases before the detailed analysis and postprocessing described in Algorithm 3.1. Recall is measured using the top k (R@k) most similar fragments of text, where $k = \{1, 5, 10, 20\}$. However, in order to increase precision, we conduct a second experiment in Section 5.2.2.2. There, as in the evaluation of Section 3.5, detections are filtered using Algorithm 3.1 to determine which cases are plagiarism. Finally, in Section 5.2.2.3 we compare the computational efficiency of the models. In both experiments of Section 5.2.2.1 and 5.2.2.2 we also include in a separated subsection the analysis of results as function of the type of obfuscation and document length of the plagiarism cases. In Table 5.2 we present the statistics of the PAN-PC-11 dataset considering these types of cases.

For this evaluation, we selected from Chapter 3 the CL-C3G, CL-ESA, CL-ASA, and CL-KGA models.[5] We also show the S2Net, BAE, and XCNN distributed representation-based models detailed in Section 5.2.1. In addition, we use our CWASA model (cf. Section 5.2.1.4) in order to represent documents by means of distributed word alignments: CWASA (S2Net), CWASA (BAE), and CWASA (XCNN). Finally, we show the performance of the original KBSim model, KBSim (VSM) from here, the isolated results of its vector component (VSM), and the results when replacing that vector component with the document vectors of the distributed representation-based models: KBSim (S2Net), KBSim (BAE), and KBSim (XCNN). We perform this combination of distributed representations and knowledge graphs be-

[5] We decided to not include the alternative CL-KGA variants presented Chapter 3 to focus on the comparison between reference models.

	Model	Spanish-English (ES-EN)				German-English (DE-EN)			
		R@1	**R@5**	**R@10**	**R@20**	**R@1**	**R@5**	**R@10**	**R@20**
(a)	CL-KGA	**0.924**	**0.952**	**0.960**	**0.963**	**0.803**	**0.871**	**0.899**	**0.916**
	VSM	0.791	0.880	0.905	0.924	0.630	0.786	0.831	0.872
	CL-ASA	0.663	0.787	0.819	0.853	0.523	0.693	0.755	0.806
	CL-ESA	0.677	0.784	0.824	0.858	0.481	0.611	0.666	0.720
	CL-C3G	0.497	0.672	0.743	0.805	0.204	0.393	0.489	0.593
(b)	S2Net	**0.637**	**0.763**	**0.809**	**0.852**	**0.508**	**0.675**	**0.744**	**0.799**
	XCNN	0.468	0.648	0.721	0.786	0.362	0.561	0.647	0.728
	BAE	0.509	0.717	0.784	0.836	0.308	0.513	0.607	0.697
(c)	CWASA (XCNN)	**0.881**	**0.921**	**0.937**	**0.946**	**0.739**	**0.823**	**0.849**	**0.873**
	CWASA (S2Net)	0.859	0.909	0.921	0.936	0.601	0.731	0.779	0.818
	CWASA (BAE)	0.536	0.695	0.754	0.803	0.543	0.701	0.760	0.806
(d)	KBSim (S2Net)	0.932	0.958	0.962	0.962	**0.818**	**0.885**	**0.904**	**0.922**
	KBSim (VSM)	**0.940**	**0.965**	**0.968**	**0.972**	0.809	0.883	**0.904**	0.921
	KBSim (BAE)	0.926	0.951	0.959	0.963	0.798	0.876	0.896	0.912
	KBSim (XCNN)	0.868	0.915	0.934	0.938	0.753	0.852	0.879	0.902

Table 5.3. ES-EN and DE-EN performance analysis in terms of R@k, where k = {1, 5, 10, 20}.

cause we believe that these representations, which separately perform better than VSM in other tasks (Yih et al., 2011), are also able to complement better the knowledge graph component of KBSim. We recall that our DCW weighting scheme of semantic relations of knowledge graphs proved its superiority for this task (cf. Section 3.5). Therefore, in this section CL-KGA and KBSim only employ the DCW weighting scheme (cf. Section 3.3.3.2).

All our tables separate the models according to their category: (a) state-of-the-art approaches; (b) continuous word representation-based approaches; (c) proposed word-vector alignment-based approaches; and (d) hybrid approaches.

5.2.2.1 Experiment A: Cross-language Similarity Ranking

In this section we compare the R@k of the models when ranking the most similar fragments of text with the plagiarism cases. First, we analyse the results of the complete PAN-PC-11 dataset. Next, in Section 5.2.2.1 we analyse the results on the basis of the type of plagiarism case. In Table 5.3 we

show the results for ES-EN and DE-EN.[6] As we can see, DE-EN similarity is more difficult to detect for all the models. Overall, no differences are found between the models with respect to the ranking order. Therefore, we can jointly analyse the differences between them. The models which employ knowledge graphs, CL-KGA and KBSim, obtain the best results. The difference between CL-KGA and other state-of-the-art models in R@1 is superior to 25% (absolute value), and highlights the potential of such type of representations. The use of bilingual vectors and the TF-IDF re-weighting benefits VSM that obtains interesting results too. It is followed, in order of performance, by CL-ASA, CL-ESA, and CL-C3G, that is the baseline in all our experiments. These results are in line with those analysed in Section 3.5.

Despite the good performance of distributed representations in other tasks (see Section 4.6), the models of group (b) offer average performance compared to the state of the art. The S2Net model obtains superior results than XCNN and BAE, specially in DE-EN. Note that S2Net and BAE directly learn representations of text using a bag-of-words format. Therefore, distributed representations of large fragments of text are still representative. In contrast, XCNN learns word-level distributed representations and hence when projecting a large fragment of text ($\sim$1000 words) the summed distributed representations flatten vectors and loose discriminative power, affecting XCNN performance. However, these comments refer to the case when the cosine similarity is employed to compare continuous vectors of documents based on the sum of word vectors. The performance differs when the word vectors are used without this sum-based composition.

The use of word alignments, i.e., by means of CWASA, produces notable improvements respect to the sum of word vectors. e.g. CWASA (XCNN) is 40% superior to XCNN even when it is employing the same word vectors. As we analyse in Section 5.2.2.1, the use of CWASA allows to successfully measure similarity between texts of any length. This allows to employ XCNN word vectors to measure similarity between fragments of text with superior results than CWASA (S2Net) and CWASA (BAE). In addition, despite CWASA does not outperform the CL-KGA model, for computational time constraints we restrict the vocabulary to 20,000 words when using distributed representations, and we are rivalling with a model that employs BabelNet, a

[6] In this experimentation, in all the tables, the best results — at type-of-model level — are highlighted in bold.

multilingual semantic network with more than 9M concepts. The vocabulary coverage of the languages is about 82% for English, 72% for Spanish, and 42% for German. This also justifies the decrease of performance in DE-EN. A higher variety of stemmed words is observed for the German agglutinative language, which is not covered by the vocabulary in the same amount than the other languages. We also note that the performance of BAE shows the highest variation from R@1 to R@5 among all models: $\sim$21%. After a manual analysis of the resulting distributed representations and the values of similarity between texts, we observe a very reduced variance: lower than $\sim 10^{-2}$. This led the model to be less precise when differentiating close elements and affects the performance of CWASA (BAE).

Finally, the combination of knowledge graphs with vectors produces the best results. Similarly to our results in cross-language document retrieval and categorization (cf. Section 4.6), thanks to the dynamic interpolation, the original KBSim (VSM) model obtains higher results than CL-KGA and VSM separately. We may appreciate that the use of distributed vector representations allows to successfully complement knowledge graphs too. KBSim (S2Net) obtains on average the highest results in this experiment. Although KBSim (XCNN) does not obtain such high results, the differences in R@5 are small. As we will see in Section 5.2.2.2, such differences are not relevant when detecting plagiarism and the models performance may change as function of the postprocessing algorithm employed. In Section 5.2.2.2 we will also study the statistical differences of all the models to analyse if the observed differences are significant from a statistical viewpoint. We note that with the current parameters of Algorithm 3.1, R@5 is the recall upper-bound for the plagiarism detection performed in Section 5.2.2.2.

Cross-language Similarity Ranking in Function of the Type of Plagiarism Cases

In this section we analyse the R@k of the models as function of the type of plagiarism. We divide plagiarism cases according to the type obfuscation — translated obfuscation and translated manual obfuscation — employed to generate the case, and according to the case length — short, medium, and

Type of obfuscation		Model	Spanish-English (ES-EN)				German-English (DE-EN)			
			R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
Translated manual obfuscation	(a)	CL-KGA	**0.860**	**0.919**	**0.939**	**0.945**	**0.721**	**0.810**	**0.857**	**0.869**
		VSM	0.696	0.796	0.841	0.877	0.549	0.721	0.781	0.832
		CL-ESA	0.607	0.737	0.795	0.837	0.406	0.548	0.614	0.686
		CL-ASA	0.533	0.662	0.712	0.756	0.387	0.569	0.643	0.713
		CL-C3G	0.450	0.599	0.674	0.738	0.231	0.420	0.537	0.642
	(b)	S2Net	**0.545**	**0.672**	**0.725**	**0.799**	**0.444**	**0.622**	**0.685**	**0.742**
		BAE	0.458	0.635	0.713	0.767	0.297	0.500	0.579	0.677
		XCNN	0.414	0.610	0.669	0.744	0.358	0.572	0.653	0.743
	(c)	CWASA (XCNN)	**0.799**	**0.864**	**0.888**	**0.899**	**0.641**	**0.749**	**0.782**	**0.808**
		CWASA (S2Net)	0.760	0.842	0.857	0.880	0.524	0.669	0.730	0.759
		CWASA (BAE)	0.459	0.623	0.689	0.760	0.345	0.494	0.566	0.653
	(d)	KBSim (S2Net)	0.863	0.917	0.928	0.941	**0.741**	**0.813**	0.844	**0.877**
		KBSim (VSM)	**0.876**	**0.930**	**0.938**	**0.943**	0.721	0.805	**0.855**	0.871
		KBSim (BAE)	0.855	0.912	0.932	0.940	0.725	0.806	0.843	0.865
		KBSim (XCNN)	0.773	0.846	0.875	0.895	0.652	0.783	0.836	**0.877**
Translated automatic obfuscation	(a)	CL-KGA	**0.935**	**0.957**	**0.963**	**0.964**	**0.805**	**0.880**	**0.902**	**0.919**
		VSM	0.799	0.886	0.910	0.928	0.638	0.793	0.837	0.876
		CL-ASA	0.674	0.797	0.828	0.861	0.537	0.706	0.767	0.816
		CL-ESA	0.682	0.788	0.826	0.860	0.488	0.617	0.671	0.723
		CL-C3G	0.500	0.678	0.749	0.810	0.201	0.390	0.485	0.588
	(b)	S2Net	**0.645**	**0.770**	**0.816**	**0.856**	**0.514**	**0.681**	**0.751**	**0.805**
		BAE	0.513	0.724	0.790	0.841	0.309	0.514	0.610	0.699
		XCNN	0.472	0.651	0.725	0.789	0.363	0.559	0.646	0.727
	(c)	CWASA (XCNN)	**0.887**	**0.925**	**0.941**	**0.949**	**0.749**	**0.831**	**0.856**	**0.879**
		CWASA (S2Net)	0.867	0.914	0.926	0.940	0.609	0.738	0.784	0.824
		CWASA (BAE)	0.543	0.701	0.760	0.806	0.409	0.557	0.620	0.682
	(d)	KBSim (S2Net)	0.939	0.962	0.965	0.966	**0.829**	**0.894**	**0.915**	**0.929**
		KBSim (VSM)	**0.941**	**0.965**	**0.967**	**0.967**	0.810	0.889	0.906	0.925
		KBSim (BAE)	0.914	0.955	0.964	**0.967**	0.814	0.890	0.907	0.923
		KBSim (XCNN)	0.875	0.920	0.925	0.941	0.762	0.857	0.882	0.903

Table 5.4. ES-EN and DE-EN performance analysis in terms of the obfuscation type for the plagiarism cases and R@k, where $k = \{1, 5, 10, 20\}$.

long.[7] Most of the highlights of Section 5.2.2.1 persist when discriminating considering the type of case. However, there are several points to note. In Table 5.4 results are reported on the basis of the obfuscation type. The translated manual obfuscation has manual correction after the automatic translation and generates cases with paraphrasing in order to hide the plagiarism. Therefore, it is more difficult to detect similarity between such type of cases.

[7]We follow the PAN-PC-11 setup and consider as short cases those with less than 700 characters. Long cases are those larger than 5,000 characters.

Type of obfuscation		Model	Spanish-English (ES-EN)				German-English (DE-EN)			
			R@1	**R@5**	**R@10**	**R@20**	**R@1**	**R@5**	**R@10**	**R@20**
Long length cases	(a)	CL-KGA	**0.944**	**0.964**	**0.967**	**0.969**	**0.819**	**0.895**	**0.911**	**0.928**
		VSM	0.820	0.903	0.925	0.939	0.655	0.802	0.842	0.881
		CL-ASA	0.701	0.820	0.847	0.878	0.554	0.719	0.779	0.828
		CL-ESA	0.707	0.808	0.841	0.872	0.503	0.631	0.681	0.729
		CL-C3G	0.508	0.690	0.761	0.822	0.197	0.382	0.475	0.580
	(b)	S2Net	**0.662**	**0.785**	**0.830**	**0.867**	**0.523**	**0.688**	**0.757**	**0.812**
		XCNN	0.486	0.663	0.735	0.800	0.351	0.545	0.634	0.717
		BAE	0.524	0.741	0.807	0.857	0.307	0.513	0.608	0.699
	(c)	CWASA (XCNN)	**0.906**	**0.941**	**0.952**	**0.958**	**0.762**	**0.840**	**0.865**	**0.888**
		CWASA (S2Net)	0.886	0.928	0.939	0.950	0.618	0.744	0.788	0.828
		CWASA (BAE)	0.559	0.715	0.772	0.818	0.419	0.560	0.620	0.679
	(d)	KBSim (S2Net)	0.952	0.968	0.970	0.971	**0.842**	**0.903**	**0.922**	**0.934**
		KBSim (VSM)	**0.959**	**0.975**	**0.976**	**0.979**	0.828	0.898	0.915	0.931
		KBSim (BAE)	0.949	0.964	0.969	0.970	0.825	0.898	0.914	0.929
		KBSim (XCNN)	0.897	0.933	0.943	0.948	0.777	0.865	0.888	0.909
Medium length cases	(a)	CL-KGA	**0.933**	**0.960**	**0.965**	**0.967**	**0.803**	**0.879**	**0.901**	**0.916**
		VSM	0.800	0.886	0.910	0.928	0.637	0.792	0.836	0.876
		CL-ASA	0.673	0.796	0.827	0.860	0.530	0.701	0.761	0.812
		CL-ESA	0.688	0.794	0.831	0.865	0.488	0.618	0.671	0.723
		CL-C3G	0.502	0.678	0.748	0.809	0.201	0.389	0.485	0.591
	(b)	S2Net	**0.647**	**0.771**	**0.815**	**0.856**	**0.516**	**0.681**	**0.749**	**0.802**
		XCNN	0.476	0.656	0.727	0.794	0.365	0.563	0.648	0.728
		BAE	0.517	0.728	0.793	0.842	0.309	0.515	0.611	0.699
	(c)	CWASA (XCNN)	**0.888**	**0.926**	**0.939**	**0.947**	**0.746**	**0.828**	**0.853**	**0.877**
		CWASA (S2Net)	0.870	0.917	0.927	0.941	0.611	0.738	0.784	0.823
		CWASA (BAE)	0.546	0.704	0.761	0.809	0.412	0.560	0.621	0.683
	(d)	KBSim (S2Net)	0.939	0.962	0.965	0.965	**0.829**	**0.894**	**0.914**	**0.928**
		KBSim (VSM)	**0.942**	**0.964**	**0.967**	**0.968**	0.814	0.886	0.905	0.922
		KBSim (BAE)	0.929	0.954	0.960	0.964	0.810	0.882	0.901	0.915
		KBSim (XCNN)	0.880	0.923	0.935	0.943	0.759	0.856	0.881	0.903
Short length cases	(a)	CL-KGA	**0.944**	**0.952**	**0.959**	**0.963**	**0.790**	**0.867**	**0.892**	**0.908**
		VSM	0.787	0.876	0.902	0.922	0.621	0.780	0.825	0.867
		CL-ASA	0.659	0.783	0.815	0.850	0.513	0.684	0.748	0.800
		CL-ESA	0.673	0.780	0.820	0.855	0.473	0.602	0.658	0.713
		CL-C3G	0.494	0.669	0.740	0.802	0.201	0.389	0.486	0.590
	(b)	S2Net	**0.633**	**0.758**	**0.806**	**0.848**	**0.501**	**0.668**	**0.738**	**0.793**
		XCNN	0.463	0.644	0.716	0.782	0.361	0.559	0.646	0.728
		BAE	0.503	0.713	0.780	0.831	0.305	0.508	0.601	0.691
	(c)	CWASA (XCNN)	**0.877**	**0.918**	**0.934**	**0.943**	**0.732**	**0.818**	**0.844**	**0.868**
		CWASA (S2Net)	0.856	0.906	0.918	0.933	0.593	0.724	0.772	0.812
		CWASA (BAE)	0.532	0.692	0.751	0.800	0.393	0.543	0.606	0.672
	(d)	KBSim (S2Net)	0.930	0.956	0.961	0.964	**0.814**	**0.881**	**0.903**	**0.921**
		KBSim (VSM)	**0.939**	**0.964**	**0.966**	**0.968**	0.799	0.874	0.896	0.914
		KBSim (BAE)	0.925	0.950	0.958	0.963	0.795	0.872	0.893	0.908
		KBSim (XCNN)	0.865	0.912	0.927	0.937	0.746	0.846	0.872	0.896

Table 5.5. ES-EN and DE-EN performance analysis in terms of plagiarism case length and R@k, where $k = \{1, 5, 10, 20\}$.

CL-ESA, that is based on a representation by similarities with a collection of documents, outperforms CL-ASA in cases with manual obfuscation. This is somehow expected due that ESA was originally meant for tasks of relatedness rather than plagiarism.

In Table 5.5 we can see the results as function of the case length. In opposition to the short cases, the similarity between long cases of plagiarism is the easiest to detect. The additional information that long cases provide, makes it easier to the models to represent and to discriminate between texts. However, those differences in performance rarely excel 2%. The exception is the CL-ASA model, that suffers a higher decay when cases became shorter. This may be produced by the document length component of the model, that is more precise normalising larger cases of plagiarism. Note that KBSim (S2Net) obtains the highest results independently of the type of obfuscation and case length analysed, which highlights its robustness for CL similarity analysis and plagiarism detection.

5.2.2.2 Experiment B: Cross-language Plagiarism Detection

In this section we compare the CWASA continuous word representation and KBSim models with several state-of-the-art approaches on the PAN-PC-11 dataset for CL plagiarism detection. We show the results in Table 5.6. Although both English and German are Germanic languages, due to their grammatical differences, the additional difficulty of the detection in DE-EN is also visible in this experiment. The decay of plagdet — the overall score for plagiarism detection — ranges between 8%-27% when comparing DE-EN with ES-EN results. The lowest results are obtained with CL-C3G, that does not find enough lexical and syntactic similarities to model the content properly using character n-grams. The CL-ESA and CL-ASA models obtain a similar recall but the latter one excels in precision and increases its plagdet. In fact, CL-ESA offers a higher number of false positives. Finally, the CL-KGA model is the best state-of-the-art approach and obtains the highest results in both ES-EN and DE-EN language pairs. Note that the best possible value of granularity is 1.0, which means that our model is not detecting a single case as multiple cases of plagiarism or vice versa. These comments are a short summary of those highlighted in Section 3.5.

As we also pointed out in Section 5.2.2.1, the continuous word representation models, which represent documents based on the sum of word vectors,

	Model	Spanish-English (ES-EN)				German-English (DE-EN)			
		Plag	**Prec**	**Rec**	**Gran**	**Plag**	**Prec**	**Rec**	**Gran**
(a)	CL-KGA	**0.651**	**0.752**	**0.574**	**1.000**	**0.564**	**0.650**	**0.495**	**1.000**
	VSM	0.564	0.630	0.517	1.010	0.414	0.524	0.362	1.048
	CL-ASA	0.517	0.690	0.448	1.071	0.406	0.604	0.344	1.113
	CL-ESA	0.471	0.535	0.448	1.048	0.269	0.402	0.230	1.125
	CL-C3G	0.373	0.563	0.324	1.148	0.115	0.316	0.080	1.166
(b)	S2Net	**0.514**	0.734	**0.440**	**1.098**	**0.379**	**0.669**	**0.304**	1.148
	XCNN	0.386	**0.738**	0.310	1.189	0.270	0.664	0.196	1.174
	BAE	0.440	0.736	0.360	1.142	0.212	0.482	0.150	**1.120**
(c)	CWASA (XCNN)	**0.609**	0.686	**0.547**	**1.001**	**0.492**	**0.611**	**0.430**	**1.037**
	CWASA (S2Net)	0.607	**0.693**	0.542	1.002	0.408	0.585	0.353	1.111
	CWASA (BAE)	0.354	0.546	0.296	1.121	0.237	0.478	0.176	1.122
(d)	KBSim (XCNN)	**0.673**	**0.793**	0.585	**1.000**	**0.586**	**0.741**	**0.485**	**1.000**
	KBSim (VSM)	0.656	0.745	**0.586**	**1.000**	0.574	0.661	0.508	**1.000**
	KBSim (S2Net)	0.652	0.741	0.583	**1.000**	0.572	0.671	0.499	**1.000**
	KBSim (BAE)	0.651	0.743	0.579	**1.000**	0.567	0.659	0.499	**1.000**

Table 5.6. ES-EN and DE-EN performance analysis in terms of plagdet (Plag), precision (Prec), recall (Rec) and granularity (Gran). This table complements the results of Tables 3.5 and 3.6.

offer an average performance in this task.[8] The S2Net model outperforms BAE and XCNN but obtains lower values than CL-KGA. We can see close values in terms of precision for S2Net and XCNN. However, S2Net's recall is 10% higher in all the tests. This, along with the highest granularity, penalises XCNN's plagdet.

The models of group (c) — where CWASA is used to measure similarity — notably improve the performance of S2Net, BAE, and XCNN. We appreciate how, especially with XCNN, recall and granularity improve with a low impact on precision. In contrast to S2Net and BAE, that use a bag-of-words format to learn vectors of documents, XCNN directly generates continuous vectors of words. These vectors find in CWASA an excellent complement in order to accurately measure the CL similarity. Note that in this experiment we use Algorithm 3.1 to analyse the similarities and to iden-

[8] Although S2Net and BAE directly learn representations of text, note that this composition is internally based on the use of a bag-of-words format, that employs the sum of word vectors.

CL-KGA
Plagiarism case in *"suspicious-document04541.txt"*, offset=101,683: You would have a **successor** Vice-**President** in cases of **dismissal**, **resignation** or death. As for the rest was great to affinity that existed between it and the **Constitutional Code sanctioned** in 1811 by **Congress Miranda** met on 2 March.
Plagiarism case in *"suspicious-document06272.txt"*, offset=43,421: The last part of my journey, **night** and raining, **dark corridors** of the **house**, the **kitchen** so **big**, so **dark** at first, then look strange in **light** of the **huge bonfire** fur and things of my **uncle**, the **woman** appeared suddenly gray, the **dark** moorland **dining room**, explored the dim **light** of **lantern** four **glasses** clouded by scab; the silence of "outside" ... worse than silence: a distant sound and intermittent rough, something which put fear into the valiant Don Quixote **chest** one **night** in near Sierra Morena, and the other silent **house** stopped talking about My **uncle** had impressed me badly.
XCNN
Plagiarism case in *"suspicious-document07684.txt"*, offset=454: **And you** better well, because we would **have** been worse **had** both fallen in deepest pit **and** most serious sin. "**I** do not regret it, **having rejected your** honor, what **I** regret is drawing him with unprecedented treachery to **reject** later."
Plagiarism case in *"suspicious-document06175.txt"*, offset=0: "**Not** like **you** go" she said. I fear something terrible happens **to you**: but go, because they want **and** can **not be** avoided. Take, however, this box, **and** very careful **not to open it**. If **you open it**, will never **be** able **to** see me again.

Table 5.7. Example of the type of cases detected by CL-KGA and XCNN. In this table, the cases detected by CL-KGA are not detected by XCNN and vice versa. The bold words highlight semantically related ones in the case of CL-KGA and frequent ones in the case of XCNN.

tify the plagiarism cases. To do this, Algorithm 3.1 retrieves the five most similar fragments with each text fragment in the other language. This penalises BAE that, as we mentioned in Section 5.2.2.1, has a low variance between continuous vectors and makes it more difficult to correctly align the text fragments.

Finally, the combination of vector representations with knowledge graphs, makes the KBSim models of group (d) to obtain on overall the highest results. In fact, KBSim (XCNN) outperforms the original KBSim (VSM), and is the best model, independently of the language pair analysed. This proves the

potential of KBSim for the tasks of CL similarity analysis and plagiarism detection. This also confirms that knowledge graphs and continuous models capture different aspects of text and complement each other. In order to illustrate this fact, we selected from the English partition four cases of plagiarism generated with translated automatic obfuscation.[9]. Two cases (referred as CL-KGA) were detected by CL-KGA and not by XCNN. The other two cases (referred as XCNN) represent the opposite situation. We can see these cases in Table 5.7. Thanks to the wide coverage of the BabelNet multilingual semantic network, our knowledge graph-based model eases the detection of cases with semantically related words. On the other hand, the XCNN model based on continuous representations covers knowledge graph shortcomings such as the out-of-vocabulary words and has the potential to take into account also their frequencies. We note that in this thesis we stemmed the input of the continuous representation models.

Despite the high R@k of some models (see Section 5.2.2.1), the final values of recall, and consequently plagdet, considerably decrease. We note that this is normal if we consider that recall must be reduced in order to obtain a precise model. This also demonstrates the potentialities and limitations of Algorithm 3.1 for plagiarism detection.

After analysing the performance of the models, we are also interested in analysing whether or not the observed differences across the obtained results are statistically significant. In order to analyse this, we use bootstrap resampling[10] (Efron and Tibshirani, 1994) to measure the plagdet of the models in ES-EN and DE-EN including also their confidence intervals. We show the results in Figure 5.3. As we can see, the KBSim and CL-KGA models do not show significant differences. Despite KBSim (XCNN) obtains on average a higher performance, these results show that CL-KGA or other KBSim models perform similarly. However, the larger confidence intervals of some models denote a higher variability in performance, e.g. with KBSim (VSM) in DE-EN. With respect to the CWASA model, CWASA (XCNN) and CWASA (S2Net) are notably superior to XCNN and S2Net. This highlights again the potential of CWASA and its alignments for distributed word-based

[9]There is no need to include the source of plagiarism because it is basically a translation into either Spanish or German.

[10]Bootstrap methods obtain generally better results in parametric tests for small datasets — as the dataset in hand — or where sample distributions are non-normal. The statistical tests are calculated with an α of 0.05 and 1,000 samplings.

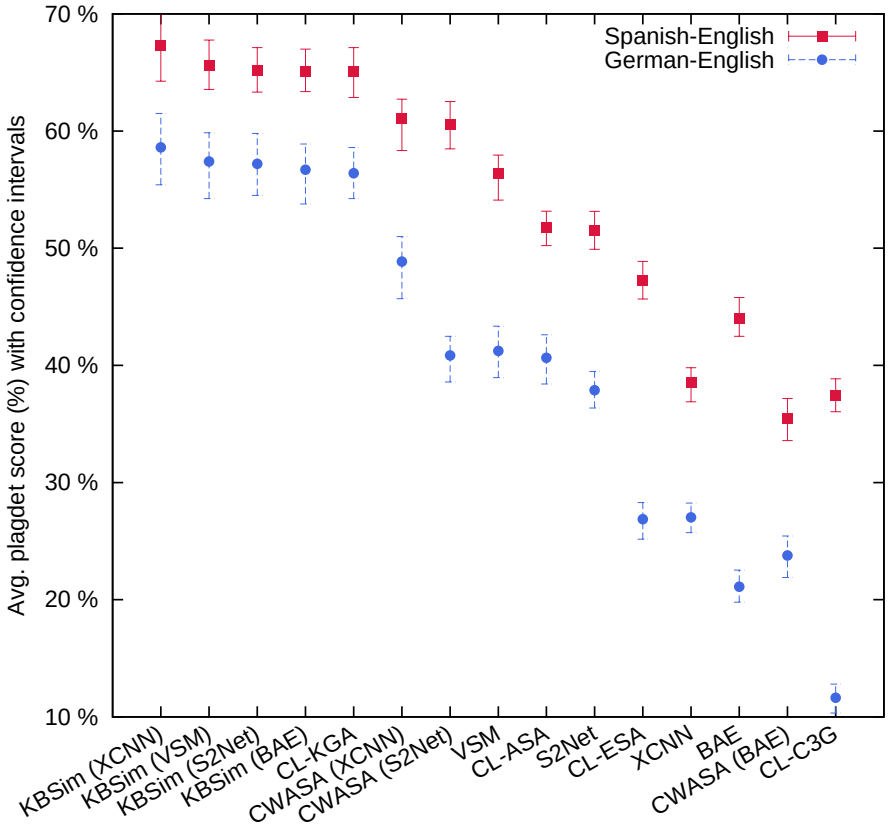

Figure 5.3. Plagdet score (%) of the compared models with confidence intervals for the Spanish-English and German-English partitions. Non-overlapped intervals among models represent statistically significant differences.

similarity analysis. In addition, CWASA (XCNN) proves to be also superior to CWASA (S2Net) in DE-EN and, therefore, the most stable. Finally, note that KBSim (XCNN) has the shortest distance between intervals of the same model across language pairs. This 3% of division suggests that the model is the most stable across languages for CL plagiarism detection.

Cross-language Plagiarism Detection as Function of the Type of Plagiarism Cases

Type of obfuscation		Model	Spanish-English (ES-EN)				German-English (DE-EN)			
			Plag	Prec	Rec	Gran	Plag	Prec	Rec	Gran
Translated manual obfuscation	(a)	CL-KGA	**0.161**	**0.175**	**0.150**	**1.000**	**0.196**	**0.171**	**0.229**	**1.000**
		VSM	0.102	0.121	0.088	**1.000**	0.109	0.147	0.086	**1.000**
		CL-ASA	0.100	0.146	0.076	**1.000**	0.085	0.137	0.062	**1.000**
		CL-ESA	0.092	0.107	0.081	**1.000**	0.078	0.122	0.057	**1.000**
		CL-C3G	0.072	0.104	0.054	**1.000**	0.042	0.053	0.035	**1.000**
	(b)	S2Net	**0.091**	0.141	**0.067**	**1.000**	**0.115**	**0.173**	**0.086**	**1.000**
		BAE	0.085	**0.191**	0.055	**1.000**	0.088	0.113	0.072	**1.000**
		XCNN	0.077	0.116	0.058	**1.000**	0.085	0.160	0.058	**1.000**
	(c)	CWASA (XCNN)	0.117	0.143	0.099	**1.000**	**0.168**	**0.212**	**0.140**	**1.000**
		CWASA (S2Net)	**0.124**	**0.147**	**0.107**	**1.000**	0.139	0.184	0.111	**1.000**
		CWASA (BAE)	0.081	0.131	0.059	**1.000**	0.056	0.095	0.040	**1.000**
	(d)	KBSim (S2Net)	**0.175**	0.174	**0.178**	**1.000**	**0.214**	0.248	**0.189**	**1.000**
		KBSim (VSM)	0.166	**0.182**	0.153	**1.000**	0.198	**0.254**	0.163	**1.000**
		KBSim (BAE)	0.159	0.165	0.154	**1.000**	0.199	0.245	0.168	**1.000**
		KBSim (XCNN)	0.151	0.173	0.134	**1.000**	0.197	0.241	0.167	**1.000**
Translated automatic obfuscation	(a)	CL-KGA	**0.706**	**0.785**	**0.641**	**1.000**	**0.607**	**0.684**	**0.547**	**1.000**
		VSM	0.603	0.673	0.553	1.011	0.445	0.562	0.391	1.053
		CL-ASA	0.552	0.736	0.479	1.077	0.439	0.652	0.373	1.125
		CL-ESA	0.503	0.571	0.479	1.052	0.288	0.431	0.247	1.137
		CL-C3G	0.398	0.602	0.347	1.160	0.122	0.343	0.085	1.183
	(b)	S2Net	**0.550**	0.784	**0.471**	**1.106**	**0.406**	**0.719**	**0.326**	1.164
		BAE	0.470	0.781	0.386	1.154	0.224	0.520	0.158	**1.132**
		XCNN	0.412	**0.791**	0.331	1.205	0.289	0.715	0.210	1.191
	(c)	CWASA (XCNN)	**0.650**	0.732	**0.585**	**1.001**	**0.525**	**0.651**	**0.460**	**1.040**
		CWASA (S2Net)	0.648	**0.739**	0.579	1.002	0.436	0.626	0.378	1.123
		CWASA (BAE)	0.377	0.581	0.316	1.131	0.255	0.517	0.190	1.134
	(d)	KBSim (XCNN)	**0.730**	**0.852**	0.639	**1.000**	**0.647**	**0.815**	0.537	**1.000**
		KBSim (S2Net)	0.709	0.773	**0.656**	**1.000**	0.608	0.691	0.543	**1.000**
		KBSim (VSM)	0.707	0.770	0.655	**1.000**	0.610	0.703	0.540	**1.000**
		KBSim (BAE)	0.708	0.781	0.649	**1.000**	0.609	0.687	**0.547**	**1.000**

Table 5.8. ES-EN and DE-EN performance analysis in terms of the obfuscation type, plagdet (Plag), precision (Prec), recall (Rec) and granularity (Gran).

In this last experiment we analyse the performance of the models considering the type of plagiarism case for CL plagiarism detection. As in Section 5.2.2.1, we divide the plagiarism cases depending on the obfuscation type employed to generate the case, and on the basis of the case length. We only note the most relevant differences among the models with respect to the general plagiarism detection analysis of the previous subsection.

In Table 5.8, depending on the obfuscation type, we note again the additional difficulty to detect cases with manual obfuscation. In this exper-

Type of obfuscation		Model	Spanish-English (ES-EN)				German-English (DE-EN)			
			Plag	**Prec**	**Rec**	**Gran**	**Plag**	**Prec**	**Rec**	**Gran**
Long length cases	(a)	CL-KGA	0.410	0.418	**0.404**	**1.000**	**0.383**	0.408	**0.361**	**1.000**
		CL-ASA	**0.411**	**0.535**	0.375	1.106	0.339	**0.513**	0.299	1.168
		VSM	0.399	0.416	0.391	1.016	0.320	0.386	0.300	1.077
		CL-ESA	0.351	0.388	0.352	1.076	0.220	0.329	0.198	1.176
		CL-C3G	0.299	0.467	0.269	1.207	0.090	0.275	0.064	1.227
	(b)	S2Net	**0.411**	0.587	**0.368**	**1.145**	**0.322**	0.589	**0.269**	**1.212**
		XCNN	0.327	**0.655**	0.271	1.253	0.230	**0.619**	0.170	1.234
		BAE	0.369	0.631	0.314	1.200	0.178	0.449	0.127	1.159
	(c)	CWASA (XCNN)	0.407	0.420	0.397	1.002	**0.361**	0.430	**0.337**	**1.063**
		CWASA (S2Net)	**0.413**	0.432	**0.398**	**1.003**	0.323	**0.470**	0.294	1.173
		CWASA (BAE)	0.283	**0.433**	0.250	1.171	0.211	0.405	0.164	1.158
	(d)	KBSim (XCNN)	**0.458**	**0.501**	0.423	**1.000**	**0.435**	**0.519**	**0.375**	**1.000**
		KBSim (BAE)	0.439	0.436	**0.442**	**1.000**	0.389	0.401	0.378	**1.000**
		KBSim (VSM)	0.438	0.439	0.438	**1.000**	0.388	0.408	0.370	**1.000**
		KBSim (S2Net)	0.432	0.437	0.429	**1.000**	0.376	0.389	0.365	**1.000**
Medium length cases	(a)	CL-KGA	**0.261**	**0.259**	**0.263**	**1.000**	**0.245**	**0.265**	**0.229**	**1.000**
		VSM	0.205	0.215	0.196	**1.000**	0.155	0.183	0.134	**1.000**
		CL-ASA	0.174	0.224	0.142	**1.000**	0.149	0.204	0.117	**1.000**
		CL-ESA	0.164	0.174	0.156	**1.000**	0.092	0.113	0.078	**1.000**
		CL-C3G	0.131	0.175	0.105	**1.000**	0.041	0.070	0.029	**1.000**
	(b)	S2Net	**0.176**	0.240	**0.139**	**1.000**	**0.135**	**0.217**	**0.098**	**1.000**
		XCNN	0.127	0.221	0.089	**1.000**	0.096	0.204	0.063	**1.000**
		BAE	0.148	**0.241**	0.107	1.000	0.072	0.126	0.051	**1.000**
	(c)	CWASA (XCNN)	**0.221**	0.223	**0.218**	**1.000**	**0.194**	**0.221**	**0.173**	**1.000**
		CWASA (S2Net)	0.219	**0.226**	0.212	**1.000**	0.155	0.196	0.129	**1.000**
		CWASA (BAE)	0.115	0.157	0.090	**1.000**	0.068	0.107	0.050	**1.000**
	(d)	KBSim (XCNN)	**0.274**	**0.293**	0.257	**1.000**	**0.261**	**0.307**	0.228	**1.000**
		KBSim (S2Net)	0.258	0.254	0.263	**1.000**	0.260	0.278	0.245	**1.000**
		KBSim (VSM)	0.260	0.256	**0.265**	**1.000**	0.254	0.270	0.240	**1.000**
		KBSim (BAE)	0.260	0.259	0.263	**1.000**	0.251	0.269	**0.267**	**1.000**
Short length cases	(a)	CL-KGA	**0.015**	**0.011**	**0.023**	**1.000**	**0.014**	**0.010**	**0.021**	**1.000**
		VSM	0.009	0.006	0.014	**1.000**	0.007	0.005	0.011	**1.000**
		CL-ESA	0.009	0.006	0.015	**1.000**	0.005	0.003	0.008	**1.000**
		CL-ASA	0.006	0.005	0.009	**1.000**	0.006	0.005	0.009	**1.000**
		CL-C3G	0.005	0.004	0.006	**1.000**	0.004	0.003	0.005	**1.000**
	(b)	S2Net	**0.008**	**0.007**	**0.010**	**1.000**	0.008	0.006	**0.010**	**1.000**
		XCNN	0.006	0.006	0.006	**1.000**	**0.009**	**0.009**	0.009	**1.000**
		BAE	0.003	0.003	0.004	**1.000**	0.005	0.004	0.007	**1.000**
	(c)	CWASA (XCNN)	0.011	0.008	**0.019**	**1.000**	**0.009**	**0.007**	**0.015**	**1.000**
		CWASA (S2Net)	**0.012**	**0.009**	0.018	**1.000**	0.007	0.005	0.011	**1.000**
		CWASA (BAE)	0.005	0.003	0.007	**1.000**	0.004	0.004	0.005	**1.000**
	(d)	KBSim (S2Net)	**0.018**	**0.014**	**0.027**	**1.000**	**0.017**	**0.013**	**0.026**	**1.000**
		KBSim (XCNN)	**0.018**	**0.014**	0.025	**1.000**	0.014	0.011	0.021	**1.000**
		KBSim (BAE)	0.017	0.013	0.023	**1.000**	0.014	0.011	0.021	**1.000**
		KBSim (VSM)	0.017	0.013	0.023	**1.000**	0.014	0.011	0.021	**1.000**

Table 5.9. ES-EN and DE-EN performance analysis in terms of plagiarism case length, plagdet (Plag), precision (Prec), recall (Rec) and granularity (Gran).

iment there is an additional handicap compared to the experiment of Section 5.2.2.1: the detailed analysis and proprocessing of Algorithm 3.1. In the statistics of Table 5.2 we observe ten times less cases with manual obfuscation. In addition, we verified that most of them are short length cases, which are generally covered by a single text fragment (see Section 3.4 for more information about the size of fragments, the division of documents with slide window and Algorithm 3.1). Therefore, Algorithm 3.1 fails in detecting most of this type of cases: it needs offset overlaps of at least two detections in the five most similar fragments. Despite this fact, we observe that KBSim is the best detector independently of the type of obfuscation, with special mention to KBSim (S2Net) in DE-EN for manual obfuscation cases of plagiarism. In contrast, the KBSim (XCNN) model obtains the best results for automatic obfuscation cases. Since these cases are more numerous, this model obtains the overall best results in Section 5.2.2.2. We also note that the 1.0 value of granularity is normal when detecting cases with large distance between them in the document. Hence the high occurrence in the tables.

In Table 5.9 we can see the results depending on the case length. It is interesting to see that CL-ASA outperforms CL-KGA for ES-EN long cases. The alignment model included in CL-ASA eases the detection of long cases — mostly composed by automatic translated cases — and increases the precision. In fact, this model was originally meant for detecting verbatim plagiarism cases. In contrast, we observe that the model does not excel for short cases of plagiarism, and is outperformed by CL-ESA. Overall, with exception of short DE-EN cases, KBSim (XCNN) obtains the best results in all the experiments. We also note its difference in performance for longer cases of plagiarism compared to KBSim (S2Net). These facts show the versatility of the KBSim (XCNN) model for the task of CL plagiarism detection.

5.2.2.3 Study of the Computational Efficiency

In this section we study the computational efficiency of the models. We specially focus on the models not studied in Section 3.5. In Table 5.10 we show the time necessary to transform the texts to the space of the models (indexing), and the time for measuring the similarity once that transformation is done. We use an Intel-i5@2.8Ghz with 16 GB of RAM to perform

System		Text indexing (texts/second)	Text similarity (texts/second)
(a)	CL-KGA	3	281
	VSM	2,083	2,291
	CL-ASA	1,741	3,627
	CL-ESA	282	1,826
	CL-C3G	3,127	2,619
(b)	XCNN	390	8,599
	S2Net	433	8,599
	BAE	380	8,598
(c)	CWASA (XCNN)	497	3,824
	CWASA (S2Net)	500	3,812
	CWASA (BAE)	510	3,784
(d)	KBSim (VSM)	3	287
	KBSim (XCNN)	3	278
	KBSim (S2Net)	3	283
	KBSim (BAE)	3	278

Table 5.10. Comparison of time required to index and estimate similarity between texts. Results are estimated as the average for processing all the ES-EN partition.

these tests over the complete ES-EN partition.[11] As we can see, there is more variability in the indexing time. The CL-KGA and KBSim knowledge graph-based models are slow due to the time required to search paths in the BabelNet multilingual semantic network. Note that it contains more than 9 million of concepts and more than 262 million of relations among them. More simple approaches such as VSM are recommended for fast document indexing. However, text indexing is usually part of the preprocessing step, being the indexing of the new documents needed only once. The XCNN, S2Net, and BAE models offer an acceptable indexing time and excel at text similarity level. Thanks to the cosine similarity between low-dimensional vectors, these three approaches are the fastest ones. Finally, the efficiency of the CWASA model make its similarity calculation also fast. This, together with its good performance in the experiments of this work, highlights its potential for large scale systems.

[11] Since the time for indexing a text does not depend on its language but its length, we only measure average times using the ES-EN partition.

5.2.3 DISCUSSION

The purpose of our research on the task of cross-language plagiarism detection was to create new and more capable models for cross-language similarity analysis. We employed the BabelNet multilingual semantic network to generate knowledge graphs as a cross-language and cross-domain representation of text and its meaning. In addition, we studied the most relevant characteristics of this representation (cf. Section 3.3.4): word sense disambiguation, vocabulary expansion, and representation by similarities with a collection of concepts.

In this section we extended the evaluation of Chapter 3. We included the following new aspects: (i) the comparison with reference cross-language distributed representation-based models; (ii) the evaluation of the KBSim model that combines knowledge graphs with document vectors; (iii) the combination of KBSim with cross-language distributed representation-based models; and (iv) the analysis of the models as function of the obfuscation type employed to generate the plagiarism case, and on the basis of the case length.

In this study we provided with sufficient evidences to prove that knowledge graphs have a strong potential as a cross-language representation of text and its meaning. In addition, after analysing the results of the evaluations of Section 3.5 and 5.2.2, including a comparison with the reference models for this task, we conclude that knowledge graph-based models offer state-of-the-art performance for the task of cross-language plagiarism detection. However, we have not found statistical significant differences between the CL-KGA and KBSim models in this task. Therefore, despite the general higher values of KBSim (XCNN), there were no differences between the KBSim model combined with the traditional VSM and the distributed representation-based ones. In consequence, in Section 5.3 we will not explore this line in the tasks of cross-language document retrieval and categorization in the extended evaluation of the Chapter 4.

With respect to the combination of knowledge graphs and distributed representations, in Section 1.1 we showed the relationship between these two representations with an example that also highlights the relationship with how our mind represents knowledge. We note that, despite the KBSim model combined with distributed representations was not statistically superior, all the knowledge graph-based models in this evaluation used the distributed concept weighting scheme proposed in Section 3.3.3.2. Therefore, all the

models with statistical significant differences on top of our ranking (cf. Figure 5.3), employ the combination of knowledge graphs and distributed representations to excel over the rest.

5.3 Cross-language Document Retrieval and Categorization Results

In Chapter 4 we presented our KBSim model for cross-language similarity analysis (cf. Section 4.3). That chapter also studied the performance of the model in the tasks of cross-language document retrieval and categorization comparing it with the state of the art obtaining good results (cf. Section 4.6). However, KBSim is a modified version of our CL-KGA model (cf. Section 3.4) and they have not been compared in these two tasks. Similarly, we have not compared KBSim with the results of its vector component separately. In addition, the original KBSim employed the weighting scheme of relations of BabelNet 1.0. That scheme was based on Dice's coefficient overlaps between gloss and Wikipedia hyperlink information. In Section 3.3.3.2 we proposed a new weighting scheme for knowledge graph semantic relations that offered better performance for cross-language plagiarism detection (cf. Section 3.5 and 5.2.2). Therefore, in order to complement the study with knowledge graphs of Chapter 4, in Section 5.3.1 we show these results together with the ones reported in Section 4.6. Finally, in Section 5.3.2 we discuss the results of this study and close the cross-language part of this thesis.

5.3.1 Complementary Evaluation of Cross-language Document Retrieval and Categorization

In this section we complete the evaluation of Section 4.6. We employ the same datasets, evaluation measures, and state of the art models. In addition, we include some additional models in order to show the complete knowledge graph picture at task level. First we show our new experiments in cross-language document retrieval. Next, in Section 5.3.1.2 we show our new experiments in cross-language document categorization.

5.3.1.1 Experiments in Cross-language Document Retrieval

For cross-language document retrieval we compare the following models. As baselines we selected the CosSim_E, CosSim_{BN} (cf. Section 4.6.1.1),

CL-C3G, CL-ESA, and VSM models.[12] We also employ the distributed representation-based[13] CL-LSI, CCA, OPCA, and S2Net models.[14] Finally, we show the results of the original KBSim model, the CL-KGA model, and the results of these models using our Distributed Concept Weighting (DCW) scheme: KBSim (DCW) and CL-KGA (DCW). We evaluate all the models over the test partition of the Wikipedia dataset described in Section 4.6.1.1 that contains 8,675 documents. We show the accuracy of ranking the real Wikipedia comparable document across languages as the most similar one — which is equivalent to estimate the R@*1* —, and the Mean Reciprocal Rank (MRR) of Eq. 5.6. We do not note again the insights of Section 4.6 and focus on the study of the results related to our knowledge graph models.

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}, \tag{5.6}$$

where $|Q|$ is the number of documents to rank and the function rank_i returns the document retrieval ranking of the document i.

The Table 5.11 shows the results[15] of this evaluation. The use of our bilingual vector representation makes VSM to excel over other vector representations such as CosSim_{BN}. However, despite VSM also outperforms CL-KGA, we can see that both offer an average performance. These results are in line with the results that we obtained during the development of our KBSim model, and were the ones that motivated the study of the CL-KGA and VSM combination and consequently, the KBSim proposal. We note that the dataset employed is distributed online using the TF-IDF weighting preprocessing and contains only 20k different words.[16] This constraint makes

[12] As we did in Section 5.2.2, VSM refers to the vector component of KBSim described in Section 4.4.

[13] Despite in the publication included as Chapter 4 of this thesis we refer to this type of models as linear projection ones, their nature is based on the distributional semantics theory (see Section 1.1). Therefore, for being consistent with the rest of this thesis, in this section we refer to them as distributed representation-based models.

[14] Since the BAE and the XCNN models did not excel over other distributed representation-based ones such as S2Net (cf. Section 5.2.2), we do not include them in this extended evaluation.

[15] In this study, statistically significant results according to a one-tailed χ^2 test with Yates' correction ($p < 0.05$) are highlighted in bold.

[16] Since the dataset is represented by the TF-IDF of isolated words, we cannot detect multi-words in order to map them to the most appropriated knowledge graph concepts.

Model	Accuracy	MRR
KBSim (DCW)	**0.754**	0.800
S2Net	**0.745**	0.797
KBSim	0.734	0.775
OPCA	0.726	0.773
VSM	0.709	0.763
CL-KGA (DCW)	0.707	0.759
CosSim$_E$	0.703	0.747
CosSim$_{BN}$	0.703	0.755
CCA	0.689	0.738
CL-KGA	0.684	0.736
CL-LSI	0.530	0.613
CL-ESA	0.266	0.330
CL-C3G	0.251	0.302

Table 5.11. Test results for comparable document retrieval in Wikipedia. This table updates the results of Table 4.1.

it more difficult to our CL-KGA model to generate representative document graphs. In consequence, more simple representations such as the CosSim$_{BN}$ and the VSM models are able to outperform it. In contrast, KBSim combines VSM and CL-KGA exploiting the amount of information in the knowledge graphs as interpolation basis (cf. Section 4.5), and it is able to work at par than the state of the art. Finally, we highlight the improvements obtained when the DCW is employed. The CL-KGA (DCW) model is at par with VSM. Moreover, KBSim (DCW) outperforms S2Net and they both are statistically significant in top of these results. All these facts manifest the quality of the DCW scheme and the KBSim model, as well as the potential of knowledge graphs for cross-language document retrieval.

5.3.1.2 Experiments in Cross-language Document Categorization

For this task we compare the same additional models included in the previous section — VSM, CL-KGA, CL-KGA (DCW), and KBSim (DCW) — with those employed in the original evaluation of Section 4.6.2: the CosSim$_E$ and CosSim$_{BN}$ baselines, the CL-LSI, CCA, OPCA and Full MT distributed rep-

Model	EN News Accuracy	ES News Accuracy
KBSim (DCW)	**0.839**	**0.722**
KBSim	0.819	0.700
Full MT	**0.848**	0.648
VSM	0.809	0.681
CL-KGA (DCW)	0.806	0.681
CosSim_{BN}	0.802	0.674
CL-KGA	0.795	0.669
OPCA	**0.841**	0.595
CCA	**0.839**	0.532
CL-LSI	**0.840**	0.510
CosSim_{E}	0.805	0.448

Table 5.12. Test results for cross-language text categorization. This table updates the results of Table 4.2.

resentation models, and our KBSim model.[17] We show the accuracy of classification of the two cross-language test partitions defined in Section 4.6.2: EN News and ES News, with 1,875 and 12,342 documents, respectively. We do not note again the insights of Section 4.6.2 and focus on the study of the results related to our knowledge graph models.

In Table 5.12 we show the results of categorization. Most of the insights about knowledge graphs of Section 5.3.1.1 persist in this evaluation. The VSM model is superior to other vector representations such as CosSim_{BN}. It outperforms CL-KGA and marginally also CL-KGA (DCW). However, when knowledge graphs and VSM are combined together in KBSim, the performance is notably superior. The comment about the TF-IDF text preprocessing also persists here. The Multilingual Reuters Collection is distributed online in that format, which makes it more difficult to create representative knowledge graphs. Finally, we note again that the results when DCW is employed are higher. KBSim (DCW) outperforms the original KBSim and is at par with the statistically significant models on top of the EN news partition. In addition, it is statistically the best model in the ES news partition. This result is specially relevant if we consider that the ES news partition is much

[17] We note that all our knowledge graph models use the *k*-NN process employed in Section 4.6.2 with KBSim for categorization.

larger than the EN one. Note that larger test collections are able to provide with more accurate results in terms of significance.

5.3.2 DISCUSSION

In Chapter 4 we introduced KBSim, a new model of cross-language similarity analysis that improves CL-KGA including a vector component that counterbalances possible knowledge graph shortcomings. We applied this new model to cross-language document retrieval and categorization and compared it with the state of the art. That included the comparison with the reference distributed representation models and other models also popular in cross-language plagiarism detection. Moreover, in the extended evaluation of this section we independently analysed the performance of the KBSim components: knowledge graphs and multilingual vectors. Finally, we evaluated the performance of KBSim and CL-KGA employing our new distributed concept weighting scheme.

Results in the tasks of cross-language document retrieval and categorization showed that KBSim using the new weighting scheme is much better than its independent components. They also showed that the model is able to work at par or better than the state of the art. All these facts manifest the quality of the DCW scheme and the KBSim model, as well as the potential of knowledge graphs for cross-language document retrieval and categorization.

To conclude the cross-language part of this thesis, we would like to summarise some important aspects. We studied our knowledge graph-based models in three representative tasks of the cross-language scenario: cross-language plagiarism detection, document retrieval, and categorization. We also analysed the characteristics that make knowledge graphs adequate for these tasks. The results of the CL-KGA and KBSim models compared to several reference models provided with state-of-the-art performance and also showed interesting insights about the models. These results were obtained employing several datasets and language pairs. All these facts prove the quality and potential of knowledge graphs as a cross-language representation of text and meaning. In addition, they show the versatility of the developed models, specially KBSim, for cross-language NLP and IR tasks.

5.4 Knowledge Graphs in Other NLP Tasks

In this thesis we showed how we successfully employ knowledge graphs as a cross-language and cross-domain representation of text and its meaning. We thoughtfully studied the tasks of single- and cross-domain polarity categorization, and the cross-language tasks of plagiarism detection, document retrieval and categorization. However, we are interested in studying if knowledge graphs can be used as a general representation of text and its meaning for other non-cross-language or cross-domain NLP tasks.

In this section we detail our experiments and results with knowledge graphs in the tasks of community question answering, native language identification, and language variety identification. We perform the question answering task with a model that could be considered an extension of KBSim in order to take advantage of training data and several similarity models. The native language and the language variety identification tasks are conducted with our KBSim (DCW) model for classification. That part of the study aims to bridge these two close language identification tasks which until now have been addressed separately by two different research communities.

The rest of this section is structured as follows. First, in Section 5.4.1 we describe our participation in the SemEval 2016 community question answering shared task. Next, in Section 5.4.2 we study together the native language and the language variety identification tasks. Finally, in Section 5.4.3 we draw some conclusions on this part of the thesis.

5.4.1 Community Question Answering

The SemEval 2016 Task 3 on Community Question Answering (CQA) (Nakov et al., 2016) is focused on automatically identifying good answers to new questions by searching a discussion forum for similar questions and by identifying, among their answers, those that answer the new question. Last edition of this task (Nakov et al., 2015) classified the answers as *good*, *bad*, or *potentially relevant* with respect to one question. However, this type of automatic systems are imperfect from the user perspective. Therefore, the 2016 edition is focused on ranking subtasks that employ the probability or relevance between the questions and comments.

The SemEval 2016 Task 3 is composed by four subtasks:

- **Subtask A) English question-comment similarity ranking.** Given one question and its 10 comments, rank them in function of their relevance.
- **Subtask B) English question-related question similarity ranking.** Given one new question, rank 10 related questions according to their relevance with respect to the original one.
- **Subtask C) English question-external comment similarity ranking.** Given one new question and 10 related questions with their 10 respective comments, rank these 100 comments according to the original question. This subtask tries to cover the complete task of CQA.
- **Subtask D) Arabic question-related question with correct answer re-ranking.** This subtask simplifies Subtask C and require users to rank 30 related questions — with one comment each one — respect to one original question.

Automatic question answering has been a popular interest of research in NLP from the beginning of the Internet (Rosso et al., 2012). The use of BOW representations allowed to correctly answer 60% of the questions of the first large-scale question answering evaluation at the TREC-8 Question Answering track (Voorhees, 1999). More complex systems used inference rules to connect expressions between questions and answers (Lin and Pantel, 2001). Similarly, Ravichandran and Hovy (2002) employed bootstrapping to generate surface text patterns in order to successfully answer questions. Other works such as Buscaldi et al. (2010) are based on the redundancy of n-grams in order to find one or more text fragments that include tokens of the original question and the answer. Jeon et al. (2005) studied the semantic relatedness between texts for question answering. They used translation obfuscation to paraphrase the text and to detect which terms are closer in meaning. Probabilistic topic models have been also useful for detecting the semantics in this task. Celikyilmaz et al. (2010) used LDA for representing questions by means of latent topics.

With respect to the previous edition of the SemEval CQA task, several teams experimented with complex solutions that included meta-learning, external resources, and linguistic features such as syntactic relations and distributed word representations. Similarly to the model described in this section, the highest performing approach employed a combination of lexical and

semantic-based similarity measures (Tran et al., 2015). Another interesting approach, Hou et al. (2015), included textual features — word lengths and punctuation — in addition to syntactical-based features — POS tags.

In this section we detail our participation in the three English-related Task 3 subtasks on CQA (subtasks A, B, and C). We describe an approach that is based on our KBSim (cf. Section 4.5) model. The latter employs knowledge graphs and document vectors to measure similarity at semantic and lexical level, respectively, and combines these values with a dynamic interpolation. Note that KBSim does not use any training data. In contrast, here we employ knowledge graphs and document vectors at monolingual level along with other additional representations for a similarity ranking task that provides with training data. Our new model is designed to take advantage of this shared task training data in order to learn the interpolation thresholds of all the similarity models employed. We first represent each instance to rank — question versus (vs.) comments, question vs. related questions, or question vs. comments of related questions — with a set of similarities computed at two different levels: lexical and semantic. Similarly to our KBSim model, this representation allows us to estimate the relatedness between text pairs in terms of what is explicitly stated and what it means. Our lexical similarities employ representations such as word and character n-grams, and BOW. The semantic similarities include the use of our CWASA model (cf. Section 5.2.1.4), distributed representations of text, knowledge graphs, and frames from the FrameNet lexical database (Baker et al., 1998).

We first detail our model for this CQA shared task in Section 5.4.1.1. Next, in Section 5.4.1.2 we analyse its results and draw some conclusions about this task.

5.4.1.1 Lexical and Semantic-based Community Question Answering

In this section we describe the model that we designed for this CQA task. Similarly to our KBSim model, this new model for CQA exploits both the verbatim and the contextual similarities between texts, i.e., questions and comments.[18]

[18]We note that all our features are similarity scores obtained with different text similarity measures. More details and examples can be found in the respective papers or sections of this thesis.

Lexical Features

The lexical features that we employ are the following:

- **Cosine Similarity.** We use cosine similarity to measure lexical similarity between two text snippets. We calculate cosine similarity based on word n-grams (n=1,2), character 3-grams and TF-IDF scores of words.

- **Word Overlap.** We use the count of common words between two texts. This count is normalized by the length.

- **Noun Overlap.** We use NLTK[19] to part-of-speech tag the text and compute the normalized count of overlapping nouns in two texts as a similarity measure.

- **N-gram Overlap.** We compute the normalized count of common n-grams (n=1,2,3) between two texts.

Semantic Features

The semantic features that we employ are the following:

- **Distributed representations of texts.** We use the continuous Skip-gram model (see Section 3.3.3.2) of the word2vec toolkit to generate distributed representations of the words of the complete English Wikipedia.[20] Next, for each text, e.g. question or comment, we average its word vectors in order to have a single representation of its content as this setting has shown good results in other NLP tasks (e.g. for language variety identification (Franco-Salvador et al., 2015c) and discriminating similar languages (Franco-Salvador et al., 2015d) (cf. Section 5.4.2.1)). Finally, the similarity between texts, e.g. question vs. comment, is estimated using the cosine similarity.

[19] `http://www.nltk.org/`

[20] We use 200-dimensional vectors, context windows of size 10, and 20 negative words for each sample.

- **Distributed word alignments.** We use our CWASA model (cf. Section 5.2.1.4) to measure the similarity by double-direction aligning distributed word representations of texts.

- **Knowledge graphs.** We use our CL-KGA model[21] (cf. Section 3.4) to measure the similarity of texts at knowledge graph space.

- **Common frames.** We use Framenet (Baker et al., 1998) to extract the frames associated with the lexical items in the text. For each frame present in the text, we calculate the common lexical items between sentences associated with this frame. The goal is to allow inference of similarity at the level of semantic roles.

As additional feature, for Subtasks B and C we also use the ranking provided by the Google search engine for the questions related to the original questions.

Data Representation and Ranking

Due to the representation of questions (composed by *subject* and *body* fields) and answers (a *comment* field) we adapt our system for the different English subtasks:

- **Subtask A (question-comment similarity ranking):** we use the aforementioned similarity-based features at three levels: question subject vs. comment, question body vs. comment, and full question vs. comment.

- **Subtask B (question-related question similarity ranking):** for this subtask we measure the similarities at body, subject, and full question level.

- **Subtask C (question-external comment similarity ranking):** we employ all the features of Subtasks A and B, plus the similarities of the

[21] We do not use KBSim to generate features because it is a combination of two representations and their similarity scores. In this task we want to automatically learn the combination of several representations by means of machine learning techniques. However, the proposed model follows the KBSim nature of combining more than one similarity measure.

original question — subject, body, and full levels — with the related question comments.

In order to rank the questions and comments, we select a variant of SVM optimized for ranking problems: SVM$_{rank}$ (Joachims, 2002).[22] We use a linear kernel and optimize the SVM cost factor parameter using Bayesian optimizations[23] (Snoek, 2013). In addition to the ranking, the task requires also to provide with a label for each instance that reflects if the question or comment is relevant to the compared question. For each subtask we optimize a threshold to determine the relevance of each instance that is based on our predicted relevance ranking. In other words, we binarize our ranking.[24]

5.4.1.2 Results and Discussion of the Community Question Answering Study

In this section we study the results of our approach in the SemEval 2016 Task 3 on CQA.

Methodology

For evaluating our approach we use the CQA-QL English corpus (version 3.2) (Nakov et al., 2016) provided for the SemEval 2016 Task 3 on CQA.[25] In Table 5.13 we can see the statistics of the corpus.

We compare the results of our approach with those provided by the random baseline and the Google search engine when ranking the questions and comments.[26] We also show the results of the best performing system for each

[22]Preproscessing steps include stopword removal, lemmatization, and stemming. However, for the distributed representation and knowledge graph-based features we do not employ stemming. These decisions are motivated for performance reasons during our prototyping.

[23]We used the Spearmint toolkit: `https://github.com/HIPS/Spearmint`

[24]Note that each subtask originally allowed to submit three runs per team. For the sake of simplicity, in our tables we show only the one with better results.

[25]Despite the Subtask A allows to use the corpus of the SemEval 2015 CQA task, we did not observe improvements when using it. Therefore, we just use the 2016 corpus. In addition, in this study we combine the two available 2016 training partitions.

[26]Some considerations about the evaluation: these subtasks employed binary classification. At testing time, *Bad* and *PotentiallyUseful* are both considered *false*. The same occurs with *PerfectMatch* and *Relevant*, which are both considered *true*. In addition, following the rules of the task, the employed measures use only the top 10 ranked instances.

	Train.	Dev.	Test
Original questions			
Total	267	50	70
Related questions			
Total (Subtask B)	2,669	500	700
PerfectMatch	235	59	81
Relevant	848	155	152
Irrelevant	1,586	286	467
Related comments			
wrt *Original question*			
Total (Subtask C)	26,690	5,000	7,000
Good	2,837	345	654
Bad	21,473	4,061	5,943
PotentiallyUseful	2,380	594	403
wrt *Related question*			
Total (Subtask A)	17,900	2,440	3,270
Good	6,651	818	1,329
Bad	8,139	1,209	1,485
PotentiallyUseful	3,110	413	456

Table 5.13. Statistics of the CQA-QL version 3.2 English corpus.

subtask.[27] The official measure of the task is the Mean Average Precision (MAP) (see Eq. 5.7), but we include also two alternative ranking measures: Average Recall (AvgRec) (see Eq. 5.9) and Mean Reciprocal Rank (MRR) (cf. Eq. 5.6). In addition, we include four classification measures: Accuracy (acc.), Precision (P), Recall (R), and F1-measure (F1). Next we detail the measures not defined in the previous chapters and sections:

$$\text{MAP} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \text{AvgPrec}(i), \tag{5.7}$$

[27] Since their development set results are not available, we only show their test set results.

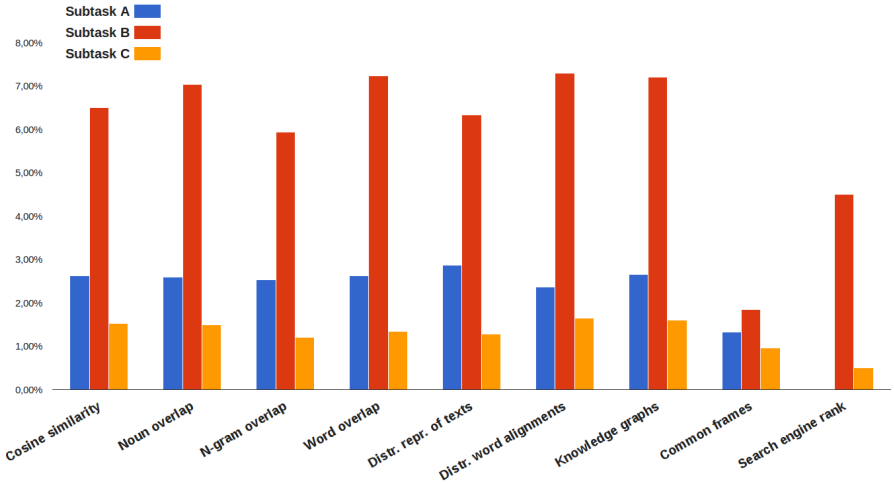

Figure 5.4. IGR of our similarity-based features for the three English CQA subtasks.

where Q is the number of documents to rank and $\text{AvgPrec}(i)$ is the average precision of the top i documents. It is estimated as follows:

$$\text{AvgPrec}(n) = \frac{1}{\min(m,n)} \sum_{k=1}^{n} P(k), \tag{5.8}$$

where $P(k)$ means the precision at cut-off k in the item list, i.e., the ratio of number of users followed up to the position k over the number k. $P(k)$ equals 0 when the k-th item is not followed upon recommendation and m is the number of relevant documents. If the denominator is zero, $\text{AvgPrec}(n)$ is set to zero. Similarly, the average recall is estimated as follows:

$$\text{AvgRec}(n) = \frac{1}{\min(m,n)} \sum_{k=1}^{n} R(k). \tag{5.9}$$

		Ranking measures			Classification measures			
	Model	**MAP**	**AvgRec**	**MRR**	**Acc.**	**P**	**R**	**F1**
Development set results								
(a)	Random baseline	0.456	0.654	0.535	0.433	0.344	**0.764**	0.475
	Search engine	0.538	0.728	0.631	n/a	n/a	n/a	n/a
(b)	Proposed approach	**0.630**	**0.811**	**0.722**	**0.672**	**0.510**	0.545	**0.527**
Test set results								
(a)	Random baseline	0.528	0.665	0.587	0.525	0.452	0.405	0.428
	Search engine	0.595	0.726	0.678	n/a	n/a	n/a	n/a
(b)	Proposed approach	0.676	0.795	0.771	0.624	0.541	0.501	0.520
(c)	Filice et al. (2016)	**0.792**	**0.888**	**0.864**	**0.751**	**0.770**	**0.553**	**0.644**

Table 5.14. Results of **Subtask A: English Question-Comment Similarity**. (a) Baselines; (b) proposed approach; (c) best performing subtask approach.

Results and Discussion

In order to study the relevance of our features, we show their Information Gain Ratio (IGR) (cf. Eq. 2.4) divided per task in Figure 5.4.[28] As we can see, the features are notably more informative for subtask B. That is in line with the results that we analyse later for the classification and ranking measures. Moreover, the comparison between the lexical and the semantic features seems to favour the latter ones, with exception of the common frames, that are not very informative. It is also interesting that the distributed word alignments of our CWASA model outperform on average the distributed representations of texts estimated by averaging word vectors. This highlights the potential of CWASA for similarity. In addition, we note that the search engine rank offers the lowest contribution for subtask B and C. Finally, we note that the similarity between knowledge graphs estimated with our CL-KGA model is on average the most informative one among all the employed features due to its stability.

[28] Since the features of our model represent similarity values at different levels (cf. Section 5.4.1.1), the total number of features for some subtasks is larger than 100. For the sake of simplicity, in this figure we ignored that level distinction and averaged that IGR values.

		Ranking measures			Classification measures			
	Model	**MAP**	**AvgRec**	**MRR**	**Acc.**	**P**	**R**	**F1**
Development set results								
(a)	Random baseline	0.559	0.732	0.622	0.488	0.443	**0.766**	0.562
	Search engine	0.713	0.861	0.766	n/a	n/a	n/a	n/a
(b)	Proposed approach	**0.755**	**0.910**	**0.817**	**0.758**	**0.714**	0.724	**0.719**
Test set results								
(a)	Random baseline	0.470	0.679	0.510	0.452	0.404	0.326	0.361
	Search engine	0.747	0.883	0.838	n/a	n/a	n/a	n/a
(b) & (c)	Proposed approach	**0.773**	**0.908**	**0.840**	**0.767**	**0.636**	**0.704**	**0.668**

Table 5.15. Results of **Subtask B: English Question-Question Similarity**. (a) Baselines; (b) proposed approach; (c) best performing subtask approach.

		Ranking measures			Classification measures			
	Model	**MAP**	**AvgRec**	**MRR**	**Acc.**	**P**	**R**	**F1**
Development set results								
(a)	Random baseline	0.138	0.096	0.160	0.284	0.070	**0.759**	0.128
	Search engine	0.306	0.346	0.360	n/a	n/a	n/a	n/a
(b)	Proposed approach	**0.383**	**0.421**	**0.425**	**0.897**	**0.252**	0.249	**0.250**
Test set results								
(a)	Random baseline	0.150	0.114	0.152	0.167	0.296	0.094	0.143
	Search engine	0.404	0.460	0.459	n/a	n/a	n/a	n/a
(b)	Proposed approach	0.434	0.480	0.484	**0.888**	**0.386**	0.327	0.354
(c)	Filice et al. (2016)	**0.556**	**0.63**4	**0.612**	0.834	0.322	**0.702**	0.442

Table 5.16. Results of **Subtask C: English Question-External Comment Similarity**. (a) Baselines; (b) proposed approach; (c) best performing subtask approach.

With respect to the ranking and classification subtasks, we can see the results of Subtask A (question-comment similarity ranking) in Table 5.14.[29] In terms of ranking measures, our system outperforms both the random and the search engine baselines. Using the development set, we observe a MAP improvement of 9.4% compared with the results obtained by the search engine. We can see similar differences with respect to the other two ranking

[29]The best results per partition and subtask are highlighted in bold. In addition, our percentage comparisons use always absolute values.

measures. Classification results are also superior. We obtain improvements in accuracy and F1 of 24.9% and 5.2%, respectively. These results show the potential of the selected lexical and semantic-based features for this subtask. However, our model is outperformed by the one of Filice et al. (2016). They learn semantic relations between questions and answers using kernels and previously-proposed features from Barrón-Cedeño et al. (2015). This contributes to strongly increase the precision and consequently the accuracy and all the ranking measures based on that metric.

Similar to Subtask A, the performance of our approach is also superior to the baselines in Subtask B (question-related question similarity ranking). As we can see in Table 5.15, using the development set, the improvement of MAP, AvgRec, and MRR is of 4.6%, 5%, and 6.4% respectively compared to the search engine baseline. In this case, the similarity between questions is easier to estimate — also for the baselines — and the improvements in performance are slightly reduced. With respect to the classification measures, we outperform the random baseline with 27.4% and 16.1 % of accuracy and F1-measure respectively. We note that our approach obtains the highest results among all the submitted systems — with considerable margin (1.04%) — for subtask B.

In Table 5.16 we can see the results of the Subtask C (question-external comment similarity ranking). In this case, we are ranking 100 comments (10 times more compared to the other subtasks). Therefore, this is the most difficult subtask. However, we obtain improvements in line with those reported for the other subtasks. Compared to the search engine baseline, the MAP, AvgRec, and MRR improves 8.7%, 8.5%, and 7.5% respectively when using the development partition. The accuracy and F1-measure improves 61.5% and 12.2% respectively. The largest number of comments to rank, and the use of top 10 results when measuring results, benefits our approach with this especially high difference in accuracy. However, Filice et al. (2016) obtains in general higher results than our approach. Their results in subtask A highlight the versatility of their model for CQA. Note that the comparison of results of all the submitted systems and task participants can be found in the task overview (Nakov et al., 2016).

In this part of the thesis we studied the performance of an approach that is based on our KBSim model but extended to employ training data and to combine more similarity models. Experimental results showed that our ap-

proach is able to outperform — with considerably differences — the random and Google search engine baselines in the three English subtasks. After the analysis of our results, we highlight that the combination of lexical and semantic-based features that we employ in this study offers a competitive performance for the CQA task. This is true also when comparing results with other task participants. Our approach obtains the highest results for subtask B, which is the most related task to classical text similarity. However, for the other two subtasks, we obtain an average ranking position. That difference in performance is logical if we consider the IGR values of Figure 5.4. Those values are three times higher for subtask B. Therefore, we conclude that our approach is more adequate for similarity tasks — as the plagiarism detection and document retrieval evaluated in this thesis — rather than question answering. In addition, on the contrary to KBSim and CL-KGA, this approach performs at monolingual level, which shows the potential of knowledge graphs for monolingual similarity tasks. Finally, as in Section 5.2 with cross-language plagiarism detection, we proved that knowledge graphs and distributed representations can be combined together to obtain additional improvements. This makes sense if we consider that both representations are related (see their relations, also with the cognitive science, in Section 1.1).

5.4.2 Bridging the Native Language and the Language Variety Identification Tasks

The task of Native Language Identification (NLI) is to determine the native language of the author (L_A) of a text which he wrote in another language (L_T); for example, deciding whether an English essay was written by a Chinese or German student. By contrast, Language Variety Identification (LVI) aims at classifying texts of different varieties of a single language; for example, distinguishing between American and British English. These two tasks are related to author profiling (Rangel et al., 2015), which identifies the linguistic profile of an author on the basis of its writing style, and determines author's traits such as gender, age, personality, or in this case native or language variety. That task is important in marketing, forensic, and educational applications. Despite the Internet destroyed frontiers among regions or traits, industry representatives are still very interested into author profiling segmentation — specially in social media. For example, when a new product is released, identifying the geographical distribution and nationality of the authors of the opinions may help to improve marketing campaigns.

The NLI task was introduced by Koppel et al. (2005), who employed features such as character and POS n-grams, as well as spelling and grammatical errors. The NLI Shared Task (Tetreault et al., 2013) allowed different systems to be directly compared for the first time. One of the submitted systems, based on string kernels, was subsequently refined to establish state of the art on several corpora (Ionescu et al., 2014). A recent study of the cross-corpora effects by Malmasi and Dras (2015b) demonstrates the existence of corpora-independent language transfer features.

The LVI task has attracted much interest in the last few years. Character n-grams and other features have been employed to identify varieties of Portuguese in news texts (Zampieri and Gebre, 2012), Arabic in blogs and forums (Sadat et al., 2014), and Spanish in tweets (Maier and Gómez-Rodríguez, 2014). Franco-Salvador et al. (2015c) proposed to use distributed representations of words to classify varieties of Spanish from blogs and journalistic texts (Franco-Salvador et al., 2015d). The Shared Task on Discriminating between Similar Languages (DSL) set the objective of classifying texts representing several sets of closely related languages and language varieties (Zampieri et al., 2014, 2015).

The starting point of this part of the thesis is the observation that, although the two tasks have been considered separately, and investigated by different teams of researchers, they share a common focus. In the NLI task, we are interested in the language of the author (L_A), which is different from the known language of the text (L_T). By contrast, in the LVI task, L_A is the same as L_T, which we want to determine. We posit that identifying L_A is the objective of both tasks, and that it can be achieved by identifying similar types of lexical and semantic characteristics of L_A.

In this section, we test our hypothesis by testing generic representations and models in both tasks without any task-specific adaptation. We employ knowledge graphs with our KBSim (DCW) model (cf. Section 4.5 and 5.3), henceforth referred to as KBSim. We also employ distributed representations and string kernels, the state of the art in LVI and NLI, respectively. Although the representations have been applied by previous works to either of the tasks, to the best of our knowledge, this is the first study to apply string kernels to LVI, distributed representations to NLI, and knowledge graphs to both tasks. We evaluate the methods on several datasets, including the data from the respective shared tasks.

The rest of this section is structured as follows. We first describe in Section 5.4.2.1 the methods that we employ to compare with our KBSim model. Next, in Section 5.4.2.2 we analyse the results and draw conclusions about this study.

5.4.2.1 Methods for NLI and LVI

String kernels and distributed representations have been shown to achieve high accuracy on the NLI and LVI tasks, respectively. In this section we briefly describe the two approaches that we compare with our KBSim model.

String Kernels

String Kernels (SK) are functions that measure the similarity of string pairs at lexical level. They have been successfully employed in text categorization (Lodhi et al., 2002), authorship attribution (Popescu and Grozea, 2012), and NLI (Ionescu et al., 2014).

In this section, we follow the formulation and implementation of Ionescu et al. (2014).[30] A simple measure of the similarity of two strings s,t is the number of shared substrings of length p. The general form of a p-grams kernel is:

$$\mathrm{k}_p(s,t) = \sum_{v \in L^p} f(\mathrm{num}_v(s), \mathrm{num}_v(t)),$$

where $\mathrm{num}_v(s)$ is the number of occurrences of string v as a substring of s over an alphabet L. Three variants of the kernel differ in the definition of the function $f(x,y)$: (i) $x \cdot y$ in the p-spectrum kernel; (ii) $\mathrm{sgn}(x) \cdot \mathrm{sgn}(y)$ in the p-grams presence bits kernel[31]; and (iii) $\min(x,y)$ in the p-grams intersection bits kernel. The kernels combine different n-gram sizes ($p = [5,8]$), and are normalized using the following formula:

$$\hat{\mathrm{k}}(s,t) = \mathrm{k}(s,t) / \sqrt{\mathrm{k}(s,s) \cdot \mathrm{k}(t,t)}.$$

We perform the classification with kernel discriminant analysis (Hastie and Tibshirani, 2003), which returns the eigenvector matrix U. We compute the feature matrices $Y = KU$ and $Y_\mathrm{t} = K_\mathrm{t}U$, where K and K_t are the training

[30] `http://string-kernels.herokuapp.com/`

[31] "sgn" is the sign function.

and test instance kernels. For each class c, we create the prototype Y_c as the average of all vectors of Y that correspond to the instances of class c. Finally, we classify each test instance by identifying the class of the prototype with the lowest mean squared error between $Y_t(i)$ and Y_c.

Distributed Representations

The use of Distributed Representations (DR) of words capture semantic relationships between words (Mikolov et al., 2013c). They have been successfully employed in classification tasks such as polarity classification (Le and Mikolov, 2014).

In this section we generate distributed representations of words using the continuous Skip-gram model and the negative sampling (Mikolov et al., 2013b) that we described in Section 3.3.3.2.[32]

We investigate two methods of deriving a distributed representation $\vec{e}$ of a document or instance d: (i) by averaging all distributed representations of words $\vec{w}_i \in d$, and (ii) via Sentence Vectors (SenVec) (cf. Section 3.3.3.2), which represent the entire sentence, and are derived using Skip-gram architecture.

5.4.2.2 Results and discussion of the NLI and LVI Study

In this section we compare the SK, DR, and KBSim models on the NLI and LVI tasks.

Datasets and Methodology

In this evaluation we measure the accuracy of classification in several datasets. The NLI task is represented by two datasets (see Table 5.17), which contain essays written by English language students. TOEFL11 (henceforth referred to as TOEFL) is the official dataset of the NLI Shared Task (Tetreault et al., 2013). It is composed by English essays of natives of the following languages: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. ICLE is a normalized version of a corpus

[32]We use 300-dimensional vectors, context window of size 10, and 20 negative words for each sample. Our preprocessing consists of tokenization, word lowercase, and removing tokens of length one.

that has been used for the NLI research in the past (Tetreault et al., 2012). In contains essays of native speakers of the following languages: Bulgarian, Chinese, Czech, French, Japanese, Russian, and Spanish. Since it lacks of a test set, we follow previous work by performing 5-fold cross validation instead.

The LVI task is also represented by two datasets. (i) DSLCC 2.0 is the official dataset of the DSL shared task (Zampieri et al., 2015). It contains journalistic texts that represent several sets of closely related languages and language varieties: Bulgarian, Macedonian, Serbian, Croatian, Bosnian, Czech, Slovak, Argentinian Spanish, Peninsular Spanish, Brazilian Portuguese, European Portuguese, Malay, Indonesian and a group containing texts written in a set of other languages. (ii) HispaBlogs (HB) is a collection of blogs from different Spanish-speaking countries (Franco-Salvador et al., 2015c): Argentina, Chile, Mexico, Peru, and Spain. Each blog contains approximately 10 posts.

We compare the SK, DR, and KBSim models with several approaches. The baseline is a BOW approach with 10,000 most frequent words represented as binary features, which is implemented as an SVM model with a linear kernel. Previous work results include the best performing systems on the respective shared tasks: Jarvis et al. (2013) for TOEFL, who use lexical and POS n-grams with SVM, and Malmasi and Dras (2015a) for DSLCC, who use an ensemble of models based on word and character n-grams. The ICLE results are from Tetreault et al. (2012), who use an ensemble of models based on features such as character, word and POS n-grams, as well as spelling errors and function word counts. For HispaBlogs, we report the results of the low dimensional representation model based on text statistics proposed by Rangel et al. (2016).

Development Experiments

During development, we experimented with several variants of the SK and DR approaches, as well as with four different combination methods.[33] We note that our KBSim model does not tune parameters. Therefore, we employ it directly with the test partitions.

[33] On the datasets that lack of separate development partitions, we tuned the parameters with 10-fold cross-validation over the training set.

	NLI		LVI	
Dataset	**TOEFL**	**ICLE**	**DSLCC**	**HB**
$\|L\|$	11	7	14	5
Training	9,900	770	252,000	2,250
Development	1,100	-	28,000	-
Test	1,100	-	14,000	1,000
avg. length	243	689	37	3,168

Table 5.17. Datasets used in this study, with the number of languages, the number of instances, and the average instance length in words.

The results of the development experiments demonstrated several trends. The p-grams presence kernel slightly outperformed the intersection kernel in most cases, with the p-spectrum kernel a distant third. The DR model based on averaging all the distributed representations of words was marginally superior to SenVec in all cases. For classification, logistic regressions worked better than SVM.

We note that the DSLCC datset includes two types of language pairs. The first type are closely related languages, such as Czech vs. Slovak, which have distinct written standards. The second type are different national variants of so-called *pluricentric languages*, such as Argentinian vs. European Spanish, which largely follow the same standard. The accuracy of our combined model is in the range of 99-100% accuracy on the first type, as opposed to 88-92% on the second type. We conclude that only the latter should be considered within the LVI task, while the former belongs to the well-studied language identification task (Gold, 1967). Nevertheless, we follow previous work in reporting the average accuracy on the entire DSLCC dataset.

Results and Discussion

Based on the development results, we selected three models for the final testing: the averaging of distributed representations of words with logistic regression, the presence bits string kernel, and our KBSim model with the k-NN process employed in Section 4.6.2 for categorization.

	NLI		LVI	
Dataset	**TOEFL**	**ICLE**	**DSLCC**	**HB**
BOW	0.601	0.782	0.903	0.527
String kernels	**0.828**	**0.892**	0.944	**0.749**
Distributed repr.	0.661	0.594	0.921	**0.722**
KBSim	0.619	0.851	0.915	0.694
Previous work	**0.836**	**0.901**	**0.955**	0.711

Table 5.18. Classification accuracy (in %) on the evaluated datasets.

The final results are shown in Table 5.18.[34] These values confirm that string kernels have the potential to achieve state-of-the-art results on the NLI task. Ionescu et al. (2014) report the accuracy in the range of 77.5–85.3% on TOEFL, and 82.3–91.3% on ICLE. However, we found that string kernels excel on the LVI task as well. They substantially outperform the best published results on HispaBlogs, and are only about 1% below the best system in the DSL shared task.

String kernels are effective on the LVI task for the similar reasons as in the NLI task. For example, the Spanish word *coger* "to take" is used frequently in European news, but not in Latin America, where it has acquired a taboo meaning. The occurrence of this word, which fits within an eight character n-gram, is a strong clue for the SK classifier.

Distributed representations of words work well on the LVI task. On HispaBlogs, our result is better than the result obtained by Rangel et al. (2016). However, their performance is poor on the NLI task, especially on ICLE, which is likely due to the small size of the dataset. Conversely, the considerable DR improvement over the baseline on HispaBlogs may be attributed to the average instance length (see Table 5.17). Another reason may be the frequency of named entities, which we estimate at 5–7% of the tokens in the English essays on general topics that make up TOEFL and ICLE vs. 11-12% in the LVI datasets.

[34] In this study, statistically significant results according to a one-tailed χ^2 test with Yates' correction ($p < 0.05$) are highlighted in bold.

The KBSim model and its combination of knowledge graphs and vector representations does not excel in these tasks and obtains average values. The NLI results are distant from those obtained with string kernels. However, it is notably superior to the distributed representations in the ICLE dataset. This highlights one of the KBSim advantages. The reduced training data of ICLE leads to obtain not very representative distributed representations of words. In contrast, KBSim, that extracts its knowledge from BabelNet, does not need any training data. The results for LVI follow a similar trend. We obtain results between those of the baseline and the distributed representations and string kernels. We note that due to the short texts of DSLCC (see Table 5.17), KBSim uses there its vector component almost at 100%.

We may conclude that these two classification tasks are mainly based on detecting lexical differences. In general, abstract representations of knowledge such as the KBSim and DR ones are less effective than lexical-based ones as SK at leveraging individual "give-away" word tokens like *coger*. In addition, they fail at detecting the lexical and grammatical regularities which are discriminant in these tasks. On the other hand, the distributed representations have the potential to take into account the frequencies of words. For example, a high frequency of the English function word *he* in TOEFL essays is more indicative of Turkish than of Arabic native speakers. In contrast, knowledge graphs excel in tasks were not much training data is available for the other models. In Section 4.6.2 we used KBSim for a cross-language classification task. However, in this section we outperform the baseline in a monolingual task. These facts highlight its potential for that type of settings. Finally, since we obtained results at par with the state of the art with string kernels on both tasks, without any task-specific adaptation, we confirm our hypothesis regarding the inherent similarity of the two tasks.

5.4.3 Discussion about Knowledge Graphs in Other NLP Tasks

After our study of knowledge graphs for single- and cross-domain classification, we focused large part of this thesis in cross-language tasks. However, during our research we employed knowledge graphs, to a lesser extent, also in other applications. In this part of the thesis we applied knowledge graphs to the NLP tasks of community question answering, native language identification, and language variety identification.

Our results in community question answering proved that our CL-KGA model can be combined together with several similarity measures employing training data in order to learn the interpolation weights of the measures. The study of the information gain ratio of these measures showed that knowledge graphs offer a high amount of information, even compared to the popular distributed representations of words. Experimental results in the SemEval 2016 community question answering shared task provided with the best results among all the participants of its subtask B — a question-question similarity task. However, the results in the other two subtasks were average. Therefore, we concluded that the proposed approach is more suitable for similarity tasks rather than, for instance, question answering. We also showed the good performance of knowledge graphs for monolingual similarity analysis.

With respect to the other two studies, we worked under the hypothesis that the native language and the language variety identification tasks are related because both are based on determining the original language of the author. We applied to both tasks our KBSim model and compared it to state of the art approaches: distributed representations of words and string kernels, respectively. That was, to the best of our knowledge, the first study to apply, without any task-specific adaptation, these representations to both tasks. We evaluated the models in several reference datasets. Our results proved that KBSim is very useful when not much training data is available. Our analysis also highlighted that the tasks are mainly based on detecting lexical differences. That produced a very high performance with string kernels on both tasks and average results for our knowledge graph-based model and the distributed representation-based one. The state of the art results of string kernels on both tasks, without any task-specific adaptation, confirmed our hypothesis regarding the relationship of the two tasks.

We conclude this part of the thesis highlighting the knowledge graphs potential for tasks or domains were not training data is available. Finally, we note that these representations worked with competitive results — especially in community question answering — in three monolingual tasks. This highlights its potential as general single- and cross-language representation of text and its meaning.

5.5 Conclusions

In this chapter we tried to connect the contents of the chapters of this thesis which have been published as research articles. We first discussed in detail the results of those chapters with respect to the objectives of this thesis, including also new results in order to complete the picture for each addressed task. Next, we showed our experiments with knowledge graphs in the NLP tasks of community questions answering, native language identification, and language variety identification.

Our main conclusion with respect to the cross-domain polarity classification task is that WSD-based features are useful and contribute to obtain additional improvements when combined with traditional ones. In contrast, the features of the vocabulary expansion only contributed at single-domain level and their cross-domain performance is questionable.

The potential of knowledge graphs for cross-language IR and NLP tasks have been proved too. The results of the CL-KGA and KBSim models compared to several reference models provided with state-of-the-art performance in the cross-language tasks of plagiarism detection, document retrieval and categorization, and also showed interesting insights about the models.

Finally, the use of knowledge graphs in other NLP tasks showed the potential of knowledge graphs for tasks or domains were training data is not available. These representations worked with competitive results in the tasks of community questions answering, native language identification, and language variety identification.

6 Conclusions

The work presented in this thesis has focused on the study of the use of knowledge graphs as a cross-domain and cross-language representation of text and its meaning. A knowledge graph is a graph that expands and relates the original concepts belonging to a set of words. The use of a wide-coverage multilingual semantic network to generate knowledge graphs provides them with a language coverage of hundreds of languages and millions human-general and -specific concepts.

In Chapter 2 we employed knowledge graph-based features — along with other traditional ones and meta-learning — for the NLP task of single- and cross-domain polarity classification. The next part of the thesis focused on cross-language IR tasks. In Chapter 3 we proposed CL-KGA, a fully knowledge graph-based model of similarity analysis for cross-language plagiarism detection. Next, in Chapter 4 we improved that approach to create the KBSim model, which covers knowledge graph shortcomings such as out-of-vocabulary words and verbal tenses. We applied it to cross-language document retrieval and categorization. Finally, in Chapter 5 we analysed the results obtained in the aforementioned chapters and completed the cross-language part of this thesis by extending the evaluation of Chapter 3 and 4 in order to investigate further knowledge graphs at task level, i.e., we evaluated and compared the proposed models in the three cross-language tasks. We finished that chapter studying the use of knowledge graphs for the other NLP tasks. We applied them to community questions answering, native language identification, and language variety identification.

We have studied in depth the cross-domain and cross-language potential of knowledge graphs in the different evaluations and discussions of this work. That study allows us to answer the research questions made in the introduction of this thesis:

Questions about the cross-domain scenario

- *What is the contribution of the knowledge graph-based features for cross-domain NLP tasks?*

  The results of Section 2.4 and 5.1 showed that WSD-based features are useful at cross-domain level and contribute to obtain additional improvements when combined with traditional ones. In contrast, the features of the vocabulary expansion only benefited at single-domain level and their cross-domain potential is questionable. However, we note that vocabulary expansion is still needed during the WSD step (see Section 2.3.1). Therefore, despite its potential as feature has not been proved, it is indirectly relevant for the cross-domain scenario.

Questions about the cross-language scenario

- *What is the contribution of the knowledge graph characteristics in cross-language similarity?*

  Similarly to the answer of the previous research question, the study of these characteristics in the task of cross-language plagiarism detection (see Section 3.5) showed that WSD is the essential component of the representation, being only necessary the use of vocabulary expansion during the WSD processing. The contribution of the language independence is directly related to the performance offered when we used it in our representation. This leads us to the next question.

- *Could knowledge graphs be employed to successfully solve cross-language similarity tasks?*

  In Section 3.5, 4.6, 5.2, and 5.3 we compared the performance of our knowledge graph-based models, CL-KGA and KBSim, for the tasks of cross-language plagiarism detection, document retrieval, and categorisation. Those results showed the strong potential and versatility of knowledge graphs as a cross-language representation of text and its meaning. In addition, they proved that knowledge graph-based models, specially KBSim, offer state-of-the-art performance in these cross-language similarity tasks.

Questions about the use of knowledge graphs in other NLP tasks

- *What is the performance of knowledge graphs in other NLP tasks?*

  The performance of our combination of similarity models for community questions answering obtained the best results among all the participants of the subtask B of the SemEval 2016 Task 3 on Community Question Answering. That combination included the CL-KGA model, which was one of the most determinant similarity measures — in terms of classification — according to our study of their information gain ratio. However, that performance was obtained in a question-question similarity subtask, which is closely related to the CL-KGA similarity nature. The performance of our system was average for the other two subtasks. Similarly, we also obtained average results when we employed KBSim for the native language and the language variety identification tasks. Our results proved that KBSim is very useful for tasks were not much training data is available. Our analysis also highlighted that these two tasks are mainly based on detecting lexical differences, which is contrary to knowledge graphs, that provide with abstract representations of meaning.

  In summary, knowledge graphs offered different performances depending on the tasks. However, we note that our representation obtained always competitive — or at least better than the average — results in three monolingual NLP tasks.

To sum up, we believe that the answers to these questions show the potential of knowledge graphs as a cross-domain and cross-language representation of text and its meaning for NLP and IR tasks. This is also supported by the scientific contributions of this thesis.

6.1 Scientific Contributions

The different contributions of this thesis have been materialised in several publications. As a result, 5 journal, 7 conference, and 2 workshop papers have been generated. Below we sum up the different scientific contributions highlighting their quality using the reference scoring system, i.e. the ERA CORE Conference Ranking and the journal impact factor, respectively.

6.1.1 CROSS-LANGUAGE PLAGIARISM DETECTION

We approached the cross-language plagiarism detection task and the BabelNet multilingual semantic network in:

- **Franco-Salvador, M.**, Gupta, P., and Rosso, P. (2012). Cross-language plagiarism detection using BabelNet's statistical dictionary. *Computación y Sistemas, Revista Iberoamericana de Computación*, 16(4):383–390.

A first version of our cross-language knowledge graph analysis model was published in:

- **Franco-Salvador, M.**, Gupta, P., and Rosso, P. (2013). Cross-language plagiarism detection using a multilingual semantic network. In *Proceedings of the 35th European Conference on Information Retrieval (ECIR'13)*, LNCS(7814), pages 710–713. Springer-Verlag. **(CORE B)**

- **Franco-Salvador, M.**, Gupta, P., and Rosso, P. (2013). Graph-based similarity analysis: a new approach to cross-language plagiarism detection. *Journal of the Spanish Society of Natural Language Processing (Sociedad Española de Procesamiento del Languaje Natural)*, num. 50.

We evaluated the knowledge graph analysis model with cross-language cases of paraphrasing in:

- **Franco-Salvador, M.**, Gupta, P., and Rosso, P. (2014). Knowledge graphs as context models: Improving the detection of cross-language plagiarism with paraphrasing. In *Ferro, N., editor, Bridging Between Information Retrieval and Databases*, volume 8173 of Lecture Notes in Computer Science, pages 227–236. Springer Berlin Heidelberg.

We improved the knowledge graph analysis model and studied its characteristics deeper in:

- **Franco-Salvador, M.**, Rosso, P., and Montes-y-Gómez, M. (2016). A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management*, 52(4):550–570. **(Impact Factor: 1.26)**

Finally, we evaluated the knowledge-based document similarity model for this task and compared it with cross-language distributed representation-based models in:

- **Franco-Salvador, M.**, Gupta, P., Rosso, P., and Banchs, R. E. (2016). Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Knowledge-Based Systems*, 111:87-99. **(Impact Factor: 2.92)**

We collaborated with the committee of the international competition of plagiarism at PAN by evaluating the user submissions of cross-language datasets in:

- **Franco-Salvador, M.**, Bensalem, I., Flores, E., Gupta, P., and Rosso, P. (2015). PAN 2015 Shared Task on Plagiarism Detection: Evaluation of Corpora for Text Alignment. In *Working Notes Papers of the CLEF 2015 Evaluation Labs*, volume 1391 of CEUR Workshop Proceedings. CLEF and CEUR-WS.org.

6.1.2 Cross-language Document Retrieval and Categorization

We extended the cross-language knowledge graph analysis model to cover knowledge graph shortcomings such as out of vocabulary words and verbal tenses with our knowledge-based document similarity model in:

- **Franco-Salvador, M.**, Rosso, P., and Navigli, R. (2014). A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 414–423. Association for Computational Linguistics. **(CORE A)**

6.1.3 Single- and Cross-domain Polarity Classification

We proposed our knowledge-enhanced meta-learning model and studied the performance of knowledge graphs for single- and cross-domain polarity classification in:

- **Franco-Salvador, M.**, Cruz, F. L., Troyano, J. A., and Rosso, P. (2015). Cross-domain polarity classification using a knowledge-enhanced meta-classifier. *Knowledge-Based Systems*, 86:46–56. **(Impact Factor: 2.92)**

We compared that approach with string kernels in:

- Giménez-Pérez, R. M., **Franco-Salvador, M.**, and Rosso, P. (2017). Single and Cross-domain Polarity Classification using String Kernels. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. Association for Computational Linguistics. **(CORE A)**

6.1.4 Language Variety Identification

We first studied how to employ embeddings for this task and compared it with the author profiling state of the art in:

- **Franco-Salvador, M.**, Rangel, F., Rosso, P., Taulé, M., and Martí, M. A. (2015). Language variety identification using distributed representations of words and documents. In *Proceeding of the 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction (CLEF 2015)*, volume LNCS(9283). Springer-Verlag.

Next we employed that model for the international shared task on discriminating between similar languages in:

- **Franco-Salvador, M.**, Rosso, P., and Rangel, F. (2015). Distributed representations of words and documents for discriminating similar languages. In *Proceeding of the Joint Workshop on Language Technology*

for Closely Related Languages, Varieties and Dialects (LT4VarDial), RANLP.

Finally we compared that model with other state-of-the-art ones in language variety identification in:

- Rangel, F., **Franco-Salvador, M.**, and Rosso, P. (2016). A low dimensionality representation for language variety identification. *In Proceedings of the 17th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2016).* Springer-Verlag.

6.1.5 NATIVE LANGUAGE IDENTIFICATION

We proved the relationship between the native language and the language variety identification tasks in:

- **Franco-Salvador, M.**, Kondrak, G., and Rosso, P. (2017). Bridging the native language and the language variety identification tasks. In *Proceedings of the 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES'17)*. **(CORE B)**

6.1.6 COMMUNITY QUESTION ANSWERING

We published the results of a knowledge graph and word embedding-based ensemble model in a SemEval 2016 community question answering shared task paper:

- **Franco-Salvador, M.**, Kar, S., Solorio, T., and Rosso, P. (2016). UH-PRHLT at SemEval-2016 Task 3: Combining lexical and semantic-based features for community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval '16)*, San Diego, California. Association for Computational Linguistics.

6.2 Future Work

Finally, we identify the future research directions we intend to explore to extend the research carried out in the framework of our Ph.D. One of these directions consists in investigating further the performance of knowledge graphs in the cross-domain and cross-language scenarios. For instance, we are interested in the task of cross-lingual text similarity described in Agirre et al. (2016).

In Chapter 3 we presented the distributed concept weighting for semantic relations between concepts. That method is based on the use of distributed representations of concept definitions. However, these definitions are monolingual and our distributed concept representations worked at multilingual level only because the concepts have a multilingual identifier. In order to obtain a completely multilingual representation of concepts, as future work we are interested in exploring other alternatives non-dependent of these definitions and, in consequence, of their language. That research would affect and hopefully improve our knowledge graph weighting scheme.

Another direction that we want to study is related to the knowledge graph construction. In this thesis we followed Navigli and Lapata (2010) to create knowledge graphs. That method is based on searching and merging paths between concepts using a knowledge base. However, more recent WSD techniques derived graph representations from a knowledge base using semantic signatures (Moro et al., 2014). Those signatures represent concepts by a bag of close concepts. Their use removes the path searching-based method, allows for a fast graph creation, and apparently provides with more precise knowledge representations for WSD. Future works will include the development of more fine-grained semantic representations of documents that combine entity-centric document understanding with explicit semantic relations from a knowledge base. We are also interested in new representations which capture events and event chains in text, and link their arguments and relations to a reference knowledge base. Additionally, knowledge base-centric text understanding also calls for new kinds of wide-coverage knowledge bases that include temporal information (e.g., at what point in time facts stated in the knowledge base are true?), and, consequently, novel time-aware semantification methods.

Finally, in order to adapt our graphs to the most common machine learning models, we intend to study vector representations derived from our knowl-

edge graphs. The use of graph kernels (Gärtner et al., 2003) is a popular technique to perform such type of transformations. These kernels are functions that compute the similarity between two graphs employing a similarity measure such as the one presented in Section 3.4. The graph-based vector is obtained by representing a graph as a function of its similarities with a collection of graphs. Note the large amount of literature using semantic tree kernels, i.e., acyclic connected graph kernels. Recent works offered good results with those kernels in open information extraction (Xu et al., 2013) and have proved their potential for similarity tasks in structured lexical similarity (Croce et al., 2011).

List of Figures

List of Tables

Bibliography

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. pages 497–511.

Agirre, E., de Lacalle, O. L., and Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Agirre, E., Martínez, D., de Lacalle, O. L., and Soroa, A. (2006). Two graph-based algorithms for state-of-the-art wsd. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, pages 585–593. Association for Computational Linguistics.

Aletras, N. and Stevenson, M. (2015). A hybrid distributional and knowledge-based model of lexical semantics. In *Proceedings of 4th Joint Conference on Lexical and Computational Semantics (*SEM'15)*, pages 20–29.

Amini, M.-R., Usunier, N., and Goutte, C. (2009). Learning from multiple partially observed views - an application to multilingual text categorization. In *Proceedings of the Annual Neural Information Processing (NIPS'09) Conference - Advances in Neural Information Processing Systems 22*, pages 28–36.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL'98) and 17th International Conference on Computational Linguistics (COLING'98)*, pages 86–90. Association for Computational Linguistics.

Balabanović, M. and Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72.

Banerjee, S. and Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 2003 International Joint Conference on Artificial Intelligence (IJCAI'03)*, volume 3, pages 805–810.

Baroni, M. and Lenci, A. (2009). One distributional memory, many semantic spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics (GEMS'09)*, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barrón-Cedeño, A. (2012). *On the mono- and cross-language detection of text re-use and plagiarism*. Ph.D. thesis, Universitat Politènica de Valèncía.

Barrón-Cedeño, A., Filice, S., Da San Martino, G., Joty, S., Marquez, L., Nakov, P., and Moschitti, A. (2015). Threadlevel information for comment classification in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL'15) and the 7th International Joint Conference on Natural Language Processing (IJCNLP'15)*, volume 15, pages 687–693.

Barrón-Cedeño, A., Gupta, P., and Rosso, P. (2013). Methods for cross-language plagiarism detection. *Knowledge-Based Systems*, 50:211–217.

Barrón-Cedeño, A., Rosso, P., Pinto, D., and Juan, A. (2008). On cross-lingual plagiarism analysis using a statistical model. In *Proceedings of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN'08)*.

Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al. (2007). Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia - a crystallization point for the web of data. *Web Semantics*, 7(3):154–165.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. (2008). Learning bounds for domain adaptation. *Advances in neural information processing systems*, pages 129–136.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 187–205.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, pages 120–128. Association for Computational Linguistics.

Bollegala, D., Weir, D., and Carroll, J. (2011). Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'11)*, pages 132–141. Association for Computational Linguistics.

Bollegala, D., Weir, D., and Carroll, J. (2013). Cross-domain sentiment classification using a sentiment sensitive thesaurus. *Knowledge and Data Engineering, IEEE Transactions on*, 25(8):1719–1731.

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., and Shah, R. (1993). Signature verification using A

"siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(4):669–688.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, pages 136–145. Association for Computational Linguistics.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49(1-47).

Buscaldi, D., Rosso, P., Gómez-Soriano, J. M., and Sanchis, E. (2010). Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems*, 34(2):113–134.

Caramazza, A. and Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, 10(1):1–34.

Cavnar, W. (1995). Using an n-gram-based document representation with a vector processing retrieval model. *National Institute of Standards and Technology (NIST), Special Publication (SP)*, pages 269–278.

Celikyilmaz, A., Hakkani-Tur, D., and Tur, G. (2010). Lda based similarity modeling for question answering. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search (NAACL-HLT'10)*, pages 1–9. Association for Computational Linguistics.

Ceska, Z., Toman, M., and Jezek, K. (2008). Multilingual plagiarism detection. In *Artificial Intelligence: Methodology, Systems, and Applications*, pages 83–92. Springer.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Cimiano, P., Schultz, A., Sizov, S., Sorg, P., and Staab, S. (2009). Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'09)*, volume 9, pages 1513–1518.

Clough, P. et al. (2003). Old and new challenges in automatic plagiarism detection. In *National Plagiarism Advisory Service, 2003; http://ir. shef. ac. uk/cloughie/index. html*. Citeseer.

Clough, P. and Stevenson, M. (2011). Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45(1):5–24.

Corezola Pereira, R., Moreira, V., and Galante, R. (2010). A new approach for cross-language plagiarism analysis. In Agosti, M., Ferro, N., Peters, C., de Rijke, M., and Smeaton, A., editors, *Multilingual and Multimodal Information Access Evaluation*, volume 6360 of *Lecture Notes in Computer Science*, pages 15–26. Springer Berlin Heidelberg.

Croce, D., Moschitti, A., and Basili, R. (2011). Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 1034–1046.

Cruz, F. L., Troyano, J. A., Pontes, B., and Ortega, F. J. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Dai, W., Xue, G.-R., Yang, Q., and Yu, Y. (2007). Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*, pages 210–219. ACM.

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., and Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature*, 80(6574):499–505.

Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 256–263.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Di Marco, A. and Navigli, R. (2013). Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. *Computational Linguistics*, 39(3):709–754.

Diamantaras, K. I. and Kung, S. Y. (1996). *Principal component neural networks*. Wiley New York.

Dredze, M., Crammer, K., and Pereira, F. (2008). Confidence-weighted linear classification. In *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*, pages 264–271. ACM.

Dumais, S. T., Furnas, G., Landauer, T., Deerwester, S., Deerwester, S., et al. (1995). Latent semantic indexing. In *Proceedings of the Text Retrieval Conference (TREC'95)*.

Dumais, S. T., Letsche, T. A., Littman, M. L., and Landauer, T. K. (1997a). Automatic cross-language retrieval using latent semantic indexing. In *Proceedings AAAI-97 Spring Symposium Series: Cross-language Text and Speech Retrieval*, pages 18–24. Hull & D. Oard (Eds.).

Dumais, S. T., Letsche, T. A., Littman, M. L., and Landauer, T. K. (1997b). Automatic cross-language retrieval using latent semantic indexing. In *Proceedings of the AAAI Spring Symposium on Cross-language Text and Speech Retrieval*, volume 15, pages 15–21.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Ehrlich, K. and Rapaport, W. J. (1997). *A computational theory of vocabulary expansion*. Department of Computer Science, State University of New York at Buffalo.

Ehrlich, K. A. (1995). *Automatic Vocabulary Expansion Through Narrative Context*. PhD thesis. UMI Order No. GAX95-25550.

Enríquez, F., Cruz, F. L., Ortega, F. J., G Vallejo, C., and Troyano, J. A. (2013). A comparative study of classifier combination applied to nlp tasks. *Information Fusion*, 14(3):255–267.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 417–422.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings*

of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'15).

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Bradford Books.

Filice, S., Croce, D., Moschitti, A., and Basili, R. (2016). Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*, San Diego, California. Association for Computational Linguistics.

Forman, G. (2008). Bns feature scaling: an improved representation over tf-idf for svm text classification. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*, pages 263–270. ACM.

Franco-Salvador, M., Bensalem, I., Flores, E., Gupta, P., and Rosso, P. (2015a). PAN 2015 Shared Task on Plagiarism Detection: Evaluation of Corpora for Text Alignment. In *Working Notes Papers of the CLEF 2015 Evaluation Labs (CLEF'15)*, volume 1391 of *CEUR Workshop Proceedings*, page n/a. CLEF and CEUR-WS.org.

Franco-Salvador, M., Cruz, F. L., Troyano, J. A., and Rosso, P. (2015b). Cross-domain polarity classification using a knowledge-enhanced meta-classifier. *Knowledge-Based Systems*, 86:46 – 56.

Franco-Salvador, M., Gupta, P., and Rosso, P. (2012). Cross-language plagiarism detection using BabelNet's statistical dictionary. *Computación y Sistemas, Revista Iberoamericana de Computación*, 16(4):383–390.

Franco-Salvador, M., Gupta, P., and Rosso, P. (2013a). Cross-language plagiarism detection using a multilingual semantic network. In *Proceedings of the 35th European Conference on Information Retrieval (ECIR'13)*, LNCS(7814), pages 710–713. Springer-Verlag.

Franco-Salvador, M., Gupta, P., and Rosso, P. (2013b). Graph-based similarity analysis: a new approach to cross-language plagiarism detection. *Journal of the Spanish Society of Natural Language Processing (Sociedad Española de Procesamiento del Languaje Natural)*, num. 50.

Franco-Salvador, M., Gupta, P., and Rosso, P. (2014a). Knowledge graphs as context models: Improving the detection of cross-language plagiarism with paraphrasing. In Ferro, N., editor, *Bridging Between Information Retrieval and Databases*, volume 8173 of *Lecture Notes in Computer Science*, pages 227–236. Springer Berlin Heidelberg.

Franco-Salvador, M., Gupta, P., Rosso, P., and Banchs, R. E. (2016a). Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Knowledge-Based Systems*, 111:87–99.

Franco-Salvador, M., Kar, S., Solorio, T., and Rosso, P. (2016b). UH-PRHLT at SemEval-2016 Task 3: Combining lexical and semantic-based features for community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*, San Diego, California. Association for Computational Linguistics.

Franco-Salvador, M., Kondrak, G., and Rosso, P. (2017). Bridging the native language and the language variety identification tasks. In *Proceedings of the 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES'17)*.

Franco-Salvador, M., Rangel, F., Rosso, P., Taulé, M., and Martí, M. A. (2015c). Language variety identification using distributed representations of words and documents. In *Proceeding of the 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction (CLEF'15)*, volume LNCS(9283). Springer-Verlag.

Franco-Salvador, M., Rosso, P., and Montes y Gómez, M. (2016c). A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management*, 52(4):550–570.

Franco-Salvador, M., Rosso, P., and Navigli, R. (2014b). A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)*, pages 414–423. Association for Computational Linguistics.

Franco-Salvador, M., Rosso, P., and Rangel, F. (2015d). Distributed representations of words and documents for discriminating similar languages.

In *Proceeding of the RANLP'15 Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, page n/a.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 2007 International Joint Conference on Artificial Intelligence (IJCAI'07)*, volume 7, pages 1606–1611.

Gärtner, T., Flach, P., and Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines*, pages 129–143. Springer.

Giménez-Pérez, R. M., Franco-Salvador, M., and Rosso, P. (2017). Single and cross-domain polarity classification using string kernels. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*. Association for Computational Linguistics.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, 10(5):447–474.

Gupta, P., Bali, K., Banchs, R. E., Choudhury, M., and Rosso, P. (2014). Query expansion for mixed-script information retrieval. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'14)*, pages 677–686.

Gupta, P., Banchs, R. E., and Rosso, P. (2015). Continuous space models for clir. *Technical Report, Universitat Politècnica de València*.

Gupta, P., Barrón-Cedeño, A., and Rosso, P. (2012). Cross-language high similarity search using a conceptual thesaurus. In *Proceedings 3rd International Conference of CLEF Initiative on Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics (CLEF'12)*, LNCS(7488), pages 67–75. Springer-Verlag.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, 13(1):307–361.

Hall, M. A. and Smith, L. A. (1998). Practical feature subset selection for machine learning. In *Proceedings of Australian Computer Science Conference (ACSC'98)*, pages 181–191.

Hamouda, A. and Rohaim, M. (2011). Reviews classification using sentiwordnet lexicon. *The Online Journal on Computer Science and Information Technology*.

Hassan, S. and Mihalcea, R. (2011). Semantic relatedness using salient semantic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'11)*.

Hastie, T. and Tibshirani, R. (2003). *The elements of statistical learning*. Springer, corrected edition, July.

Haveliwala, T., Kamvar, S., and Jeh, G. (2003). An analytical comparison of approaches to personalizing pagerank. Technical Report 2003-35, Stanford InfoLab.

He, X. (2007). Using word dependent transition models in hmm based word alignment for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation (ACL'07)*, pages 80–87. Association for Computational Linguistics.

Heck, L. P., Hakkani-Tür, D., and Tür, G. (2013). Leveraging knowledge graphs for web-scale unsupervised semantic parsing. In *Proceedings 14th Annual Conference of the International Speech Communication Association (Interspeech'13)*, pages 1594–1598.

Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61.

Hou, Y., Tan, C., Wang, X., Zhang, Y., Xu, J., and Chen, Q. (2015). Hitszicrc: Exploiting classification approach for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*, volume 15, pages 196–202. Association for Computational Linguistics.

Hovy, E. H., Navigli, R., and Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Hull, D. (1994). Improving text retrieval for the routing problem using latent semantic indexing. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 282–291. Springer.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL'15)*.

Ionescu, R.-T., Popescu, M., and Cahill, A. (2014). Can characters reveal your native language? A language-independent approach to native language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 1363–1373.

Ito, M., Nakayama, K., Hara, T., and Nishio, S. (2008). Association thesaurus construction methods based on link co-occurrence analysis for wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*, pages 817–826. ACM.

Jackson, D. A., Somers, K. M., and Harvey, H. H. (1989). Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? *American Naturalist*, pages 436–453.

Jarvis, S., Bestgen, Y., and Pepper, S. (2013). Maximizing classification accuracy in native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–118.

Jeon, J., Croft, W. B., and Lee, J. H. (2005). Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM International Conference on Information and knowledge Management (CIKM'05)*, pages 84–90. ACM.

Jia, L., Yu, C., and Meng, W. (2009). The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*, pages 1827–1830. ACM.

Jiang, J. and Zhai, C. (2007). Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting on Association for Computational Linguistics (ACL'07)*, volume 7, pages 264–271.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features*. Springer.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, pages 133–142. ACM.

Jolliffe, I. T. (1986). *Principal component analysis*, volume 487. Springer-Verlag New York.

Koppel, M., Schler, J., and Zigdon, K. (2005). Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics*, pages 209–217. Springer.

Landauer, T. K. and Littman, M. L. (1994). Computerized cross-language document retrieval using latent semantic indexing. US Patent 5,301,109.

Lauly, S., Boulanger, A., and Larochelle, H. (2014a). Learning multilingual word representations using a bag-of-words autoencoder. *CoRR*, abs/1401.1803.

Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V. C., and Saha, A. (2014b). An autoencoder approach to learning bilingual word representations. In *Proceedings of the Annual Neural Information Processing (NIPS'14) Conference - Advances in Neural Information Processing Systems 27*, pages 1853–1861.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*.

Li, S. and Zong, C. (2008). Multi-domain sentiment classification. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (ACL'08)*, pages 257–260. Association for Computational Linguistics.

Lin, D. and Pantel, P. (2001). Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(04):343–360.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002). Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444.

Maier, W. and Gómez-Rodríguez, C. (2014). Language variety identification in spanish tweets. In *Proceedings of the EMNLP'14 Workshop on Language Technology for Closely Related Languages and Language Variants (EMNLP'14)*, pages 25–35, Doha, Qatar. Association for Computational Linguistics.

Malmasi, S. and Dras, M. (2015a). Language identification using classifier ensembles. In *Proceeding of the RANLP Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, Hissar, Bulgaria.

Malmasi, S. and Dras, M. (2015b). Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'15)*.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*, volume 999. MIT Press.

Martín-Valdivia, M.-T., Martínez-Cámara, E., Perea-Ortega, J.-M., and Ureña-López, L. A. (2013). Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10):3934 – 3942.

Matuschek, M. and Gurevych, I. (2013). Dijkstra-WSA: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics (TACL)*, 1:151–164.

Maurer, H. A., Kappe, F., and Zaka, B. (2006). Plagiarism-a survey. *J. UCS*, 12(8):1050–1084.

Mayfield, J. and McNamee, P. (1999). Indexing using both n-grams and words. *National Institute of Standards and Technology (NIST), Special Publication (SP)*, pages 419–424.

McCallum, A. and Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *Proceedings of the AAAI workshop on learning for text categorization (AAAI'98)*, volume 752, pages 41–48. Citeseer.

McDonald, S. and Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society (COGSCI'01)*, pages 611–616.

Mcnamee, P. and Mayfield, J. (2004). Character n-gram tokenization for european language text retrieval. *Information Retrieval*, 7(1-2):73–97.

Mihalcea, R. and Radev, D. (2011). *Graph-based natural language processing and information retrieval*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations (ICLR'13)*, pages 1–12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the Annual Neural Information Processing (NIPS'13) Conference - Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 12th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'13)*, pages 746–751.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, pages 880–889. Association for Computational Linguistics.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Mnih, A. and Teh, Y. W. (2012). A fast and simple algorithm for training neural probabilistic language models. *arXiv preprint arXiv:1206.6426*.

Montes y Gómez, M., Gelbukh, A. F., López-López, A., and Baeza-Yates, R. A. (2001). Flexible comparison of conceptual graphs. In *Proceedings of the 12th International Conference on Database and Expert Systems Applications (DEXA'01)*, pages 102–111.

Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *Proceedings of the international workshop on artificial intelligence and statistics (Aistats'05)*, pages 246–252. Citeseer.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Nakov, P., Màrquez, L., Magdy, W., and Moschitti, A. (2015). Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*, pages 269–281. Association for Computational Linguistics.

Nakov, P., Màrquez, L., Magdy, W., Moschitti, A., Glass, J., and Randeree, B. (2016). SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*, San Diego, California. Association for Computational Linguistics.

Nastase, V. and Strube, M. (2013). Transforming wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85.

Navigli, R. (2009). Word Sense Disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69.

Navigli, R., Jurgens, D., and Vannella, D. (2013). Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the 7^{th} International Workshop on Semantic Evaluation (SemEval'13), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM'13)*, pages 222–231.

Navigli, R. and Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692.

Navigli, R. and Ponzetto, S. P. (2012a). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Navigli, R. and Ponzetto, S. P. (2012b). BabelRelate! a joint multilingual approach to computing semantic relatedness. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Ohana, B. and Tierney, B. (2009). Sentiment classification of reviews using sentiwordnet. In *Proceedings of the 9th Information Technology and Telecommunications Conference (IT&T'09)*, page 13.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., and Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pages 751–760. ACM.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, pages 79–86. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 1532–1543.

Pilehvar, M. T. and Navigli, R. (2014). A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. *Computational Linguistics*, 1(1).

Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., and Rosso, P. (2009). A statistical approach to crosslingual natural language tasks. *Journal of Algorithms*, 64(1):51–60.

Platt, J. C., Toutanova, K., and Yih, W.-t. (2010). Translingual document representations from discriminative projections. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*, pages 251–261. Association for Computational Linguistics.

Ponzetto, S. P. and Navigli, R. (2010). Knowledge-rich Word Sense Disambiguation rivaling supervised system. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 1522–1531.

Popescu, M. and Grozea, C. (2012). Kernel methods and string kernels for authorship analysis. In *Proceeding of the 3th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction (CLEF'12)*.

Popping, R. (2003). Knowledge graphs and network text analysis. *Social Science Information*, 42(1):91–106.

Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., and Rosso, P. (2010a). Overview of the 2nd international competition on plagiarism detection. In *Braschler M., Harman D., and Pianta E.(Eds.), Notebook Papers of CLEF 2010 LABs and Workshops (CLEF'10)*.

Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011a). Cross-language plagiarism detection. *Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis*, 45(1):45–62.

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011b). Overview of the 3rd international competition on plagiarism detection. In *Petras V., Forner P., Clough P. (Eds.), Notebook Papers of CLEF 2011 LABs and Workshops (CLEF'11)*.

Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., and Stein, B. (2014). Overview of the 6th international competition on plagiarism detection. In *Working Notes for CLEF 2014 Conference (CLEF'14)), Sheffield, UK, September 15-18, 2014.*, pages 845–876.

Potthast, M., Hagen, M., Göring, S., Rosso, P., and Stein, B. (2015). Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment. In *Working Notes Papers of the CLEF 2015 Evaluation Labs (CLEF'15)*, volume 1391 of *CEUR Workshop Proceedings*, page n/a. CLEF and CEUR-WS.org.

Potthast, M., Stein, B., and Anderka, M. (2008). A wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval*, pages 522–530. Springer.

Potthast, M., Stein, B., Barrón-Cedeño, A., and Rosso, P. (2010b). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 997–1005. Association for Computational Linguistics.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Quinlan, J. R. (1996). Bagging, boosting, and c4.5. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'96)*, pages 725–730.

Raileanu, L. E. and Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1):77–93.

Rangel, F., Franco-Salvador, M., and Rosso, P. (2016). A low dimensionality representation for language variety identification. In *Proceedings of*

the 17th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'16). Springer-Verlag.

Rangel, F., Rosso, P., Potthast, M., Stein, B., and Daelemans, W. (2015). Overview of the 3rd author profiling task at pan 2015. In *Proceeding of the 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction (CLEF'15)*, volume 1391. CEUR-WS.org.

Ravichandran, D. and Hovy, E. (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, pages 41–47. Association for Computational Linguistics.

Reyes, A. and Rosso, P. (2013). On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, pages 1–20.

Rosso, P., Hurtado, L.-F., Segarra, E., and Sanchis, E. (2012). On the voice-activated question answering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(1):75–85.

Sadat, F., Johnson, H., Agbago, A., Foster, G., Martin, J., and Tikuisis, A. (2005). Portage: A phrase-based machine translation system. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts (ACL'05)*, Ann Arbor, USA.

Sadat, F., Kazemi, F., and Farzindar, A. (2014). Automatic identification of arabic language varieties and dialects in social media. In *Proceeding of the 1st International Workshop on Social Media Retrieval and Analysis (SoMeRa'14)*.

Salakhutdinov, R. and Hinton, G. (2009). Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Salton, G., Fox, E. A., and Wu, H. (1983). Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Shen, H., Bunescu, R., and Mihalcea, R. (2013). Coarse to fine grained sense disambiguation in wikipedia. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM'13), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 22–31. Association for Computational Linguistics.

Snoek, J. (2013). *Bayesian Optimization and Semiparametric Models with Applications to Assistive Technology*. PhD thesis, University of Toronto.

Sowa, J. F. (2006). Semantic networks. *Encyclopedia of Cognitive Science*.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with +20 languages. In *Proceedings 5th International Conference on language resources and evaluation (LREC'06)*.

Tetreault, J., Blanchard, D., and Cahill, A. (2013). A report on the first native language identification shared task. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA'13)*, pages 48–57. Citeseer.

Tetreault, J. R., Blanchard, D., Cahill, A., and Chodorow, M. (2012). Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*, pages 2585–2602.

Thompson, B. (2005). Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*.

Tran, Q. H., Tran, V., Vu, T., Nguyen, M., and Pham, S. B. (2015). Jaist: Combining multiple features for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*, volume 15, pages 215–219. Association for Computational Linguistics.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, pages 417–424. Association for Computational Linguistics.

Tyler, L., Bright, P., Dick, E., Tavares, P., Pilgrim, L., Fletcher, P., Greer, M., and Moss, H. (2003). Do semantic categories activate distinct cortical regions? evidence for a distributed neural semantic system. *Cognitive Neuropsychology*, 20(3-6):541–559.

Van Halteren, H., Zavrel, J., and Daelemans, W. (1998). Improving data driven wordclass tagging by system combination. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL'98) and 17th International Conference on Computational Linguistics (COLING'98)*, pages 491–497. Association for Computational Linguistics.

Voorhees, E. M. (1999). The trec-8 question answering track report. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, volume 99, pages 77–82.

Vossen, P. (2004). EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. *International Journal of Lexicography*, 17(2):161–173.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Xu, Y., Kim, M.-Y., Quinn, K., Goebel, R., and Barbosa, D. (2013). Open information extraction with tree kernels. In *Proceedings of the 12th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'13)*, pages 868–877.

Yarlett, D. and Ramscar, M. J. (2008). *Language learning through similarity-based generalization*. PhD thesis.

Ye, Z., Huang, X., and Lin, H. (2009). A graph-based approach to mining multilingual word associations from wikipedia. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*, pages 690–691. ACM.

Yih, W.-t., Toutanova, K., Platt, J. C., and Meek, C. (2011). Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL'11)*, pages 247–256. Association for Computational Linguistics.

Zampieri, M. and Gebre, B. G. (2012). Automatic identification of language varieties: The case of portuguese. In *Proceedings of the 11th Conference on Natural Language Processing (KONVENS'12)*, pages 233–237. Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI).

Zampieri, M., Tan, L., Ljubešić, N., and Tiedemann, J. (2014). A report on the dsl shared task 2014. In *Proceedings of the COLING First Joint Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., and Nakov, P. (2015). Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, Hissar, Bulgaria.