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

Marc Franco-Salvador[a,*], Paolo Rosso[a], Manuel Montes-y-Gómez[b]

[a] *Pattern Recognition and Human Language Technology (PRHLT) Research Center*
Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain
[b] *Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica,*
Luis Enrique Erro 1, Puebla 72840. Mexico

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

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

[*] Corresponding author.
Email address: mfranco@prhlt.upv.es (Marc Franco-Salvador)

Preprint submitted to Information Processing & Management *December 3, 2015*

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 dictionaries 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., 2013), 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 dur-*

ing 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 2 we provide an overview of the state of the art in CL plagiarism detection and distributed representations of concepts. In Section 3 we describe the knowledge graphs, their weighting schemes, including our new approach, and their main characteristics. In Section 4 we describe the CL-KGA model for CL plagiarism detection. Finally, in Section 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.

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.

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 3.4.4, 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 et al., 2008; Pinto et al., 2009; Barrón-Cedeño, 2012) 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.

[1] `http://www.illc.uva.nl/EuroWordNet/`

[2] `http://eurovoc.europa.eu/`

The Cross-Language Knowledge Graph Analysis (CL-KGA) model (Franco-Salvador et al., 2013) 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, 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., 1997) 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 (Potthast et al., 2015; Franco-Salvador et al., 2015a), 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

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

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

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.

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.

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)

[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/`

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.

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 3.3.2 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. 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, 2012c), WSD (Navigli and Ponzetto, 2012b), 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., 2013), and CL document retrieval and categorisation (Franco-Salvador et al., 2014b). In Figure 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.

[7] Each word has a number of senses. We define "concept" as any of those senses, which may be represented via synsets (see Section 3.1).

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

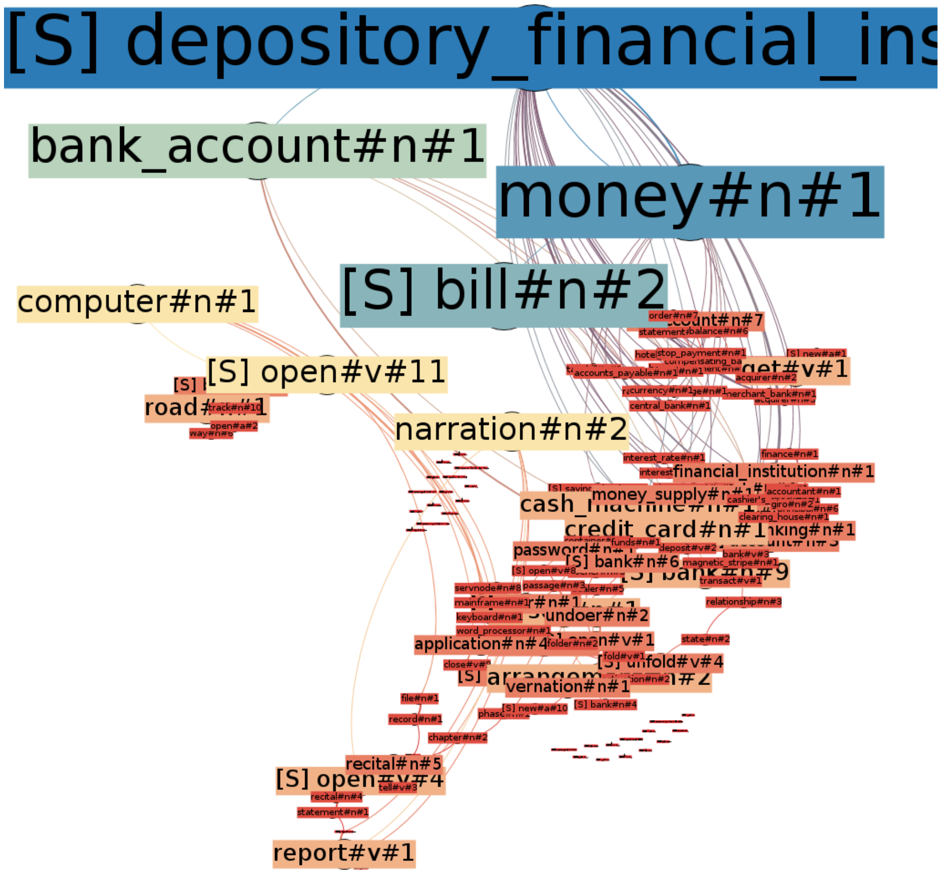

Figure 1: Knowledge graph built from the sentence "I opened a new bank account" (source words: ("open#v, new#a, bank#n, account#n")). Larger boxes represent concepts with higher connectivity.

3.1. BabeNet

BabelNet[9] 2.5 (Navigli and Ponzetto, 2012b) is a multilingual semantic network whose concepts and relations are obtained from the automatic mapping onto WordNet of Wikipedia, OmegaWiki,[10] Wiktionary,[11] Wiki-

[9] http://babelnet.org

[10] http://omegawiki.org

[11] http://wiktionary.org

data,[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. 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 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).

3.2. 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:

[12] http://wikidata.org

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

[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.

$$S_K = \bigcup_{t \in T} \mathrm{Synsets}_L(t), \quad (1)$$

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

(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 $v \to v_1 \to \cdots \to v_n \to v'$, 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 (2012b), 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, 2012b). 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. 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 (2012b) in order to measure this degree of association between synsets. Next, in Section 3.3.2 we present our

[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 5.3).

new method based on distributed representations of concepts for weighting their relations.

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:

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

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:

$$\text{Semantic Relatedness}(w, w') = \frac{2f_{w,w'}}{f_w + f_{w'}} \tag{3}$$

Using this weighting scheme, we depict in Figure 2 a histogram of the distribution of BabelNet's relation weights. We observe that only ∼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 3.3.2).

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

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

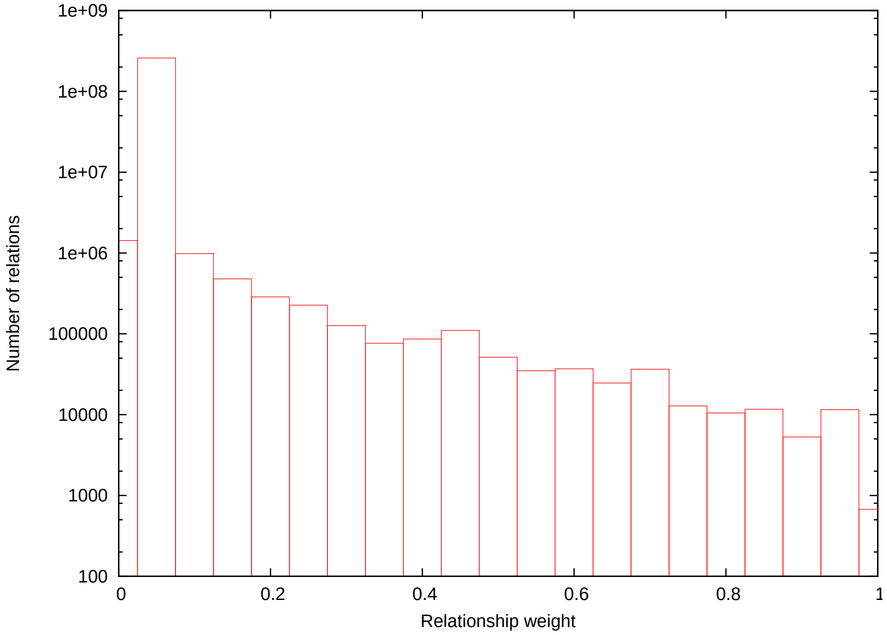

Figure 2: Distribution of relation weights in BabelNet using the Dice's coefficient-based weighting.

3.3.2. Distributed representations of concepts for computing semantic relatedness

The weighting described in Section 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

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

representations in scenarios in which the word was never seen before, but its 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:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log \mathrm{p}(w_{t+j}|w_t) \tag{4}$$

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)

[21] Although the approach described in Section 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.

of the iterated words of a sentence. Next we detail the 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 (Le and Mikolov, 2014; Franco-Salvador et al., 2015c,d).

(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 $\vec{v}_{s'}$:

[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.

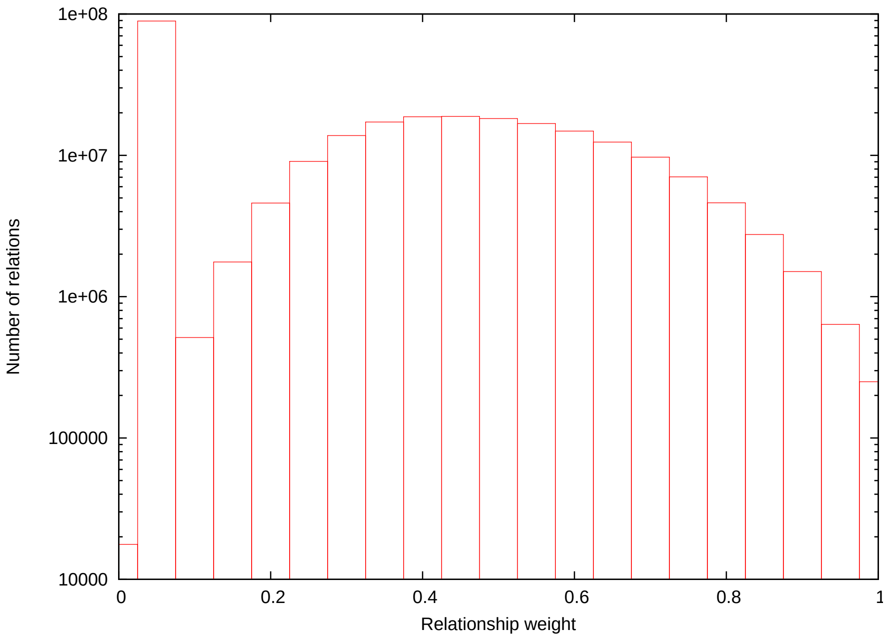

Figure 3: Distribution of relation weights in BabelNet using distributed concept weighting.

$$\text{Semantic Relatedness}(s, s') = \frac{\vec{v}_s \cdot \vec{v}_{s'}}{\|\vec{v}_s\| \|\vec{v}_{s'}\|} \tag{5}$$

In Figure 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 3.3.1. In addition, if we observe Figure 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.

3.4. 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.

3.4.1. Word sense disambiguation

Knowledge graphs have been successfully used in the past to perform WSD (Navigli and Ponzetto, 2012b). As we stated, the graphs created in Section 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 6 to select the disambiguation s_{WSD} among the S_K synsets of each tuple, where 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.

$$s_{\text{WSD}} = \underset{s \in S_K}{\text{argmax}} \quad \text{score}(s) \tag{6}$$

(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.

3.4.2. 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 3.4.1.

3.4.3. Language independence

As we mentioned at the beginning of Section 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 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 4 and 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.

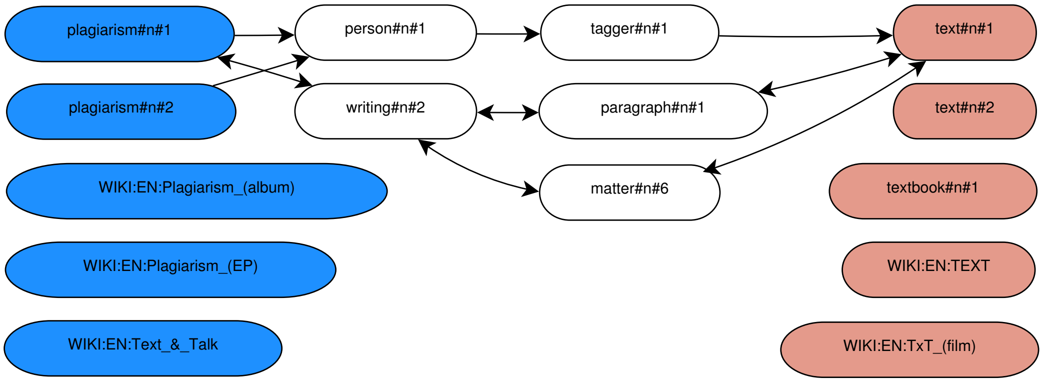

Figure 4: Knowledge graph built from the English sentence “text with plagiarism” (source words: (“text#n”,“plagiarism#n”)). The coloured nodes are the different senses of the original words.

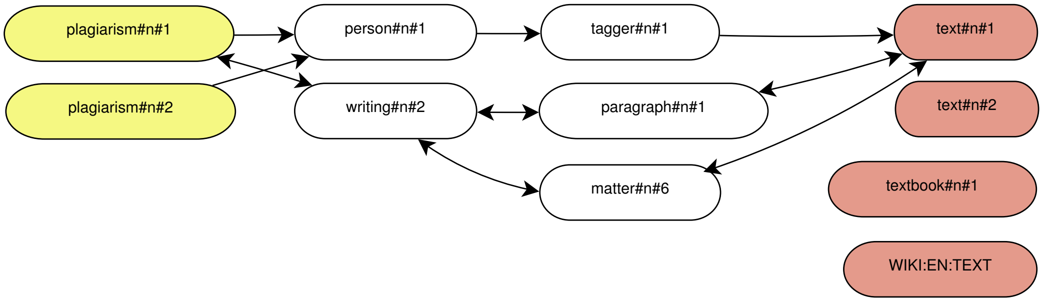

Figure 5: Knowledge graph built from the Spanish sentence “texto con plagio” (source words: (“texto#n”,“plagio#n”)).

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

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

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

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 5 we compare our model with CL-ESA to show the differences in performance at detecting CL plagiarism.

4. Cross-language knowledge graph analysis

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. (2013) 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 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 3.2 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:

$$\mathrm{S}_c(G,G') = \frac{2 \cdot \sum_{c \in \mathrm{V}(G) \cap \mathrm{V}(G')} \mathrm{w}(c)}{\sum_{c \in \mathrm{V}(G)} \mathrm{w}(c) + \sum_{c \in \mathrm{V}(G')} \mathrm{w}(c)}, \tag{7}$$

where $\mathrm{w}(c)$ is the weight of a concept c (see Section 3.2). Likewise, we calculate the similarity between the relations as:

$$\mathrm{S}_r(G,G') = \frac{2 \cdot \sum_{r \in \mathrm{E}(G) \cap \mathrm{E}(G')} \mathrm{w}(r)}{\sum_{r \in \mathrm{E}(G)} \mathrm{w}(r) + \sum_{r \in \mathrm{E}(G')} \mathrm{w}(r)}, \tag{8}$$

where $\mathrm{w}(r)$ is the weight of a semantic relation r (see Section 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:

$$\mathrm{S}_g(G,G') = a \cdot \mathrm{S}_c(G,G') + b \cdot \mathrm{S}_r(G,G'), \tag{9}$$

where a and b, $a+b=1$, are the parameters of the relevance of concepts and relations respectively. In Figure 6 we can see the differences among CL-KGA, CL-C3G, and CL-ASA when detecting CL plagiarism. Thanks to the aforementioned characteristics (see Section 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 (Potthast et al., 2011a; Barrón-Cedeño et al., 2013), 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 (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

[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.

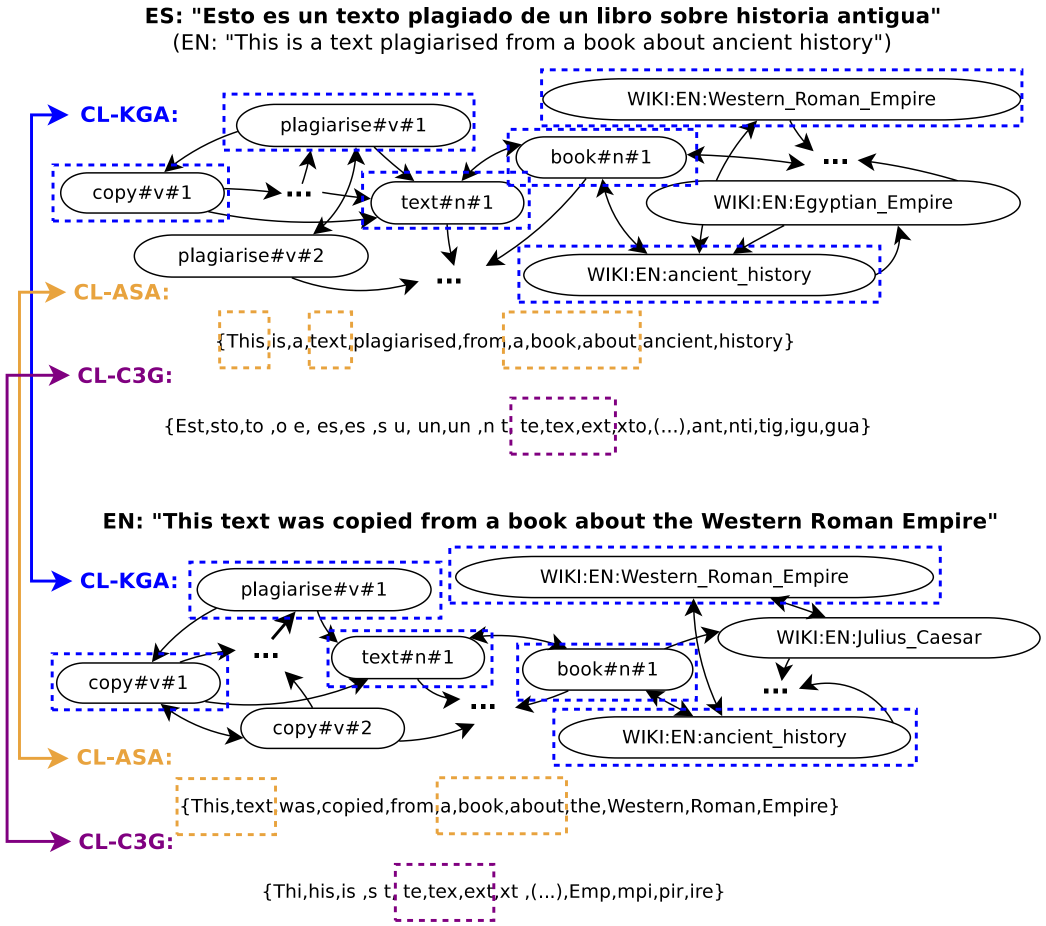
ES: "Esto es un texto plagiado de un libro sobre historia antigua"
(EN: "This is a text plagiarised from a book about ancient history")
CL-KGA:
plagiarise#v#1
WIKI:EN:Western_Roman_Empire
book#n#1
copy#v#1
...
text#n#1
WIKI:EN:Egyptian_Empire
plagiarise#v#2
...
WIKI:EN:ancient_history
CL-ASA:
{This,is,a,text,plagiarised,from,a,book,about,ancient,history}
CL-C3G:
{Est,sto,to ,o e, es,es ,s u, un,un ,n t, te,tex,ext,xto,(...),ant,nti,tig,igu,gua}
EN: "This text was copied from a book about the Western Roman Empire"
CL-KGA:
plagiarise#v#1
WIKI:EN:Western_Roman_Empire
WIKI:EN:Julius_Caesar
...
text#n#1
book#n#1
copy#v#1
...
...
copy#v#2
WIKI:EN:ancient_history
CL-ASA:
{This,text,was,copied,from,a,book,about,the,Western,Roman,Empire}
CL-C3G:
{Thi,his,is ,s t, te,tex,ext,xt ,(...),Emp,mpi,pir,ire}

Figure 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.

described in Algorithm 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

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

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

of the plagiarism case. This algorithm has been used for evaluating all the models compared in the evaluation section.

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

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)

[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/`

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 1: Statistics of PAN-PC-10 and PAN-PC-11 cross-language plagiarism detection partitions.

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 5.1 we can see the statistics of the datasets.

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

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

r. Based on these representations, the precision and the recall at character level of R under S are measured as follows:

$$\text{precision}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\bigcup_{s \in S}(s \sqcap r)|}{|r|}; \tag{10}$$

$$\text{recall}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\bigcup_{r \in R}(s \sqcap r)|}{|s|}, \tag{11}$$

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:

$$\text{granularity}(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|, \tag{12}$$

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):

$$\text{plagdet}(S, R) = \frac{F_1(S, R)}{\log_2(1 + \text{granularity}(S, R))} \tag{13}$$

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., 2013) — CL-KGA (BabelNet 1.0) from here —, and those obtained by the CL-KGA

[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).

	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 3.3.1).
	CL-KGA (DCW)	CL-KGA using the distributed concept weighting (see Section 3.3.2).
	CL-KGA (WSD path filter)	CL-KGA keeping only paths related to WSD concepts (see Section 3.4.1).
	CL-KGA (WSD concepts)	CL-KGA keeping only weighted WSD concepts (see Section 3.4.1).
	CL-KGA (WSD concepts w/o weighting)	CL-KGA keeping only a BOW of WSD concepts (see Section 3.4.1).
	CL-KGA (DCW) (WSD concepts w/o weighting)	Same as previous using the distributed concept weighting.

Table 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).

variations introduced in Section 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 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 2 we can

[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.

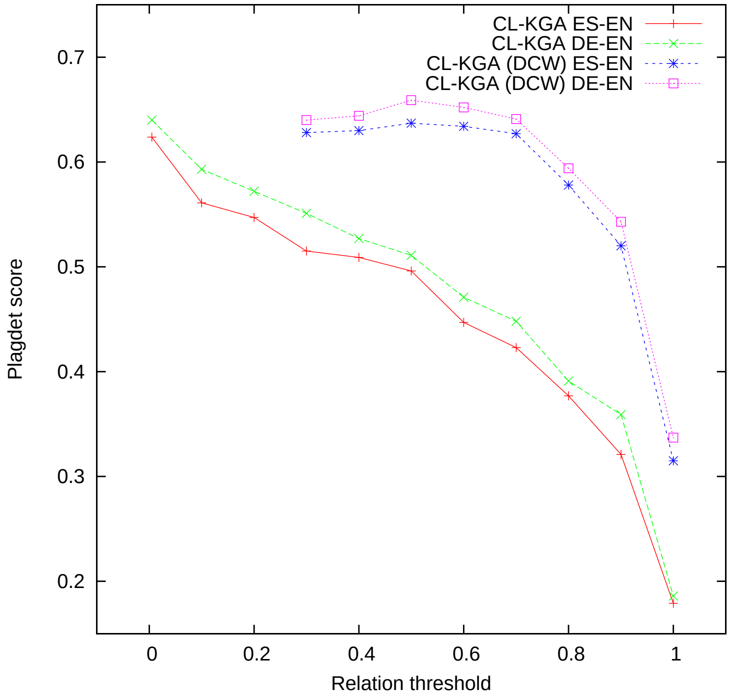

Figure 7: Plagdet score in PAN-PC-10 dataset in function of the threshold between relations.

find a summary of all the models included in the evaluation.

The experiments were divided into three subsections: (i) in Section 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 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 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.

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 3.3.1) and the new method using distributed representations of concepts (cf. Section 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 3.2). Next, we selected the values of relevance for concepts and relations used with CL-KGA (cf. Section 4) for both weightings.

To determine the threshold of the semantic relations, we tested values between 0.001 and 1.[39] In Figure 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 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 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

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

[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%.

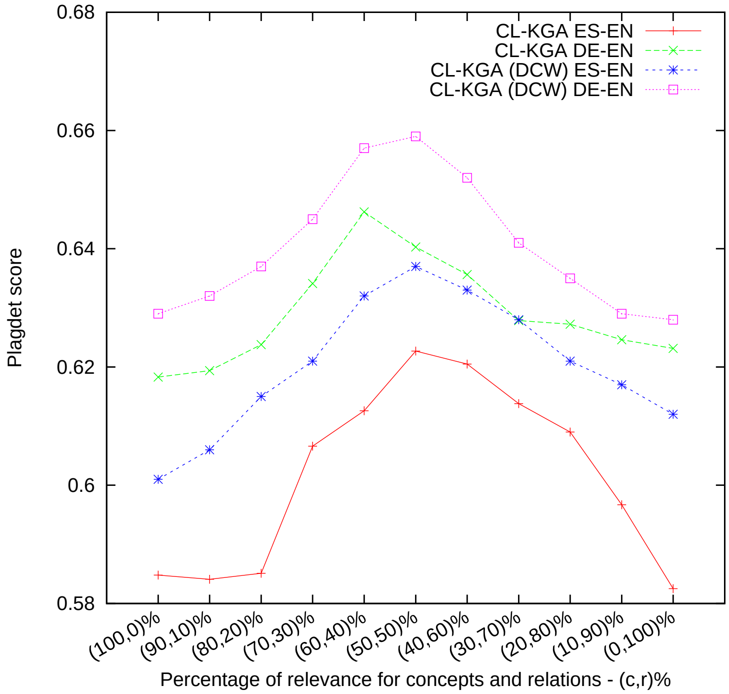

Figure 8: Plagdet score in the PAN-PC-10 dataset as function of percentage of relevance of concepts and relations.

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.

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

[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: Results of PAN-PC-11 Spanish-English partition using the CL-KGA variants.

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.

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 3.3.1, 3.3.2 and 3.4.1) of the CL-KGA model and study the characteristics of our approach.

In Table 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., 2013). Similarly to the results with PAN-PC-10 of Section 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

[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
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 4: Results of PAN-11 German-English partition using the CL-KGA variants.

perform WSD offers an accuracy close to 70% (Navigli and Ponzetto, 2012b). 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 3.2 and 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 3.4.2), 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.

The results for German-English were a similar. In Table 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.

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 2) using the PAN-PC-11 dataset for CL plagiarism detection.

In Table 5 we show the results obtained for Spanish-English. The lowest results were obtained by CL-C3G. This is unsurprising if we consider that

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

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 possible 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.

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

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 3.4.4: 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 6 we can see the results. There are some interesting aspects to highlight. CL-C3G 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.

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

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

5.5.1. Detecting cross-language plagiarism detection with paraphrasing

As we mentioned in Section 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

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

Tables 7 and 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 Algorithm 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.

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 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 ∼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.

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.

References

Aletras, N., 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). pp. 20–29.

Baker, C. F., Fillmore, C. J., Lowe, J. B., 1998. The berkeley framenet project. In: Proceedings of the 17th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, pp. 86–90.

Banerjee, S., Pedersen, T., 2003. Extended gloss overlaps as a measure of semantic relatedness. In: IJCAI. Vol. 3. pp. 805–810.

Barrón-Cedeño, A., 2012. On the mono- and cross-language detection of text re-use and plagiarism. Ph.D. thesis, Universitat Politènica de València.

Barrón-Cedeño, A., Gupta, P., Rosso, P., 2013. Methods for cross-language plagiarism detection. Knowledge-Based Systems 50, 211–217.

Barrón-Cedeño, A., Rosso, P., Pinto, D., Juan, A., 2008. On cross-lingual plagiarism analysis using a statistical model. In: Proc. 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.

Ceska, Z., Toman, M., Jezek, K., 2008. Multilingual plagiarism detection. In: Artificial Intelligence: Methodology, Systems, and Applications. Springer, pp. 83–92.

Clough, P., Stevenson, M., 2011. Developing a corpus of plagiarised short answers. Language Resources and Evaluation 45 (1), 5–24.

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.

Corezola Pereira, R., Moreira, V., Galante, R., 2010. A new approach for cross-language plagiarism analysis. In: Agosti, M., Ferro, N., Peters, C., de Rijke, M., Smeaton, A. (Eds.), Multilingual and Multimodal Information Access Evaluation. Vol. 6360 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 15–26.

Dumais, S. T., Letsche, T. A., Littman, M. L., Landauer, T. K., 1997. Automatic cross-language retrieval using latent semantic indexing. In: AAAI spring symposium on cross-language text and speech retrieval. Vol. 15. pp. 15–21.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., 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).

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

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

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

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

Franco-Salvador, M., Rangel, F., Rosso, P., Taulé, M., 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 2015). Vol. LNCS(9283). Springer-Verlag, p. n/a.

Franco-Salvador, M., Rosso, P., 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. Association for Computational Linguistics, pp. 414–423.

Franco-Salvador, M., Rosso, P., Rangel, F., 2015d. 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. p. n/a.

Gabrilovich, E., Markovitch, S., 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI. Vol. 7. pp. 1606–1611.

Gupta, P., Barrón-Cedeño, A., Rosso, P., 2012. Cross-language high similarity search using a conceptual thesaurus. In: Proc. 3rd Int. Conf. of CLEF Initiative on Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics. LNCS(7488). Springer-Verlag, pp. 67–75.

Gutmann, M. U., 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.

Heck, L. P., Hakkani-Tür, D., Tür, G., 2013. Leveraging knowledge graphs for web-scale unsupervised semantic parsing. In: INTERSPEECH. pp. 1594–1598.

Iacobacci, I., Pilehvar, M. T., 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.

Ito, M., Nakayama, K., Hara, T., 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. ACM, pp. 817–826.

Jackson, D. A., Somers, K. M., Harvey, H. H., 1989. Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? American Naturalist, 436–453.

Le, Q. V., Mikolov, T., 2014. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning.

Maurer, H. A., Kappe, F., Zaka, B., 2006. Plagiarism-a survey. J. UCS 12 (8), 1050–1084.

Mcnamee, P., Mayfield, J., 2004. Character n-gram tokenization for European language text retrieval. Information Retrieval 7 (1), 73–97.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. In: Proceedings of Workshop at International Conference on Learning Representations. pp. 1–12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26. pp. 3111–3119.

Mnih, A., 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., Baeza-Yates, R. A., 2001. Flexible comparison of conceptual graphs. In: Proc. of the 12th International Conference on Database and Expert Systems Applications (DEXA). pp. 102–111.

Morin, F., Bengio, Y., 2005. Hierarchical probabilistic neural network language model. In: Proceedings of the international workshop on artificial intelligence and statistics. Citeseer, pp. 246–252.

Moro, A., Raganato, A., Navigli, R., 2014. Entity linking meets word sense disambiguation: A unified approach. Transactions of the Association for Computational Linguistics (TACL) 2, 231–244.

Navigli, R., Jurgens, D., Vannella, D., 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In: Proceedings of the 7^{th} International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013). pp. 222–231.

Navigli, R., 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., 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., Ponzetto, S. P., 2012b. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250.

Navigli, R., Ponzetto, S. P., 2012c. 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., Ney, H., 2003. A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51.

Page, L., Brin, S., Motwani, R., Winograd, T., 1998. The PageRank Citation Ranking: Bringing Order to the Web. Tech. rep., Stanford Digital Library Technologies Project.

Pennington, J., Socher, R., Manning, C. D., 2014. Glove: Global vectors for word representation. Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014) 12, 1532–1543.

Pinto, D., Civera, J., Barrón-Cedeno, A., Juan, A., Rosso, P., 2009. A statistical approach to crosslingual natural language tasks. Journal of Algorithms 64 (1), 51–60.

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., Rosso, P., 2010a. Overview of the 2nd international competition on plagiarism detection. In: CLEF (Notebook Papers/LABs/Workshops).

Potthast, M., Barrón-Cedeño, A., Stein, B., 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., Rosso, P., 2011b. Overview of the 3rd int. competition on plagiarism detection. In: CLEF (Notebook Papers/Labs/Workshop).

Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B., 2014. Overview of the 6th international competition on plagiarism detection. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 845–876.

Potthast, M., Hagen, M., Göring, S., Rosso, P., Stein, B., Sep. 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. Vol. 1391 of CEUR Workshop Proceedings. CLEF and CEUR-WS.org, p. n/a.

Potthast, M., Stein, B., Anderka, M., 2008. A wikipedia-based multilingual retrieval model. In: Advances in Information Retrieval. Springer, pp. 522–530.

Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P., 2010b. An evaluation framework for plagiarism detection. In: Proceedings of the 23rd international conference on computational linguistics: Posters. Association for Computational Linguistics, pp. 997–1005.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D., 2006. The jrc-acquis: A multilingual aligned parallel corpus with +20 languages. In: Proc. 5th Int. Conf. on language resources and evaluation (LREC'06).

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.

Ye, Z., Huang, X., 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. ACM, pp. 690–691.

Yih, W.-t., Toutanova, K., Platt, J. C., Meek, C., 2011. Learning discriminative projections for text similarity measures. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, pp. 247–256.