Franco-Salvador, M.; Gupta, PA.; Rosso, P.; Banchs, R. (2016). Cross-language Plagiarism Detection over Continuous-space- and Knowledge Graph-based Representations of Language. Knowledge-Based Systems. 111:87-99. doi:10.1016/j.knosys.2016.08.004.

Cross-language plagiarism detection over continuous-space representations of language

Marc Franco-Salvador[a,1,*], Parth Gupta[a,1], Paolo Rosso[a], Rafael E. Banchs[b]

[a] *Pattern Recognition and Human Language Technology (PRHLT) Research Center Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain*
[b] *Institute for Infocomm Research, Singapore 138632*

Abstract

Cross-language (CL) plagiarism detection aims at detecting plagiarised fragments of text among documents in different languages. In this work we perform a comparison of different methods that make use of continuous-space representations of language to perform the task of CL plagiarism detection. We also present continuous word alignment-based similarity analysis, a new model to estimate similarity between text fragments. In addition, we study the combination of our continuous representations with the knowledge-based similarity analysis model. We compare the aforementioned approaches with several state-of-the-art models and studied their performance in detecting different length and obfuscation types of plagiarism cases. We conduct experiments over Spanish-English and German-English datasets. Experimental results show that continuous representations allow the continuous word alignment-based similarity analysis model to obtain competitive results and the knowledge-based similarity analysis model to outperform existing state-of-the-art in CL plagiarism detection.

Keywords: Cross-language, Plagiarism detection, Continuous representations, Knowledge graphs, Multilingual Semantic Network, Evaluation

*Corresponding author.
Email addresses: mfranco@prhlt.upv.es (Marc Franco-Salvador), pgupta@dsic.upv.es (Parth Gupta)

[1] The first two authors contributed equally to this work.

1. Introduction

Automatic plagiarism detection refers to the task of automatically identifying which fragment of text is plagiarised. Automatic plagiarism detection task involves finding plagiarised fragments f_q from a suspicious document d_q along with the source fragment f_s from a collection of source documents D. In cross-language setting, suspicious and source documents are written in different languages, referred as cross-language plagiarism detection (Potthast et al., 2011a; Barrón-Cedeño et al., 2013). This study aims to study different dimensions of external plagiarism detection techniques for CL plagiarism detection.

There exist many approaches to CL plagiarism detection (Potthast et al., 2008; Barrón-Cedeño, 2012; Gupta et al., 2012; Franco-Salvador et al., 2013, 2016). For the first time, we analyse the performance of continuous models such as Siamese neural network (S2Net) (Yih et al., 2011), bilingual autoencoder (BAE) (Gupta et al., 2014; Lauly et al., 2014), external data composition neural network (XCNN) (Gupta et al., 2015) for cross-language representation for CL plagiarism detection. We also investigate an alternative for continuous word composition when measuring similarity between texts. The Continuous Word Alignment-based Similarity Analysis (CWASA) employs directed word alignments on top of the word embeddings to measure the distance between two texts. This study also investigates the Knowledge-Based Similarity analysis (KBSim) model and its combination with the aforementioned continuous models. KBSim combines relevance cues from knowledge graphs — generated by means of a multilingual semantic network —, and vector space models (VSM) for estimating cross-language similarity. The objective of this study is to analyse the viability and performance of continuous models as the vector component of KBSim instead of the VSM. Compared with that representation, continuous models offered in the past higher performance when measuring text (Platt et al., 2010) similarity and may increase KBSim performance.

We carry experiments on standard plagiarism dataset PAN-PC-2011 for two languages Spanish-English (ES-EN) and German-English (DE-EN) in two settings: *i)* entirely plagiarised suspicious-source document linking (Expt. A); and *ii)* plagiarised fragments identification within entire documents (Expt. B). We also present an extensive analyses on performance of these algorithms for different length and types of plagiarism cases. Our experiments show that though continuous models have a small coverage (20k words) and trained on

a parallel corpus of a limited size, they show strong performance, especially when composed with CWASA, compared to many vector space models which have full coverage. Moreover, when combined together using Knowledge Base similarity analysis (KBsim) (Franco-Salvador et al., 2014), performance is superior than any other model alone which points to the fact that knowledge graph and continuous-based models capture different aspects of cross-lingual similarity for CL plagiarism detection.

The rest of the paper is structured as follows: In Section 2 we present related work on cross-language plagiarism detection and continuous models for cross-language similarity estimation. We detail state-of-the-art approaches for CL plagiarism detection and continuous representation-based models in Section 3 and 4 respectively. Section 5 covers the details on KBSim model. We present our experimental framework with results and analyses in Section 6. Finally, in Section 7 we summarise concluding remarks.

2. Related work

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 the Romance and the 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 schema 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 represent each text by its similarities with a multilingual document collection. Using a multilingual document collection with comparable documents across languages, e.g. Wikipedia, the resulting vectors from different languages can be compared directly.

The use of parallel corpora has been explored too. 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 to then perform an alignment of the texts. The alignment takes into account the probabilities of translations and the differences of length of equivalent texts in different languages.

Other type of approaches exploit multilingual knowledge bases or semantic thesauri. 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 excelling specially for Spanish-English. In contrast, the Cross-Language Knowledge Graph Analysis (CL-KGA) model (Franco-Salvador et al., 2013, 2016) uses a multilingual semantic network to create knowledge graphs in order to model the context of documents. This knowledge graph representation covers aspects such as concept relatedness, vocabulary expansion or word sense disambiguation. CL-KGA obtained state-of-the-art results for CL plagiarism detection, also in cases with paraphrasing. In addition, the Knowledge-Based Similarity analysis (KBSim) (Franco-Salvador et al., 2014) model has been presented as an improved version of CL-KGA for CL document retrieval and categorisation. Apart from knowledge graphs, this model also includes a vector component. However, it has not been evaluated yet for 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 in order to compete at identifying the 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. 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 put forward a heavy dependence on availability of Machine Translation (MT) systems in the involved languages 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 which do not depend on complete MT systems.[4]

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

[3] `pan.webis.de/`

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

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 concludes that T+MA is superior in short cases of plagiarism but very close to CL-ASA, which offers a higher precision in all the experiments and better performance for long cases of plagiarism.

In (Franco-Salvador et al., 2016) CL-KGA was compared with CL-CNG, CL-ESA and CL-ASA obtaining the highest results in Spanish-English and German-English plagiarism detection. In addition, 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 in function of the task, languages, and dataset employed. For instance, CL-ESA and CL-CNG were more stable across datasets, obtaining a higher performance on the comparable Wikipedia dataset. In contrast, CL-ASA obtained better results on the parallel JRC-Acquis dataset. Finally, CL-CNG reduced the quality for language pairs without lexical and syntactic similarities. Therefore, for the sake of completeness in this work we decided to compare our continuous word representation models with the CL-CNG, CL-ESA, CL-ASA, and CL-KGA models plus KBSim, the CL-KGA refinement.

There is also a significant advancement in the area of coninuous representation learning for text generally referred as latent semantic models. Latent semantic models map the high dimensional term vectors into a low dimensional abstract space referred to as latent space. There are broadly two categories of approaches: *i)* generative topic models, and *ii)* projection based models. Generative topic models, like latent dirichlet allocation (LDA), represent the high dimensional term vectors in a low-dimensional latent space of hidden topics. The projection based methods, like latent semantic analysis (LSA), learn a projection operator to map high-dimensional term vectors to low-dimensional latent space (Deerwester et al., 1990; Blei et al., 2003; Hinton and Salakhutdinov, 2006; Mikolov et al., 2013b). There also exist cross-lingual variants of these models which try to learn embeddings of text in cross-language space. Cross-language latent semantic indexing (CL-LSI) is a cross-linguage extension of latent semantic indexing (LSI) (Dumais et al., 1997). Oriented principle component analysis (OPCA) tries to learn 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., 2014; Chandar A. P. et al., 2014) and composition neural networks (Gupta et al., 2015). In this study we examine the performance of many of these models for CL plagiarism detection and more details on them are presented in Section 4.

3. Methods for cross-language plagiarism detection

In this section we describe more in detail the state-of-the-art methods for cross-language plagiarism detection. In order to detect plagiarised sections of text between two documents d_L and $d'_{L'}$ written in languages L and L', we first segment them to obtain the sets of fragments FC $\in d_L$ and FC$' \in d'_L$.[5] Next, we use a model to obtain the set of cross-language similarities SF $= \{\mathrm{S}(F, F')\}$ between all the pairs of text fragments (F, F'), $F \in$ FC and $F' \in$ FC$'$. Once we obtain the set SF, 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 is described in Algorithm 1.[6] This algorithm has been used for evaluating all the models compared in the second experiment of the Section 6.4.

In this work, to obtain the set of cross-language similarities SF, we compare our proposed approaches (see sections 4 and 5) with several state-of-the-art models: CL-CNG, CL-ESA, CL-ASA and CL-KGA. We also employ an approach we call Vector Space Model (VSM), which is later employed in Section 5 as part of the knowledge-based similarity analysis model.

3.1. Cross-Language Character n-Grams

CL-CNG is a very simple model which decomposes the text in two languages into smaller units such as character n-grams. Standard normalisation techniques are applied such as lower-casing and diacritics removal. CL-CNG was originally proposed by Mcnamee and Mayfield (2004) for cross-language information retrieval. We used $n = 3$ for our experiments and similarity

[5] We use a 5-sentence sliding window with a 2-sentence step to make the segmentation into fragments.

[6] 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 $\mathrm{thres}_2 = 2$.

Algorithm 1: Detailed analysis and postprocessing.

Input : the set of similarities $\mathrm{SF} = \{\mathrm{S}(F, F')\}$ between all the pairs of text fragments (F, F'), $F \in \mathrm{FC}$ and $F' \in \mathrm{FC}'$, $\mathrm{FC} \in d_L$ and $\mathrm{FC}' \in d'_{L'}$

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

```
PlagCases ← {}
// DETAILED ANALYSIS STEP:
foreach F ∈ FC do                          // For each text fragment of d_L...
    P_F ←argmax^5_{F'∈FC'} S(F, F')
    // P_F contains the top 5 most similar fragments to d'_L'
    // POSTPROCESSING STEP:
    repeat                                 // Repeat until convergence...
        // For each combination between fragments in P_F...
        foreach p_i ∈ P_F do
            foreach p_j ∈ P_F, i ≠ j do
                /* if the distance in d_L between two fragments of P_F
                   is lower than thres_1, merge them:                  */
                if δ(p_i, p_j) < thres_1 then
                    merge_fragments(p_i, p_j)
                /* δ(.,.) returns the distance in characters between two
                   fragments using their beginning and end offsets.    */
    until no change
    /* Select as plagiarism cases those in P_F which combine more than
       thres_2 fragments:                                              */
    PlagCases = PlagCases ∪ {offsets(p ∈ P_F | |p| > thres_2)}
    /* offsets(·) returns the beginning and end offsets of a
       plagiarism case.                                                */
return PlagCases
```

between such text representation is computed using cosine similarity as recommended in Potthast et al. (2011a).

3.2. Cross-language explicit semantic analysis

Cross-Language Explicit Semantic Analysis (CL-ESA) (Potthast et al., 2008) extends the explicit semantic analysis (Gabrilovich and Markovitch,

2007) model 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 document is qualified using the reference collection D. Even though the indexing with D is performed at monolingual level, using a multilingual document collection with comparable documents across languages, e.g. Wikipedia[7], the resulting vectors from different languages can be compared directly. Formally, having a matrix D_L where rows represent documents of a collection in a language L, a document d_L is indexed as follows:

$$d_{D_L} = D_L \cdot d_L^T, \tag{1}$$

where d_{D_L} denotes the resulting indexed vector of document d_L in D_L. Documents represented in d_L and D_L use a vector representation such as VSM with term frequency-inverse document frequency (Salton et al., 1983) (TF-IDF) weighting. The similarity between two documents d_L and $d'_{L'}$ is estimated as $\varphi(d_{D_L}, d'_{D_{L'}})$, where φ is a vector similarity function such as the cosine similarity, and d_{D_L} and $d_{D_{L'}}$ are comparable document collections across L and L' languages.

3.3. Cross-language alignment-based similarity analysis

CL-ASA measures the similarity between two document by adapting Bayes' rule for machine translation — composition of language model and translation model (Barrón-Cedeño et al., 2008). It computes the likelihood of d' to be translation of d as shown in Eq. 2.

$$\mathrm{S}(d, d') = \varrho(d')\ \mathrm{p}(d \mid d'). \tag{2}$$

CL-ASA uses $\varrho(d')$ component as length model which captures the translation length factor as defined in (Pouliquen et al., 2003). The translation model depicted by conditional probability $\mathrm{p}(d \mid d')$ in Eq. 2 is replaced by a statistical bilingual dictionary score and computed as shown in Eq. 3

$$\rho(d \mid d') = \sum_{x \in d} \sum_{y \in d'} \mathrm{p}(x, y)\ \ , \tag{3}$$

where, $\rho(d \mid d')$ no longer represents a probability measure and the dictionary $\mathrm{p}(x, y)$ defines the likelihood of word x of being a valid translation of y. The

[7] `https://es.wikipedia.org/`

CL-ASA model is trained as per the parameters reported in Barrón-Cedeño et al. (2013).

3.4. Cross-language knowledge graph analysis

Cross-language Knowledge Graph Analysis (CL-KGA) (Franco-Salvador et al., 2013, 2016) represents documents in a semantic graph space by means of knowledge graphs. A knowledge graph is created as a subset of a multilingual semantic network, e.g. BabelNet (Navigli and Ponzetto, 2012), focused in the concepts belonging to a text. These graphs include several interesting characteristics such as Word Sense Disambiguation (WSD), vocabulary expansion and language independence. Therefore, knowledge graphs created from documents in different languages can be directly compared. Formally, having a pair of graphs (G, G'), $G \in d_L$ and $G' \in d'_{L'}$, the similarity $S_g(G, G')$ between them is separately estimated for concepts and relations. The similarity between the concepts is calculated using Dice's coefficient (Jackson et al., 1989):

$$\mathrm{S}_c(G, G') = \frac{2 \cdot \sum_{c \in \mathrm{V}(G) \cap \mathrm{V}(G')} \mathrm{w}(c)}{\sum_{c \in \mathrm{V}(G)} \mathrm{w}(c) + \sum_{c \in \mathrm{V}(G')} \mathrm{w}(c)}, \tag{4}$$

where $\mathrm{V}(G)$ is the set of concepts in the graph and $\mathrm{w}(c)$ is the weight of a concept c. Likewise, the similarity between the relations is calculated as:

$$\mathrm{S}_r(G, G') = \frac{2 \cdot \sum_{r \in \mathrm{E}(G) \cap \mathrm{E}(G')} \mathrm{w}(r)}{\sum_{r \in \mathrm{E}(G)} \mathrm{w}(r) + \sum_{r \in \mathrm{E}(G')} \mathrm{w}(r)}, \tag{5}$$

where $\mathrm{E}(G)$ is the set of relations in the graph and $\mathrm{w}(r)$ is the weight of a semantic relation r. Finally, the two above measures of conceptual (S_c) and relational (S_r) similarity are interpolated to obtain an integrated measure $\mathrm{S}_g(G, G')$ between knowledge graphs:

$$\mathrm{S}_g(G, G') = a \cdot \mathrm{S}_c(G, G') + b \cdot \mathrm{S}_r(G, G'), \tag{6}$$

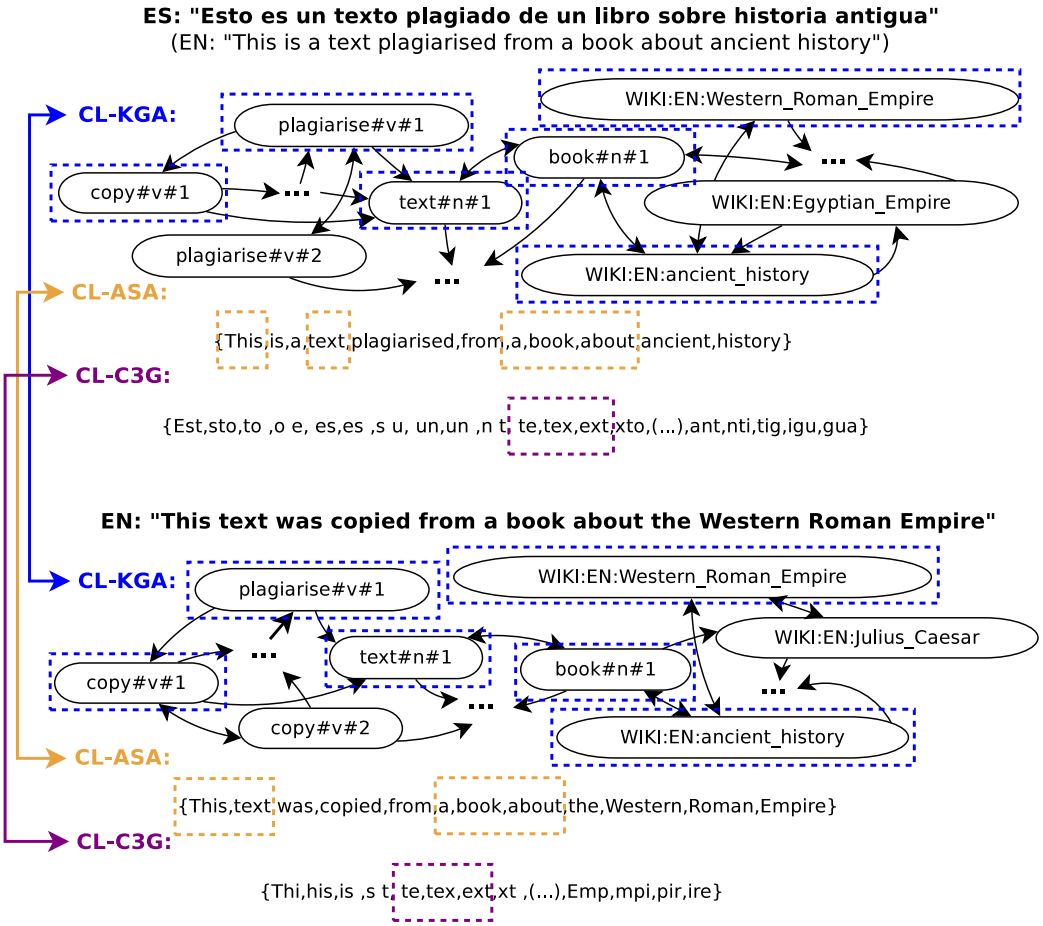

Figure 1: Toy example to illustrate the capability of detection of the CL-KGA model compared to the CL-ASA and the CL-C3G models. Higher intersection of same-coloured boxes between languages represents a higher potential plagiarism case retrieval. © Elsevier. Reproduced with permissions.

where a and b, $a+b=1$, are the parameters of the relevance of concepts and relations respectively.[8]

In this work concepts are weighted using their graph outdegree (Navigli and Ponzetto, 2012). In contrast, relations are weighted using the original weights between relations provided in BabelNet. These weights where

[8]In this work we used the optimal values provided in Franco-Salvador et al. (2016) for concepts and relations which correspond with equal relevance for a and b.

calculated using an extension of the extended gloss overlap measure (Banerjee and Pedersen, 2003) which weights semantic relations between WordNet (Fellbaum, 1998) and Wikipedia concepts.[9] In Figure 1 we can see the differences among CL-KGA, CL-C3G, and CL-ASA when detecting CL plagiarism. Thanks to the aforementioned characteristics, the use of knowledge graphs allows to detect similarity even when the paraphrasing is employed and the languages are not syntactically and semantically related. For more details about the CL-KGA model please refer to the original works (Franco-Salvador et al., 2013, 2016).

3.5. Vector space model

Vector Space Model (VSM) is a simple approach which represents documents using the TF-IDF weighting[10] and compares them using the cosine similarity. In order to make this approach cross-lingual and to counterbalance possible translation errors, we followed Franco-Salvador et al. (2014) and represented each document d_L in a bilingual form $d_{LL'}$ by concatenating the vector d_L with the vector $d_{L'}$ which contains its translations using an statistical dictionary[11] with TF-IDF re-weighting in function of the probabilities of translation of the words.

We included this representation in order to later in Section 5 provide the knowledge-based similarity measure with VSM as its vector component.

4. Continuous representations for cross-language plagiarism detection

This Section presents details of the continuous represenation learning algorithms for cross-language similarity estimation. These models are usually categorised according to the objective function they optimise and the type of data they take in. Most of these models learn cross-lingual embeddings 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

[9] Although Franco-Salvador et al. (2016) provided a new weighting schema for relations based on continuous representations of concepts, their knowledge graph construction was penalised in terms of computational time when using them. Therefore, in this work we decided to use the classical ones in order to speed up the process.

[10] Preprocessing includes tokenisation, and stop-word and punctuation removal.

[11] We used the same dictionary employed with CL-ASA.

and English-German parallel sentences from DGT-Translation Memory distributed by JRC[12]. For monolingual pre-initialisation in XCNN we use CLEF ad-hoc retrieval corpus document titles.

4.1. Similarity Learning via Siamese Neural Network (S2Net)

Following the general Siamese neural network architecture (Bromley et al., 1993), S2Net trains two identical neural networks concurrently. The S2Net takes in 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 cosine similarity. The projection operation can be described as follows:

$$y_d = W * x_d \tag{7}$$

where, x_d is the input term vector for 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 according to the details provided in Yih et al. (2011).

4.2. Bilingual Autoencoder (BAE)

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 (Gupta et al., 2014; Lauly et al., 2014; Chandar A. P. et al., 2014). These networks learn cross-language associations by optimising the reconstruction error of the cross-language data.

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 Fig. 2. For more details, please refer to Gupta et al. (2014).

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

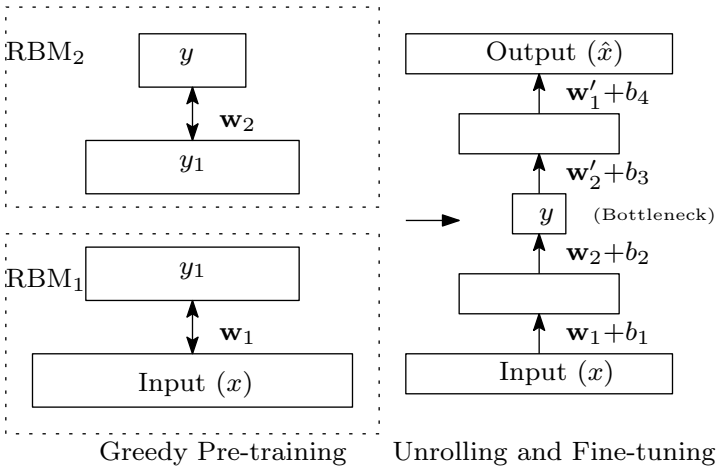

Figure 2: **Left panel**: pre-training of stacked RBMs where the upper RBM takes as input the output of the lower RBM. **Right panel**: After pre-training the structure is "unrolled" to create a multi-layer network which is fine-tuned by means of backpropagation to learn an identity function $\hat{x} \approx x$.

4.3. External data Composition Neural Networks (XCNN)

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 only 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 crosslingual 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.

The architecture of XCNN model training is shown in Fig. 3. XCNN learns word embeddings in cross-lingual setting using objective function defined in Eq. 8. It maximises the cosine similarity for a training example for a positive sample and miminises it for a negative sample. The network parameters are updated through backpropagation technique.

$$J_{cl}(\theta) = \cos(y_{l_1}, y^+_{l_2}) - \cos(y_{l_1}, y^-_{l_2}) \tag{8}$$

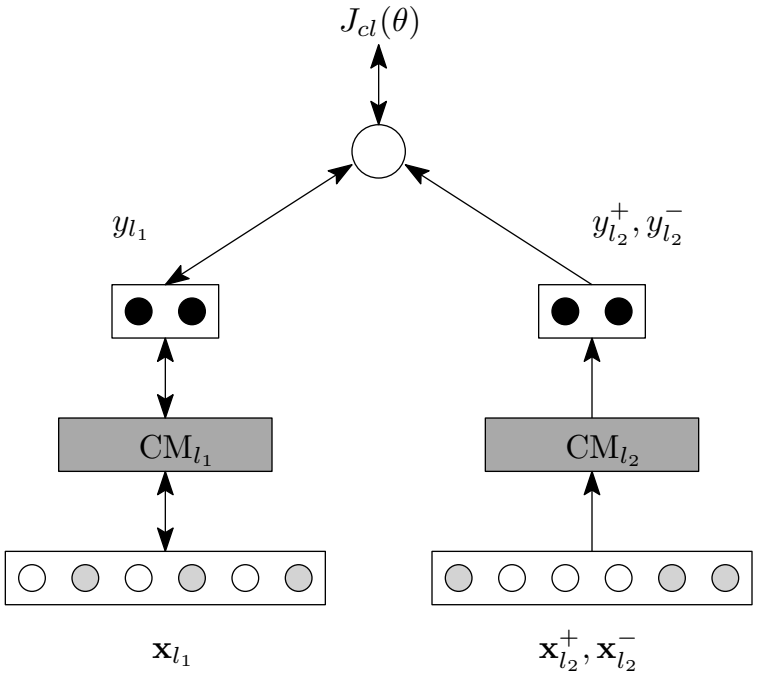

Figure 3: Architecture of external-data composition neural network model for cross-lingual training.

The representation for an input text is obtained through an addition composition function as described below.

$$\begin{aligned} y_i^{(l_1)} &= \text{g}(W_1 * x_i + b_1) \\ y_i^{(l_j)} &= \text{g}(W_j * y_i^{l_{(j-1)}} + b_j), j = 2, ..., m \\ y &= \sum_{i=1}^{n} y_i^{(l_m)} \end{aligned} \tag{9}$$

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

4.4. Continuous word alignment-based similarity analysis

The aforementioned continuous 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 explain an alternative method to combine word level vectors by means of alignments to represent text. The Continuous Word Alignment-based Similarity Analysis (CWASA) model modifies the text-to-text relatedness proposed by (Hassan and Mihalcea, 2011) in order to estimate the similarity between documents by efficiently aligning their continuous 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 both direction the closest to *First_Lady*. Formally, the similarity $\mathrm{S}(d, d')$ between two documents d and d' is estimated as follows:

$$\mathrm{S}(d, d') = \frac{1}{|\Phi|} \sum_{c_k \in \Phi} c_k, \tag{10}$$

where $d = (x_1, ..., x_n)$ and $d' = (y_1, ..., y_m)$ are represented as lists of continuous words, and Φ is generated from the list $\Phi' = \{c'_1, ..., c'_{n+m}\}$ that satisfies Equation 11:

$$c'_k = \begin{cases} \arg \max_{i=k, x_i \in d, y_j \in d'} \varphi(x_i, y_j), & \text{if } k \leq n \\ \arg \max_{j=k-n, x_i \in d, y_j \in d'} \varphi(x_i, y_j), & \text{otherwise} \end{cases} \tag{11}$$

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[13] that represents the strongest semantic pairing between the continuous words of documents d and d'.

Basically, in Eq. 11 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. 10 to estimate the similarity score between d and d' as the average of the different alignments. We note that this problem can be efficiently solved by dynamic programming. In addition, although this work is focused in a cross-lingual setting, CWASA could be directly employed with monolingual continuous

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

word representations (Mikolov et al., 2013a,b). We compare our CWASA model with the classical bag-of-words sum representation in Section 6.

5. Hybrid models for cross-language plagiarism detection

The knowledge graphs generated with CL-KGA will only cover and relate the most central concepts of a document. This is produced for the use of a knowledge base (e.g. BabelNet) as core of the model. In general, this is adequate when measuring similarity among documents of different topics, but may not be enough to detect similarities in verbal tenses, out-of-vocabulary words, or punctuation, e.g. similarity between "I wait 4 u" and "I ' m waiting 4 u"). In contrast, more traditional representations such as the VSM will be able to detect this small differences but will fail when detecting similarity between topical-related documents, e.g. similarity between "I love the capital of France in July" and "I like Paris in summer").

The Knowledge-Based Similarity analysis (KBSim) (Franco-Salvador et al., 2014) model extends the CL-KGA model in order to combine both the benefits of the knowledge graph and the multilingual vector-based representations. Key to this approach is the combination of both representations in function of the relevance of the knowledge graphs. This allows to increase the contribution of multilingual vectors in case of non-informative graphs. 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:

$$\mathrm{S}(d, d') = \mathrm{c}(G)\mathrm{S}_g(G, G') + (1 - \mathrm{c}(G))\mathrm{S}_v(\vec{v}, \vec{v}'), \tag{12}$$

where $\mathrm{S}_g(G, G')$ is the knowledge graph similarity of Eq. 6, $\mathrm{S}_v(\vec{v}, \vec{v}')$ is the vector-based similarity, and $\mathrm{c}(G)$ is an interpolation factor calculated as the edge density of knowledge graph G:

$$\mathrm{c}(G) = \frac{|\mathrm{E}(G)|}{|\mathrm{V}(G)|(|\mathrm{V}(G)| - 1)}. \tag{13}$$

Note that, using the factor $\mathrm{c}(G)$ to interpolate the two similarities in Eq. 12, the relevance for the knowledge graphs and the multilingual vectors is determined in a dynamic way. Indeed, $\mathrm{c}(G)$ makes the contribution of graph similarity depend on the richness of the knowledge graph.

The vector-based similarity $\mathrm{S}_v(\vec{v}, \vec{v}')$ was originally calculated with the VSM introduced in Section 3.5. However, in this work we are also comparing

vector representations based on continuous words. In consequence, we are also interested into analyse if the combination with such representations complements knowledge graphs better than VSM. Therefore, in Section 6 we will compare in total four additional models: KBSim (VSM), KBSim (S2Net), KBSim (BAE), and KBSim (XCNN).

6. Evaluation

In this section we compare the different models in the task of CL plagiarism detection. We first describe the datasets and methodology employed. Next, we present the results and analysis of the experiments.

6.1. Datasets

To evaluate the models we selected the PAN-PC-11[14] dataset that was created for the 2011 CL plagiarism detection competition of PAN at CLEF.[15] The dataset consist of Spanish-English (ES-EN) and German-English (DE-EN) partitions for CL plagiarism detection. The plagiarism cases were generated using translation obfuscation with Google translate.[16] In addition, PAN-PC-11 contains also cases of plagiarism with manual obfuscation after automatic translation.[17] These cases are CL paraphrasing cases of plagiarism. In Table 6.1 we can see the statistics of the dataset.

6.2. Methodology

In order to evaluate the models, employing always both ES-EN and DE-EN language partitions, we perform two different experiments. In Section 6.3, 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 detailed in Algorithm 1. The 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, in

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

[15] `http://www.clef-initiative.eu/`

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

[17] Although there exists an alternative dataset, PAN-PC-10 (`http://www.uni-weimar.de/en/media/chairs/webis/corpora/corpus-pan-pc-10/`), we selected PAN-PC-11 due to this type of cases, which were not present in the 2010 edition.

Spanish-English documents		**German-English 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 1: Statistics of PAN-PC-11 cross-language plagiarism detection partitions.

Section 6.4 we conduct a second experiment. There, detections are filtered using Algorithm 1 to determine what cases are really involved in plagiarism. As evaluation metric of the experiment 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:

$$\text{precision}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\bigcup_{s \in S}(s \sqcap r)|}{|r|}; \tag{14}$$

$$\text{recall}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\bigcup_{r \in R}(s \sqcap r)|}{|s|}, \tag{15}$$

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:

$$\text{granularity}(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|, \tag{16}$$

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 $R_s = \{r | r \in R \wedge r$

detects $s\}$. The three previous measures were integrated together in order to obtain an overall score for plagiarism detection (plagdet):

$$\text{plagdet}(S, R) = \frac{F_1(S, R)}{\log_2(1 + \text{granularity}(S, R))}. \quad (17)$$

In both experiments of Section 6.3 and 6.4 we also included in a separated subsection the analysis of results in function of the type of obfuscation and document length of the plagiarism cases.

We compare the continuous word representation models, S2Net, BAE, and XCNN (cf. Section 4) with the state-of-the-art CL-C3G[18], CL-ESA,[19] CL-ASA,[20], VSM, and CL-KGA[21] models (cf. Section 3). In addition, we use our CWASA model (cf. Section 4.4) in order to represent documents by means of continuous word alignments: CWASA (S2Net), CWASA (BAE), and CWASA (XCNN). Finally, we show the performance of the original KB-Sim model, KBSim (VSM), and the results when replacing the vector component (VSM) for the document vectors of the continuous word representation models: KBSim (S2Net), KBSim (BAE), and KBSim (XCNN). 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.

6.3. 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 6.3.1 we analyse the results in function of the type of plagiarism case. In Table 2 we

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

[19]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.

[20]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.

[21]Details about the tuning of the paramenters of CL-KGA and CL-ESA are provided in Franco-Salvador et al. (2016)

	Model	Spanish-English				German-English			
		R@1	**R@5**	**R@10**	**R@20**	**R@1**	**R@5**	**R@10**	**R@20**
(a)	CL-KGA	**0.917**	**0.946**	**0.956**	**0.961**	**0.786**	**0.865**	**0.893**	**0.911**
	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)	0.920	0.949	0.956	0.961	**0.809**	**0.878**	**0.901**	**0.921**
	KBSim (VSM)	**0.927**	**0.955**	**0.961**	**0.965**	0.794	0.871	0.896	0.915
	KBSim (BAE)	0.917	0.945	0.956	0.962	0.791	0.870	0.893	0.911
	KBSim (XCNN)	0.858	0.907	0.924	0.935	0.741	0.843	0.872	0.897

Table 2: Spanish-English and German-English performance analysis in terms of R@k, where $k = \{1, 5, 10, 20\}$.

show the results for ES-EN and DE-EN.[22] As we can see, DE-EN similarity has been 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, obtained 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 benefited VSM that obtained interesting results too. It is followed, in order of performance, by CL-ASA, CL-ESA, and CL-C3G, that has been the baseline in all our experiments.

Compared to the state-of-the-art, the continuous representation models of group (b) offered and average performance. The S2Net model obtained 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.

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

Therefore, embeddings of large fragments of text are still representative. In contrast, XCNN learns word-level embeddings and hence when projecting a large fragment of text ($\sim$1000 words) the summed embeddings flattens vectors and loose discriminative power, affecting XCNN performance. However, these comments refer 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, produced 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 6.3.1, the use of CWASA allows to successfully measure similarity between texts of any length. This allowed 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 is not outperforming the CL-KGA model, for computational time constraints we restricted the vocabulary to 20,000 words when using continuous representations, and we are rivalling with a model that employs BabelNet, a 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 languages. A higher variety of stemmed words was observed for the German agglutinative language, which have not been covered by the vocabulary in the same amount than the other languages. We also note that performance of BAE shows the highest variation from R@1 to R@5 among all models: $\sim$21% After the manual analysis of the resulting embeddings and the values of similarity between texts, we observed a very reduced variance: lower than $\sim 10^{-2}$. This led the model to be less precise when differentiating close elements and affected the performance of CWASA (BAE).

Finally, the combination of knowledge graphs with vectors produced the best results. Thanks to the dynamic interpolation, the original KBSim (VSM) model obtained higher results than CL-KGA and VSM separately. We appreciate that the use of continuous vector representations allows to successfully complement knowledge graphs too. KBSim (S2Net) obtained in average the highest results of this experiment. Although KBSim (XCNN) do not obtained such high results, the differences in R@5 are small. As we will see in Section 6.4, such differences are not relevant when detecting plagiarism and the models performance may change in function of the postprocessing

Type of obfuscation		Model	Spanish-English R@1	R@5	R@10	R@20	German-English R@1	R@5	R@10	R@20
Translated manual obfuscation	(a)	CL-KGA	**0.846**	**0.908**	**0.930**	**0.939**	**0.710**	**0.801**	**0.851**	**0.864**
		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)	0.853	0.912	0.925	0.940	**0.729**	**0.806**	**0.839**	**0.875**
		KBSim (VSM)	**0.867**	**0.924**	**0.935**	**0.942**	0.714	0.800	0.853	0.870
		KBSim (BAE)	0.847	0.901	0.925	0.939	0.715	0.799	0.839	0.863
		KBSim (XCNN)	0.764	0.840	0.874	0.893	0.641	0.774	0.830	0.873
Translated automatic obfuscation	(a)	CL-KGA	**0.922**	**0.948**	**0.958**	**0.962**	**0.794**	**0.872**	**0.897**	**0.916**
		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)	0.925	0.952	0.958	0.962	**0.817**	**0.885**	**0.908**	**0.925**
		KBSim (VSM)	**0.932**	**0.957**	**0.963**	**0.966**	0.802	0.879	0.900	0.920
		KBSim (BAE)	0.923	0.948	0.958	0.964	0.799	0.877	0.899	0.916
		KBSim (XCNN)	0.865	0.912	0.928	0.938	0.751	0.850	0.877	0.899

Table 3: Spanish-English performance analysis in terms of type of obfuscation for the plagiarism cases and R@k, where $k = \{1, 5, 10, 20\}$.

algorithm employed. In Section 6.4 we will also study the statistical differences of all the models to analyse if these conduct to significant differences when detecting plagiarism. We note that with the current parameters of Algorithm 1, R@5 is the recall upper-bound for the plagiarism detection performed in Section 6.4.

Case length		Model	Spanish-English R@1	R@5	R@10	R@20	German-English R@1	R@5	R@10	R@20
Long length cases	(a)	CL-KGA	**0.935**	**0.957**	**0.963**	**0.966**	**0.807**	**0.883**	**0.905**	**0.924**
		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)	0.938	0.958	0.962	0.967	**0.831**	**0.896**	**0.917**	**0.932**
		KBSim (VSM)	**0.944**	**0.963**	**0.967**	**0.969**	0.817	0.890	0.909	0.928
		KBSim (BAE)	0.937	0.956	0.964	0.968	0.812	0.888	0.907	0.924
		KBSim (XCNN)	0.888	0.927	0.939	0.947	0.767	0.857	0.883	0.906
Medium length cases	(a)	CL-KGA	**0.920**	**0.948**	**0.957**	**0.962**	**0.792**	**0.870**	**0.895**	**0.913**
		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)	0.924	0.951	0.957	0.961	**0.816**	**0.884**	**0.906**	**0.924**
		KBSim (VSM)	**0.931**	**0.957**	**0.962**	**0.965**	0.801	0.876	0.898	0.917
		KBSim (BAE)	0.921	0.948	0.957	0.963	0.799	0.874	0.896	0.913
		KBSim (XCNN)	0.868	0.914	0.929	0.939	0.749	0.848	0.876	0.900
Short length cases	(a)	CL-KGA	**0.913**	**0.943**	**0.954**	**0.959**	**0.779**	**0.860**	**0.888**	**0.907**
		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)	0.917	0.947	0.954	0.959	**0.802**	**0.873**	**0.897**	**0.917**
		KBSim (VSM)	**0.924**	**0.953**	**0.959**	**0.963**	0.787	0.866	0.891	0.911
		KBSim (BAE)	0.914	0.943	0.954	0.961	0.785	0.865	0.889	0.907
		KBSim (XCNN)	0.853	0.903	0.921	0.933	0.735	0.838	0.868	0.893

Table 4: Spanish-English performance analysis in terms of plagiarism case length and R@k, where $k = \{1, 5, 10, 20\}$.

6.3.1. Cross-language similarity ranking in function of the type of plagiarism cases

In this section we analyse the R@k of the models in function of the type of plagiarism case. We divide plagiarism cases in function of the type obfuscation — translated obfuscation and translated manual obfuscation — employed to generate the case, and in function of the case length — short, medium, and long[23]. Most of the highlights of Section 6.3 persist when discriminating in function of the type of case. However, there are several points to note. In Table 3, attending to the obfuscation type, note the difficulty of detection. The translated manual obfuscation has manual correction after the automatic translation and generates cases with paraphrasing in order to hide the plagiarism. Therefore, it has been more difficult to detect similarity between such type of cases. Attending to the models, CL-ESA, based on a representation by similarities with a collection of documents, outperformed CL-ASA in cases with manual obfuscation. Logical if we consider that ESA was originally meant for tasks of relatedness rather than plagiarism.

In Table 4 we can see the results in function of the case length. In opposition to the short cases, the similarity between long cases of plagiarism has been the easiest to detect. The additional information that long cases provided, made easier the models to represent and to discriminate between texts. However, those differences in performance rarely exceed 2%. The exception was the CL-ASA model, that suffered a higher decay when cases became shorter. This may be produced for the document length component included inside the model, that is more precise normalising larger cases of plagiarism. Note that KBSim (S2Net) obtained the highest results independently of the type of obfuscation and case length analysed, which highlights its performance for CL similarity analysis and plagiarism detection.

6.4. Experiment B: Cross-language plagiarism detection

In this section we compare the continuous word representation, CWASA, and KBSim models with several state-of-the-art approaches using the PAN-PC-11 dataset for CL plagiarism detection. We show the results in the Table 5. Although both English and German are Germanic languages, due to their grammatical differences, the additional difficulty of detection in DE-EN

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

	Model	Spanish-English				German-English			
		Plag	**Prec**	**Rec**	**Gran**	**Plag**	**Prec**	**Rec**	**Gran**
(a)	CL-KGA	**0.620**	**0.696**	**0.558**	**1.000**	**0.520**	0.601	**0.460**	**1.004**
	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)	**0.644**	**0.765**	0.556	**1.000**	**0.561**	**0.723**	0.463	1.010
	KBSim (S2Net)	0.623	0.701	**0.560**	**1.000**	0.536	0.614	**0.477**	**1.002**
	KBSim (VSM)	0.621	0.697	0.559	**1.000**	0.523	0.599	0.465	**1.002**
	KBSim (BAE)	0.622	0.704	0.557	**1.000**	0.521	0.592	0.468	1.004

Table 5: Spanish-English and German-English performance analysis in terms of plagdet (Plag), precision (Prec), recall (Rec) and granularity (Gran).

is also present 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 were obtained with CL-C3G, that did not found enough lexical and syntactic similarities to model the content properly using character n-grams. The CL-ESA and CL-ASA models obtained a similar recall but the latter one excelled in precision and increased its plagdet. In fact, CL-ESA offered a higher number of false positives and highlighted again its semantic relatedness nature beyond plagiarism. Finally, the CL-KGA model was the best state-of-the-art approach and obtained 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 single as multiple cases of plagiarism or vice versa. Note also that CL-ASA and CL-KGA are in tie in terms of precision. However, CL-KGA excelled in recall (specially in DE-EN).

As we also pointed out in Section 6.3, the continuous word representation models which represent documents based on the sum of word vectors offered

an average performance for this task.[24] The S2Net model outperformed BAE and XCNN but remained inferior CL-KGA. We can see close values in terms of precision between S2Net and XCNN. However, S2Net's recall has been higher than 10% in all the tests. This, joint to the highest granularity, penalised XCNN's plagdet.

The models of group (c) — using CWASA to measure similarity — notably improved the performance of S2Net, BAE, and XCNN. We appreciate how, specially with XCNN, the recall and granularity improved with a low impact in the precision. In contrast to S2Net and BAE, that use a bag-of-words format to learn vectors of documents, XCNN directly generated continuous vectors of words. These vectors found in CWASA and excellent complement in order to accurately measure CL similarity. Note that in this experiment we used Algorithm 1 to analyse the similarities and to determine what is plagiarism. To do this, Algorithm 1 retrieved the five most similar fragments with each text fragment in the other language. This penalised BAE that, as we mentioned in Section 6.3, has a low variance between continuous vectors and made more difficult to correctly align them.

Finally, the combination of vector representations with knowledge graphs, made the KBSim models of group (d) to obtain the overall highest results. In fact, KBSim (XCNN) outperformed the original KBSim (VSM), and was the best model of the evaluation, independently of the language pair analysed. This confirms that knowledge graphs and continuous models capture different aspects of text and complement each other. This also proves the potential of KBSim for the tasks of CL similarity analysis and plagiarism detection.

Despite the high R@k of some models (see Section 6.3), the final values of recall, and consequently plagdet, considerably decreased. 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 potential and limitations of Algorithm 1 for plagiarism detection.

After analysing the performance of the models, we are also interested in analysing which differences are statistically significant. In order to analyse this, we used bootstrap resampling[25] (Efron and Tibshirani, 1994) to mea-

[24] 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.

[25] Bootstrap methods are generally superior to parametric tests for small datasets — as the dataset in hand — or where sample distributions are non-normal. The statistical tests

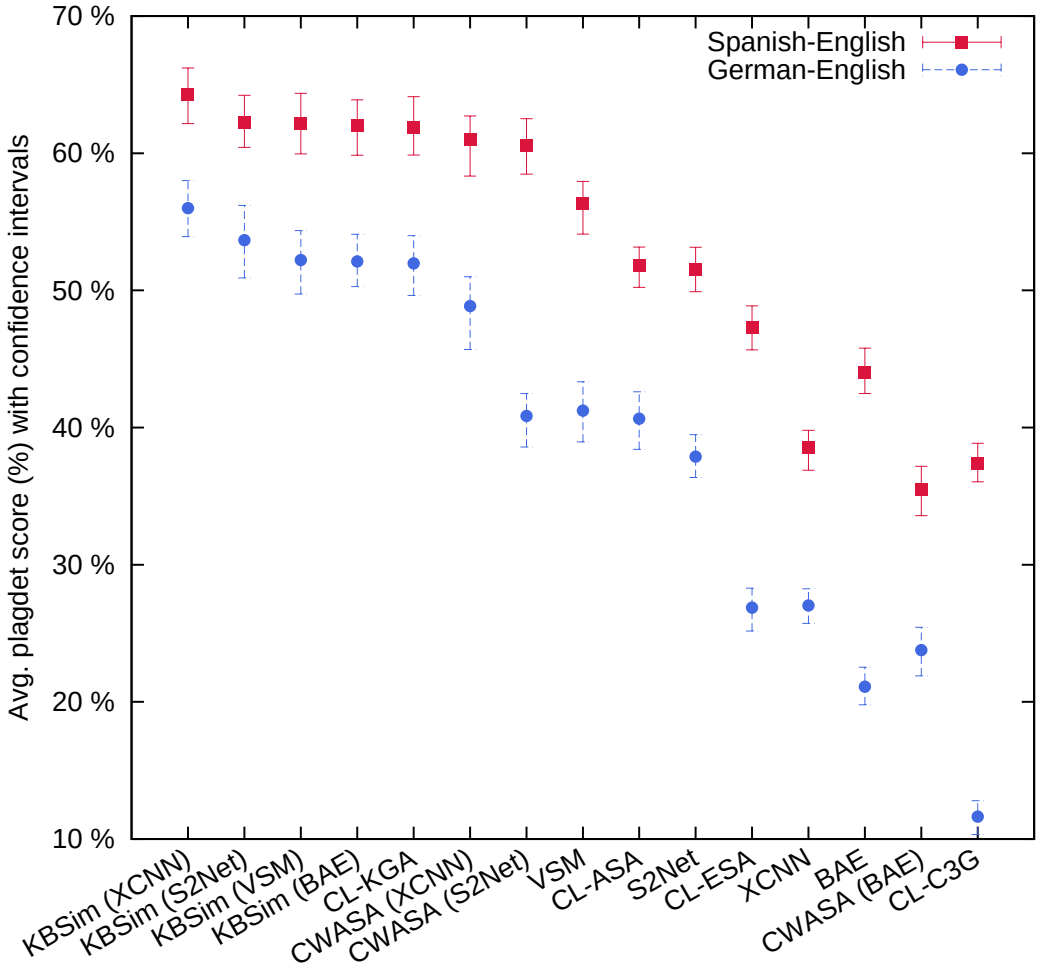

Figure 4: Plagdet score (%) of the compared models with confidence intervals for the Spanish-English and German-English partitions. Non-overlapped intervals among models represent statistically significant differences.

sure the plagdet of the models in ES-EN and DE-EN including also their confidence intervals. We show the results in Figure 4. The KBSim and CL-KGA models do not show significant improvements between them for ES-EN. Despite the average higher KBSim (XCNN) performance, these results show that CL-KGA or other KBSim models could perform at the same level or

were calculated with an α of 0.05 and 1,000 samplings.

higher. In contrast, KBSim (XCNN) and KBSim (S2Net) offer significant differences in DE-EN. However, the larger confidence intervals for DE-EN with KBSim (S2Net) denote a higher variability in performance. 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 alignments for continuous word-based similarity analysis. In addition, CWASA (XCNN) proved to be also superior to CWASA (S2Net) in DE-EN and therefore the most stable. Finally, note that KBSim (XCNN) had the shortest distance between intervals of the same model across language pairs. This 4% division manifests that the model is the most stable across languages for CL plagiarism detection.

6.4.1. Cross-language plagiarism detection in function of the type of plagiarism cases

In this last experiment we analyse the performance of the models in function of the type of plagiarism case for CL plagiarism detection. As in Section 6.3.1, we divide plagiarism cases in function of the type obfuscation employed to generate the case, and in function of the case length. We note the highlights of the models with respect to the general plagiarism detection analysis of Section 6.4.

In Table 6, attending to the obfuscation type, we note again the additional difficulty for cases with manual obfuscation. In this experiment there is an additional handicap compared to the experiment of Section 6.3.1: the detailed analysis and proprocessing Algorithm 1. In the statistics of Table 6.1 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 to more information about the size of fragments, the division of documents with slide window and the Algorithm 1). Therefore, Algorithm 1 fails detecting most of this type of cases because of its nature: it needs offset overlaps of at least two detections in the five most similar fragments. Despite this fact, we observe that KBSim has been 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 obtained the best results for automatic obfuscation cases. Since these cases are more numerous, this model obtained the overall best results in Section 6.4. We also note that the 1.0 value of granularity is normal when detecting cases with large distance between them. Hence the high occurrence in the tables.

Type of obfuscation		Model	Spanish-English Plag	Prec	Rec	Gran	German-English Plag	Prec	Rec	Gran
Translated manual obfuscation	(a)	CL-KGA	**0.139**	**0.158**	**0.124**	**1.000**	**0.169**	**0.207**	**0.143**	**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)	0.139	0.151	**0.129**	**1.000**	**0.196**	0.224	**0.174**	**1.000**
		KBSim (VSM)	**0.143**	**0.166**	0.126	**1.000**	0.176	**0.229**	0.143	**1.000**
		KBSim (BAE)	0.132	0.152	0.116	**1.000**	0.183	0.226	0.155	**1.000**
		KBSim (XCNN)	0.129	0.154	0.111	**1.000**	0.176	0.222	0.145	**1.000**
Translated automatic obfuscation	(a)	CL-KGA	**0.660**	**0.742**	**0.595**	**1.000**	**0.556**	0.642	**0.493**	**1.004**
		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)	**0.688**	**0.816**	0.594	**1.000**	**0.600**	**0.775**	0.496	1.012
		KBSim (S2Net)	0.663	0.747	**0.596**	**1.000**	0.571	0.653	**0.508**	**1.002**
		KBSim (VSM)	0.661	0.742	**0.596**	**1.000**	0.559	0.637	0.498	**1.002**
		KBSim (BAE)	0.663	0.751	0.594	**1.000**	0.556	0.630	0.500	1.005

Table 6: Spanish-English and German-English performance analysis in terms of type of obfuscation, plagdet (Plag), precision (Prec), recall (Rec) and granularity (Gran).

In Table 7 we can see the results in function of the case length. It is interesting that CL-ASA outperformed CL-KGA for ES-EN long cases. The alignment model included in CL-ASA eased the detection of long cases — mostly composed by automatic translated cases — and increased the precision. In fact, this model was originally meant for detecting verbatim plagiarism cases. In contrast, we observe that the model did not excelled for short cases of plagiarism, and was outperformed by CL-ESA. Overall, with exception of short DE-EN cases, KBSim (XCNN) obtained the best results in all the experiments. We also note its difference in performance for long cases of plagiarism compared to KBSim (S2Net). These facts manifest KBSim

Case length		Model	Spanish-English Plag	Prec	Rec	Gran	German-English Plag	Prec	Rec	Gran
Long length cases	(a)	CL-KGA	0.406	0.414	**0.398**	**1.000**	**0.366**	0.392	**0.347**	**1.006**
		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)	**0.431**	**0.467**	**0.400**	**1.000**	**0.410**	**0.499**	**0.356**	1.019
		KBSim (BAE)	0.408	0.418	**0.400**	**1.000**	0.367	0.387	0.352	1.008
		KBSim (S2Net)	0.406	0.413	**0.400**	**1.000**	0.365	0.386	0.348	**1.003**
		KBSim (VSM)	0.407	0.414	**0.400**	**1.000**	0.364	0.384	0.347	**1.003**
Medium length cases	(a)	CL-KGA	**0.224**	**0.224**	**0.225**	**1.000**	**0.211**	**0.231**	**0.193**	**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)	**0.237**	**0.254**	0.221	**1.000**	**0.225**	**0.276**	0.190	**1.000**
		KBSim (S2Net)	0.221	0.218	0.223	**1.000**	0.221	0.240	**0.205**	**1.000**
		KBSim (VSM)	0.223	0.222	**0.225**	**1.000**	0.214	0.232	0.198	**1.000**
		KBSim (BAE)	0.224	0.224	0.224	**1.000**	0.210	0.227	0.196	**1.000**
Short length cases	(a)	CL-KGA	**0.012**	**0.009**	**0.021**	**1.000**	**0.011**	**0.008**	**0.018**	**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)	**0.015**	0.010	**0.025**	**1.000**	**0.013**	**0.010**	**0.023**	**1.000**
		KBSim (XCNN)	**0.015**	**0.011**	0.022	**1.000**	0.010	0.007	0.017	**1.000**
		KBSim (BAE)	0.013	0.009	0.021	**1.000**	0.012	0.008	0.019	**1.000**
		KBSim (VSM)	0.012	0.009	0.021	**1.000**	0.011	0.008	0.018	**1.000**

Table 7: Spanish-English and German-English performance analysis in terms of plagiarism case length, plagdet (Plag), precision (Prec), recall (Rec) and granularity (Gran).

(XCNN) versatility for the task of CL plagiarism detection.

7. Conclusions

In this paper we applied multilingual continuous representations for the task of cross-language plagiarism detection. We compared existing S2Net, BAE, and XCNN models with several state-of-the-art approaches. In addition, we introduced CWASA, a new continuous word alignment-based model for tasks of similarity analysis. Finally, we studied the convergence between knowledge-based and continuous representation-based methods. We integrated the latter models in the state-of-the-art KBSim model. Thanks to this combination, KBSim (XCNN) offered state-of-the-art performance on the Spanish-English and German-English partitions of the PAN-PC-11 dataset. The study of the model in function of the type of plagiarism case — translated obfuscation, translated manual obfuscation, short, medium, and long cases — proved also its superiority. This confirms that knowledge graphs and continuous models capture different aspects of text and complement each other. This also proves the potential of KBSim for the tasks of CL similarity analysis and plagiarism detection. The comparison of the continuous representation models showed that S2Net is the best alternative when the document representation is a vector. However, without outperforming KBSim (XCNN), the use of CWASA notably increased the results of these models. CWASA (XCNN), completely designed to generate continuous representations of words, was best alternative.

For future work we will continue exploring the use of knowledge graphs, multilingual continuous representations, and how to combine them for tasks of cross-language similarity analysis. In addition, we will evaluate the performance of CWASA for monolingual similarity with other continuous word representations such as the popular continuous skip-gram and continuous bag-of-words models (Mikolov et al., 2013a,b). Finally, we are interested into explore detailed analysis and postprocessing alternatives in order to detect plagiarism more accurately.

Acknowledgements

This research has been carried out in the framework of the DIANA-APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project. We would like to thank Martin Potthast, Daniel

Ortiz-Martínez, and Luis A. Leiva for their support and comments during this research.

References

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.

Blei, D. M., Ng, A. Y., Jordan, M. I., Mar. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
URL `http://dl.acm.org/citation.cfm?id=944919.944937`

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R., 1993. Signature verification using A "siamese" time delay neural network. IJPRAI 7 (4), 669–688.
URL `http://dx.doi.org/10.1142/S0218001493000339`

Chandar A. P., S., Lauly, S., Larochelle, H., Khapra, M. M., Ravindran, B., Raykar, V. C., Saha, A., 2014. An autoencoder approach to learning bilingual word representations. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. pp. 1853–1861.

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.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A., 1990. Indexing by latent semantic analysis. JASIS 41 (6), 391–407.

Dumais, S., Landauer, T. K., Littman, M. L., 1997. Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing. In: AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval. pp. 18–24.

Efron, B., Tibshirani, R. J., 1994. An introduction to the bootstrap. CRC press.

Fellbaum, C., 1998. WordNet: An electronic lexical database. Bradford Books.

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., Rosso, P., 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. Association for Computational Linguistics, pp. 414–423.

Franco-Salvador, M., Rosso, P., y Gómez, M. M., 2016. A systematic study of knowledge graph analysis for cross-language plagiarism detection. Information Processing & Management (Accepted for publication).

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

Gupta, P., Bali, K., Banchs, R. E., Choudhury, M., 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, Gold Coast , QLD, Australia - July 06 - 11, 2014. pp. 677–686.

Gupta, P., Banchs, R. E., Rosso, P., 2015. Continuous space models for clir. Technical Report, Universitat Politècnica de València.

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.

Hassan, S., Mihalcea, R., 2011. Semantic relatedness using salient semantic analysis.

Hinton, G., Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504 – 507.

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.

Lauly, S., Boulanger, A., Larochelle, H., 2014. Learning multilingual word representations using a bag-of-words autoencoder. CoRR abs/1401.1803.

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.

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

Och, F. J., Ney, H., 2003. A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51.

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.

Platt, J. C., Toutanova, K., tau Yih, W., 2010. Translingual document representations from discriminative projections. In: EMNLP. pp. 251–261.

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., 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.

Pouliquen, B., Steinberger, R., Ignat, C., 2003. Automatic Identification of Document Translations in Large Multilingual Document Collections. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003). Borovets, Bulgaria, pp. 401–408.

Salakhutdinov, R., Hinton, G., Jul. 2009. Semantic hashing. Int. J. Approx. Reasoning 50 (7), 969–978.
URL http://dx.doi.org/10.1016/j.ijar.2008.11.006

Salton, G., Fox, E. A., Wu, H., 1983. Extended boolean information retrieval. Communications of the ACM 26 (11), 1022–1036.

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).

Yih, W., 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, CoNLL 2011, Portland, Oregon, USA, June 23-24, 2011. pp. 247–256.