Knowledge-Based Systems 111 (2016) 87–99

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Cross-language plagiarism detection over continuous-space- and knowledge graph-based 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, 138632, Singapore

ARTICLE INFO

Article history:
Received 13 January 2016
Revised 21 July 2016
Accepted 5 August 2016
Available online 6 August 2016

Keywords:
Cross-language
Plagiarism detection
Continuous representations
Knowledge graphs
Multilingual semantic network

ABSTRACT

Cross-language (CL) plagiarism detection aims at detecting plagiarised fragments of text among documents in different languages. The main research question of this work is on whether knowledge graph representations and continuous space representations can complement to each other and improve the state-of-the-art performance in CL plagiarism detection methods. In this sense, we propose and evaluate hybrid models to assess the semantic similarity of two segments of text in different languages. The proposed hybrid models combine knowledge graph representations with continuous space representations aiming at exploiting their complementarity in capturing different aspects of cross-lingual similarity. We also present the continuous word alignment-based similarity analysis, a new model to estimate similarity between text fragments. We compare the aforementioned approaches with several state-of-the-art models in the task of CL plagiarism detection and study 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 document similarity model to outperform the state-of-the-art in CL plagiarism detection.

1. Introduction

Automatic plagiarism detection refers to the task of automatically identifying which fragment of text is plagiarised. It involves finding plagiarised fragments f_q from a suspicious document d_q along with the source fragments f_s from a collection of source documents D. In the cross-language setting, suspicious and source documents are written in different languages [3,36]. This work aims at studying the combination of knowledge graph and continuous representation-based methods for the task of CL plagiarism detection.

There exist many approaches to CL plagiarism detection (CLPD) [2,13,14,19,39]. Current state-of-the-art approaches to CLPD are based on vector space model representations which operate in high dimensional spaces. Most of these approaches try to establish semantic similarity using external resources such as parallel data, comparable data, or semantic networks. The Knowledge-Based document Similarity (KBSim) model [15] combines relevance cues from knowledge graphs — generated by means of a multilingual semantic network —, and Vector Space Models (VSM) for capturing aspects of text such as out-of-vocabulary words and estimating CL similarity. However, KBSim has not yet been evaluated for CLPD. Compared with the VSM representation, latent semantic — continuous space — models have been shown to offer a higher performance when measuring text similarity [34]. The main research question of this work is on whether knowledge graph representations and continuous space representations can complement to each other and improve the state-of-the-art performance in CLPD. In this sense, we propose and evaluate hybrid models to assess the semantic similarity of two segments of text in different languages. The proposed hybrid models employ KBSim to combine knowledge graph representations with continuous space representations aiming at exploiting their complementarity in capturing different aspects of cross-lingual similarity. We analyse the quality of the KBSim model when it employs continuous models

* Corresponding author.
E-mail addresses: mfranco@prhlt.upv.es (M. Franco-Salvador), pgupta@dsic.upv.es (P. Gupta).

[1] The first two authors are ordered alphabetically and their contributions are the following: Marc Franco-Salvador implemented the state-of-the-art and designed the CL-KGA, KBSim, and CWASA models. He also carried out the evaluation of the models and contributed in the paper writing. Parth Gupta implemented S2Net, designed the BAE and XCNN continuous space models and contributed in the paper writing.

http://dx.doi.org/10.1016/j.knosys.2016.08.004
0950-7051/

such as Siamese Neural Network (S2Net) [44], Bilingual Autoencoder (BAE) [17,25], and eXternal data Composition Neural Network (XCNN) [18] for CL plagiarism detection. In addition, this study aims at filling the research gap of performance analysis of these continuous models for the CLPD task and also evaluates their independent performance. Finally, we 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 continuous word representations to measure the distance between two texts.

We carry out 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 analysis on performance of these algorithms for different lengths and types of plagiarism cases, and a study of the computational efficiency of the evaluated approaches. Our experiments show that, though continuous models have a small coverage (20k words) and have been trained on a parallel corpus of a limited size, they exhibit robust performance, especially when composed with CWASA, compared to VSM representations that have full coverage. Moreover, when combined together using KBSim, the performance is superior than the one of any other model alone. This 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 methods for CL plagiarism detection and continuous representation-based models in Sections 3 and 4, respectively. Section 5 covers the details about the KBSim model. We present our experimental framework with results and analysis in Section 6. Finally, in Section 7 we draw some conclusions.

2. Related work

The Cross-Language Character *n*-Gram (CL-CNG) model [28] follows the architecture of some monolingual models for plagiarism [7,27]. It employs vectors of character *n*-grams to represent texts and uses a measure of similarity between vectors such as the cosine similarity to compare them. This model proved to be effective for Romance and Germanic languages that share lexical and syntactic similarities.

There exist several approaches designed to measure CL similarity between distant languages. The Cross-Language Explicit Semantic Analysis (CL-ESA) [39] model adapts the well-known ESA [16] architecture to represent texts by their similarities with a multilingual collection of documents. The use of a multilingual collection such as Wikipedia, with comparable documents across languages, allows to directly compare vectors generated from distinct languages.

Models based on parallel corpora have also been proposed. This type of corpora allows to create statistical bilingual dictionaries. The Cross-Language Alignment-based Similarity Analysis (CL-ASA) model [2,4,33] uses them to translate and align words. The alignments are based on the translation probabilities and also account for the difference in length of equivalent texts in distinct languages.

The use of multilingual knowledge resources or semantic networks have been explored too. The Cross-Language Conceptual Thesaurus based Similarity (CL-CTS) model [19] uses the Eurovoc conceptual thesaurus[2] to measure the similarity between texts in terms of the number of concepts and named entities that they share. With respect to CL-CNG and CL-ASA, it provided with an average performance and excelled for Spanish-English. On the other hand, the Cross-Language Knowledge Graph Analysis (CL-KGA) model [13,14] employs a multilingual semantic network to represent the context of documents by means of knowledge graphs. This representation includes characteristics such as word sense disambiguation, concept relatedness, or vocabulary expansion. This model excels even in cases with paraphrasing and represents the state of the art in CL plagiarism detection. In recent years, several improvements over the CL-KGA architecture made the Knowledge-Based document Similarity (KBSim) model [15] an interesting alternative for CL document retrieval and categorisation. This model complements knowledge graphs with a vector component to cover knowledge graph shortcomings such as out-of-vocabulary words and verbal tenses. However, KBSim has not yet been evaluated for CL plagiarism detection neither combined with continuous representations.

Recently, the Conference and Labs of the Evaluation Forum (CLEF) actively covered the plagiarism detection task in the framework of the evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN).[3] The shared task in plagiarism detection [38] provides with a corpus of plagiarism and allows participants to create systems in order to compete at detecting its plagiarism cases. The datasets of the 2010 and 2011 editions [35,37] included also German-English and Spanish-English CL partitions. The most popular methods to detect CL plagiarism at PAN followed [8] and used machine translation techniques to convert the problem into a monolingual one. However, this puts forward a heavy dependence on availability of Machine Translation (MT) systems in the involved languages and their quality. In addition, we believe that the nature of those methods is not purely cross-lingual and can be considered monolingual with a MT preprocessing. Therefore, all the models employed in this work for CL plagiarism detection do not depend on full MT systems. Nevertheless, in [3] authors show a comparison of the CL-ASA and CL-CNG models with an approach (T+MA) employing MT to analyse the similarities at monolingual level. That study shows that T+MA is superior in short cases of plagiarism but similar to CL-ASA, that always obtains a higher precision and better results for long cases of plagiarism. Considering that [3] included an evaluation setting very similar to ours (see Section 6), we decided to not include T+MA again in this work.

In [14] 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 [36]. Different performances were observed depending on the languages, and the 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 performance for language pairs without lexical and syntactic similarities. Therefore, for the sake of completeness, in this work we decided to compare our KBSim model based on continuous representations against the CL-CNG, CL-ESA, CL-ASA, and CL-KGA models.

With respect to the continuous space representations of texts, often referred to as embeddings, the advancement in the area has been quite limited. The high dimensional representation of text in vector space is projected into a low dimensional space

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

[3] pan.webis.de/.

by means of dimensionality reduction techniques. Such representation is dense and continuous in nature and often referred to as continuous space representation of text or text embeddings. If such representations are at word-level, they are referred to as continuous word representations or word embeddings. 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 [5,9,22,30]. 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-lingual extension of Latent Semantic Indexing (LSI) [10]. Oriented Principle Component Analysis (OPCA) tries to learn a trans-lingual projection matrix by solving a generalised eigen value problem [34]. Similarly, Siamese neural network based S2Net learns the same projection matrix through backpropagation error of distance between parallel sentence pairs [44]. There also exist non-linear deep neural network based solutions to learn such cross-lingual embeddings through deep autoencoders [17,25,26] and composition neural networks [18]. In this study, we analyse the performance of some of these models separately and also when integrated in the KBSim model. More details about them are given in Section 4.

3. Methods for cross-language plagiarism detection

In this section we describe more in detail the cross-language plagiarism detection process. This task is usually performed in two steps: i) candidate retrieval and ii) detailed analysis and postprocessing [36]. The candidate retrieval provides with the list of possible text fragments involved in plagiarism. In order to detect the candidates between two documents d_L and $d'_{L'}$ written in languages L and L', the documents are first segmented to obtain the sets of fragments $\text{FC} \in d_L$ and $\text{FC}' \in d'_{L'}$. Next, a model is used to obtain the set of cross-language similarities $\text{SF} = \{\text{S}(F, F')\}$ between all the pairs of text fragments (F, F'), $F \in \text{FC}$ and $F' \in \text{FC}'$. Once the set SF is obtained, the detailed analysis and postprocessing is employed to analyse the values and determine which fragments of text are plagiarism cases. We employ the method introduced in [2,3] that is described in Algorithm 1 . In this work, we used the original thresholds tuned in [2,3] for the PAN-PC-10 and PAN-PC-11 datasets: $\text{thres}_1 = 1500$ and $\text{thres}_2 = 2$. Specifically, the paragraph size of 5 sentences and the step size of 2 sentences are aimed at achieving a small size of overlapping paragraphs between texts of considerable length. These values can be tuned for different corpora as explained in [2,3]. 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 a VSM approach, which is later employed in Section 5 as part of the KBSim model. Following we give details of the CL-KGA and VSM models, whose descriptions help later to describe KBSim easily. For more details about CL-CNG, CL-ASA, and CL-ESA, please refer to their original works.

3.1. Cross-language knowledge graph analysis

Cross-language Knowledge Graph Analysis (CL-KGA) [13,14] 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 [31], focused on the concepts belonging to a text and the semantic relations between them. As stated in [14], these graphs have several interesting characteristics such as Word Sense Disambiguation (WSD), vocabulary expansion, and concept language independence. Note that concepts are represented in BabelNet by means of multilingual sets of synonyms. 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 the Dice's coefficient [24]:

$$\text{S}_c(G, G') = \frac{2 \cdot \sum_{c \in \text{V}(G) \cap \text{V}(G')} \text{w}(c)}{\sum_{c \in \text{V}(G)} \text{w}(c) + \sum_{c \in \text{V}(G')} \text{w}(c)}, \quad (1)$$

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

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

where $\text{E}(G)$ is the set of relations (edges) in the graph and $\text{w}(r)$ 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 $S_g(G, G')$ between knowledge graphs:

$$\text{S}_g(G, G') = \alpha \cdot \text{S}_c(G, G') + (1 - \alpha) \cdot \text{S}_r(G, G'), \quad (3)$$

Algorithm 1: Detailed analysis and postprocessing.

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

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 of 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
```

where α determines the relevance of the conceptual and the relational similarities.[4]

In this work concepts are weighted using their graph outdegree [31]. In contrast, relations are weighted using the original weights between relations provided in BabelNet. These weights where calculated using an extension of the extended gloss overlap measure [1] which weights semantic relations between WordNet [12] and Wikipedia concepts.[5] For more details about the CL-KGA model, please refer to its original works [13,14].

3.2. Vector space model

The Vector Space Model (VSM) represents documents using the TF-IDF weighting scheme and compares them using the cosine similarity. Often the text is normalised by means of tokenisation, stop-word and punctuation removal. In order to make this approach cross-lingual and to counterbalance possible translation errors, we followed [15], where each document d_L is represented 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[6] with TF-IDF re-weighting as function of the probabilities of translation of the words.

4. Continuous representations for cross-language plagiarism detection

This section presents details of the continuous representation 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 receive as input. Most of these models learn cross-lingual embeddings using parallel or comparable corpus. For a fair comparison, in all the experiments presented in this work, these models are trained using the same parallel corpus. We used 250k English-Spanish and English-German parallel sentences from DGT-Translation Memory distributed by JRC.[7] For monolingual pre-initialisation in XCNN (Section 4.3) we used CLEF ad-hoc retrieval corpus document titles.

4.1. Similarity learning via siamese neural network

Following the general Siamese neural network architecture [6], Similarity Learning via Siamese Neural Network (S2Net) trains two identical neural networks concurrently. The S2Net receives as input parallel data with binary or real-valued similarity score and updates the model parameters accordingly [44]. It optimises a dynamic objective function which is directly modelled by using the cosine similarity. The projection operation can be described as follows:

$$y_d = W * x_d, \tag{4}$$

where, x_d is the input term vector for the document d, W is the learnt projection matrix (represented by the model parameters) and y_d is the latent representation of document d. The parameters of the S2Net are tuned accordingly to the details provided in [44].

[4] In this work we used the optimal values provided in [14] for concepts and relations: $a = b = 0.5$.

[5] A new weighting scheme for relations based on continuous representations of concepts was introduced in [14], but their knowledge graph construction was penalised in terms of computational time. Therefore, in this work we decided to use the classical weighting scheme in order to speed up the process.

[6] We used the same dictionary employed in CL-ASA.

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

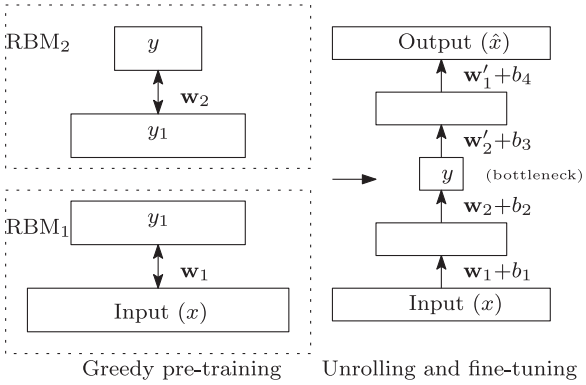

Fig. 1. Left panel: pre-training of stacked RBMs where the upper RBM takes as input the output of the lower RBM. Right panel: After pre-training the structure is "unrolled" to create a multi-layer network which is fine-tuned by means of back-propagation to learn an identity function $\hat{x} \approx x$.

4.2. Bilingual autoencoder

Salakhutdinov and Hinton [41] demonstrated that semantic modelling by means of dimensionality reduction through deep autoencoders lead to superior performance compared to the conventional LSA approach. Deep autoencoders were extended to model cross-language data and are referred to as Bilingual Autoencoders (BAE) [17,25,26]. 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. 1. For more details on training, the reader may refer to [17].

The representation through RBM is obtained as shown below:

$$\begin{aligned} y_1 &= \sigma(\mathbf{w}_1 * x + b_1) \\ y &= \sigma(\mathbf{w}_2 * y_1 + b_2) \end{aligned} \tag{5}$$

where, $\mathbf{w}_i$ and b_i are weight and bias parameters of the ith RBM, and σ is the logistic function which provides non-linearity. The output of the bottom layer is supplied as input to the layer above in order to create a stacked structure as shown in Fig. 1.

4.3. External-data composition neural networks

External-data Composition Neural Network (XCNN) is based on a composition function that is implemented on top of a deep neural network that provides a distributed learning framework [18]. Different from many other models including S2Net and BAE, which solely rely on parallel/comparable data for training, XCNN exploits also monolingual data for model training purposes. Specifically, it incorporates external relevance signals such as pseudo-relevance data or clickthrough data into the learning framework. The main motivation behind this strategy is that, monolingual models can be initialised from such largely available relevance data and then, with the help of a smaller amount of parallel data, the cross-lingual model can be trained. This property helps to gain more confidence for under-represented terms in parallel data, i.e. terms with very low frequency.

The architecture of XCNN model training is shown in Fig. 2. XCNN learns word embeddings in cross-lingual setting using the objective function defined in Eq. 6. It maximises the cosine similarity φ for a training example for a positive sample and miminises it for a negative sample. The network parameters are

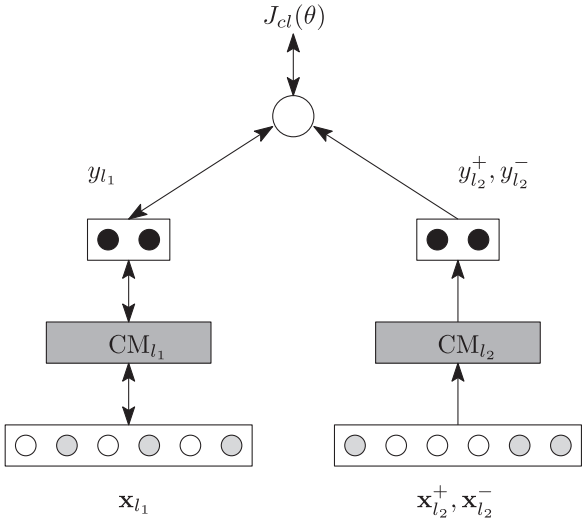

Fig. 2. Architecture of external-data composition neural network model for cross-lingual training.

updated through backpropagation.

$$J_{cl}(\theta) = \varphi(y_{l_1}, y^+_{l_2}) - \varphi(y_{l_1}, y^-_{l_2}) \tag{6}$$

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

$$\begin{aligned} y_i^{(l_1)} &= g(W_1 * x_i + b_1) \\ y_i^{(l_j)} &= g(W_j * y_i^{l_{(j-1)}} + b_j),\ j = 2, \ldots, m \\ y &= \sum_{i=1}^{n} y_i^{(l_m)} \end{aligned} \tag{7}$$

where $y_i^{(lj)}$ represents the ith term x_i in text in the layer j of a neural network, l_m represents the output layer. More details about XCNN can be found in [18].

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.[8] The Continuous Word Alignment-based Similarity Analysis (CWASA) model modifies the text-to-text relatedness proposed by Hassan and Mihalcea [20] in order to estimate the similarity between text fragments or 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 the closest to *First_Lady* in both directions. Formally, the similarity $S(d, d')$ between two documents d and d' is estimated as follows:

$$S(d, d') = \frac{1}{|\Phi|} \sum_{c_k \in \Phi} c_k, \tag{8}$$

where $d = (x_1, \ldots, x_n)$ and $d' = (y_1, \ldots, y_m)$ are represented as lists of continuous words, and Φ is generated from the list $\Phi' = \{c'_1, \ldots, c'_{n+m}\}$ that satisfies Eq. 9:

$$c'_k = \begin{cases} \arg \max_{i=k, x_i \in d, y_j \in d'} \varphi(x_i, y_j), & \text{if } k \le n \\ \arg \max_{j=k-n, x_i \in d, y_j \in d'} \varphi(x_i, y_j), & \text{otherwise} \end{cases} \tag{9}$$

where $1 \le i \le n$, $1 \le j \le m$, $1 \le k \le n+m$, φ is the cosine similarity function, and being $\Phi = \{c_1, \ldots, c_z \mid \max(n, m) \le z \le n+m\}$, $\Phi \subseteq \Phi'$, the set of cosine similarities without pairing repetitions[9] that represents the strongest semantic pairing between the continuous words of documents d and d'.

Basically, in Eq. 9 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. 8 to estimate the similarity score between d and d' as the average of the different alignments. We note that this problem has been efficiently solved by dynamic programming.[10] In addition, although this work is focused on a cross-lingual setting, CWASA could be directly employed with monolingual continuous word representations [29,30]. 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 these 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 document Similarity (KBSim) model [15] extends CL-KGA 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:

$$S(d, d') = c(G) S_g(G, G') + (1 - c(G)) S_v(\vec{v}, \vec{v}'), \tag{10}$$

where $S_g(G, G')$ is the knowledge graph similarity of Eq. 3, $S_v(\vec{v}, \vec{v}')$ is the vector-based similarity, and $c(G)$ is an interpolation factor calculated as the edge density of knowledge graph G:

$$c(G) = \frac{|E(G)|}{|V(G)|(|V(G)| - 1)} \tag{11}$$

Note that, by using the factor $c(G)$ to interpolate the two similarities in Eq. 10, the relevance for the knowledge graphs and the multilingual vectors is determined in a dynamic way. Indeed, $c(G)$ weights the contribution of graph similarity depending on the richness of the knowledge graph.

[8] A continuous word alignment establishes a relationship between two continuous word representations.

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

[10] We pre-computed the cosine similarities between words and used a loop to detect the closest word of d' for each one in d. The same loop served to detect the closest words in the opposite direction employing an auxiliary vector. Therefore, at similarity computing time, the cost of the model is $O(n \cdot m)$.

Table 1
Statistics of PAN-PC-11 cross-language plagiarism detection partitions.

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	1506	– Translated automatic obfuscation	5142
– Medium length cases	2118	– Translated manual obfuscation	433
– Short length cases	1951		

The vector-based similarity $S_v(\vec{v}, \vec{v}')$ was originally calculated with the VSM introduced in Section 3.2. However, in this work we are also comparing continuous space representations. As consequence, we are also interested in analysing 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 their analysis.

6.1. Datasets

To evaluate the models we selected the PAN-PC-11[11] dataset that was created for the 2011 CL plagiarism detection competition of PAN at CLEF.[12] The dataset consists 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.[13] In addition, PAN-PC-11 contains also cases of plagiarism with manual obfuscation after automatic translation. These cases are CL paraphrasing cases of plagiarism. Amazon Mechanical Turk was employed to generate them. In Table 1 we present the statistics of the dataset. More details on the dataset can be found in [37].

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 described in Algorithm 1. Recall is measured using the top k (R@k) most similar fragments of text, where $k = \{1, 5, 10, 20\}$. However, in order to increase precision, we conduct a second experiment in Section 6.4. There, detections are filtered using Algorithm 1 to determine what cases are plagiarism. As evaluation metric of the experiment we selected the measures employed in the PAN shared task: precision, recall, granularity, and plagdet [40]. Let S denote the set of plagiarism cases in the suspicious documents, and let R denote the set of plagiarism detections that the detector reports for these documents. A plagiarism case $s \in S$ is represented with a reference to the characters that forms it, i.e., its offsets. Likewise, $r \in R$ represents a plagiarism detection. Based on these representations, the precision and the recall at character level of R under S are measured as follows:

$$\text{precision}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\bigcup_{s \in S}(s \sqcap r)|}{|r|}; \tag{12}$$

$$\text{recall}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\bigcup_{r \in R}(s \sqcap r)|}{|s|}, \tag{13}$$

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:

$$\text{granularity}(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|, \tag{14}$$

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):

$$\text{plagdet}(S, R) = \frac{F_1(S, R)}{\log_2(1 + \text{granularity}(S, R))}. \tag{15}$$

In both experiments of Sections 6.3 and 6.4 we also included in a separated subsection the analysis of results as function of the type of obfuscation and document length of the plagiarism cases. Finally, in Section 6.5 we compared the computational efficiency of the models.

We compare the continuous word representation models — S2Net, BAE, and XCNN (cf. Section 4) — with the state-of-the-art CL-C3G, CL-ESA, CL-ASA, VSM, and CL-KGA models (cf. Section 3). We note the following details about these models. CL-C3G is CL-CNG using character 3-grams, as recommended in [36]. For CL-ESA, we used 10,000 Spanish-German-English comparable Wikipedia pages as document collection. All pages contain more than 10,000 characters in length and were represented using the 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. For CL-ASA, we used a statistical dictionary trained using the word-alignment model IBM M1 [32] on the JRC-Acquis corpus [43] along with the length model tuned as per the instructions in [3]. Details about the tuning of the parameters of CL-KGA and CL-ESA are provided in [14]. The continuous models are trained with DGT translation memory parallel data with EN-ES and EN-DE pairs.[14] For XCNN monolingual preinitialisation, we used the document collection from CLEF adhoc retrieval track. The positive sample of a sentence in XCNN is the corresponding translation. Following the literature [21,23,42], the negative sample is randomly chosen from all the sentences.

We also 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). In addition, we show the performance of the original KBSim 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.

[11] http://www.uni-weimar.de/en/media/chairs/webis/corpora/corpus-pan-pc-11/.
[12] http://www.clef-initiative.eu/.
[13] https://translate.google.com/.
[14] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory.

Table 2
ES-EN and DE-EN performance analysis in terms of R@k, where $k = \{1, 5, 10, 20\}$.

	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)	XCNN	0.468	0.648	0.721	0.786	0.362	0.561	0.647	0.728
	S2Net	**0.637**	**0.763**	**0.809**	**0.852**	**0.508**	**0.675**	**0.744**	**0.799**
	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 (VSM)	**0.927**	**0.955**	**0.961**	**0.965**	0.794	0.871	0.896	0.915
	KBSim (XCNN)	0.858	0.907	0.924	0.935	0.741	0.843	0.872	0.897
	KBSim (S2Net)	0.920	0.949	0.956	0.961	**0.809**	**0.878**	**0.901**	**0.921**
	KBSim (BAE)	0.917	0.945	0.956	0.962	0.791	0.870	0.893	0.911

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 on the basis of the type of plagiarism case. In Table 2 we show the results for ES-EN and DE-EN with best results for each category in boldface. As we can see, DE-EN similarity has been more difficult to detect for all the models. However, no differences are found between the models in ES-EN and DE-EN with respect to the ranking order. Therefore, we can jointly analyse the differences between them without stating the language pair. 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 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. 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 (∼1000 words) the summed embeddings flattens vectors and lose 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 in CWASA without this sum-based composition.

The use of word alignments, i.e., by means of CWASA, produced notable improvements with 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. A higher variety of stemmed words was observed for the German agglutinative language, which has not been covered by the vocabulary in the same amount than the other languages. We also note that the performance of BAE shows the highest variation from R@1 to R@5 among all models: ∼ 21%. After a 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 on average the highest results in this experiment. Although KBSim (XCNN) does 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 algorithm employed. In Section 6.4 we will also study the statistical differences of all the models to analyse if the observed differences are significant from the statistical point of view. 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.

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 as function of the type of plagiarism case. We divide plagiarism cases according to the type obfuscation — translated obfuscation and translated manual obfuscation — employed to generate the case, and according to the case length — short, medium, and long.[15] Most of the highlights of Section 6.3 persist when discriminating considering the type of case. However, there are several points to note. In Table 3 results are reported on the basis of the obfuscation type. The translated manual obfuscation has manual correction after the automatic translation and generates cases with paraphrasing in order to hide the plagiarism. Therefore, it has been more difficult to detect similarity between such type of cases. CL-ESA, that is based

[15] We followed the PAN-PC-11 setup and considered as short cases those with less than 700 characters. Long cases are those larger than 5000 characters.

Table 3
ES-EN and DE-EN performance analysis in terms of type of obfuscation for the plagiarism cases and R@k, where k = {1, 5, 10, 20}.

Type of obfuscation		Model	Spanish-English				German-English			
			R@1	R@5	R@10	R@20	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-ASA	0.533	0.662	0.712	0.756	0.387	0.569	0.643	0.713
		CL-ESA	0.607	0.737	0.795	0.837	0.406	0.548	0.614	0.686
		CL-C3G	0.450	0.599	0.674	0.738	0.231	0.420	0.537	0.642
	(b)	XCNN	0.414	0.610	0.669	0.744	0.358	0.572	0.653	0.743
		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
	(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 (VSM)	**0.867**	**0.924**	**0.935**	**0.942**	0.714	0.800	0.853	0.870
		KBSim (XCNN)	0.764	0.840	0.874	0.893	0.641	0.774	0.830	0.873
		KBSim (S2Net)	0.853	0.912	0.925	0.940	**0.729**	**0.806**	**0.839**	**0.875**
		KBSim (BAE)	0.847	0.901	0.925	0.939	0.715	0.799	0.839	0.863
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)	XCNN	0.472	0.651	0.725	0.789	0.363	0.559	0.646	0.727
		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
	(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 (VSM)	**0.932**	**0.957**	**0.963**	**0.966**	0.802	0.879	0.900	0.920
		KBSim (XCNN)	0.865	0.912	0.928	0.938	0.751	0.850	0.877	0.899
		KBSim (S2Net)	0.925	0.952	0.958	0.962	**0.817**	**0.885**	**0.908**	**0.925**
		KBSim (BAE)	0.923	0.948	0.958	0.964	0.799	0.877	0.899	0.916

Note: ES-EN and DE-EN with best results for each category in boldface.

on a representation by similarities with a collection of documents, outperformed CL-ASA in cases with manual obfuscation. This was somehow expected due 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 which is consistent with previous studies [3]. Compared to the rest, the CL-ASA model suffered a higher decay when cases became shorter. This may be produced by the document length component of the model, that is more precise normalising larger cases of plagiarism. Note that KBSim (S2Net) obtained the highest results independently of the type of obfuscation and case length analysed, which highlights its robustness for CL similarity analysis in 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 on the PAN-PC-11 dataset for CL plagiarism detection. We show the results in Table 5. Although both English and German are Germanic languages, due to their grammatical differences, the additional difficulty of the detection in DE-EN is also visible in this experiment. The decay of plagdet — the overall score for plagiarism detection — ranges between 8%–27% when comparing DE-EN with ES-EN results. The lowest results were obtained with CL-C3G, that did not find 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 a single case 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 (especially 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 in this task. The S2Net model outperformed BAE and XCNN but obtained lower values than CL-KGA. We can see close values in terms of precision for S2Net and XCNN. However, S2Net's recall has been 10% higher in all the tests. This, along with the highest granularity, penalised XCNN's plagdet.

The models of group (c) — where CWASA was used to measure similarity — notably improved the performance of S2Net, BAE, and XCNN. We appreciate how, especially with XCNN, the recall and granularity improved with a low impact on the precision. In contrast to S2Net and BAE, that used a bag-of-words format to learn vectors of documents, XCNN directly generated continuous vectors of words. These vectors found in CWASA an excellent complement in order to accurately measure the CL similarity. Note that in this experiment we used Algorithm 1 to analyse the similarities and to identify the plagiarism cases. 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 the text fragments.

Finally, the combination of vector representations with knowledge graphs, made the KBSim models of group (d) to obtain on overall the highest results. In fact, KBSim (XCNN) outperformed the original KBSim (VSM), and was the best model, independently of the language pair analysed. This proves the potential of KBSim for the tasks of CL similarity analysis and plagiarism detection. This

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

Case length	Model		Spanish-English				German-English			
			R@1	R@5	R@10	R@20	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)	XCNN	0.486	0.663	0.735	0.800	0.351	0.545	0.634	0.717
	(b)	S2Net	**0.662**	**0.785**	**0.830**	**0.867**	**0.523**	**0.688**	**0.757**	**0.812**
		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 (VSM)	**0.944**	**0.963**	**0.967**	**0.969**	0.817	0.890	0.909	0.928
		KBSim (XCNN)	0.888	0.927	0.939	0.947	0.767	0.857	0.883	0.906
		KBSim (S2Net)	0.938	0.958	0.962	0.967	**0.831**	**0.896**	**0.917**	**0.932**
		KBSim (BAE)	0.937	0.956	0.964	0.968	0.812	0.888	0.907	0.924
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)	XCNN	0.476	0.656	0.727	0.794	0.365	0.563	0.648	0.728
		S2Net	**0.647**	**0.771**	**0.815**	**0.856**	**0.516**	**0.681**	**0.749**	**0.802**
		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 (VSM)	**0.931**	**0.957**	**0.962**	**0.965**	0.801	0.876	0.898	0.917
		KBSim (XCNN)	0.868	0.914	0.929	0.939	0.749	0.848	0.876	0.900
		KBSim (S2Net)	0.924	0.951	0.957	0.961	**0.816**	**0.884**	**0.906**	**0.924**
		KBSim (BAE)	0.921	0.948	0.957	0.963	0.799	0.874	0.896	0.913
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)	XCNN	0.463	0.644	0.716	0.782	0.361	0.559	0.646	0.728
		S2Net	**0.633**	**0.758**	**0.806**	**0.848**	**0.501**	**0.668**	**0.738**	**0.793**
		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 (VSM)	**0.924**	**0.953**	**0.959**	**0.963**	0.787	0.866	0.891	0.911
		KBSim (S2Net)	0.917	0.947	0.954	0.959	**0.802**	**0.873**	**0.897**	**0.917**
		KBSim (XCNN)	0.853	0.903	0.921	0.933	0.735	0.838	0.868	0.893
		KBSim (BAE)	0.914	0.943	0.954	0.961	0.785	0.865	0.889	0.907

Note: ES-EN and DE-EN with best results for each category in boldface.

Table 5
ES-EN and DE-EN performance analysis in terms of plagdet (Plag), precision (Prec), recall (Rec) and granularity (Gran). The results of the CL-ASA, CL-ESA, and CL-KGA models are from [14].

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)	XCNN	0.386	**0.738**	0.310	1.189	0.270	0.664	0.196	1.174
	S2Net	**0.514**	0.734	**0.440**	**1.098**	**0.379**	**0.669**	**0.304**	1.148
	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 (VSM)	0.621	0.697	0.559	**1.000**	0.523	0.599	0.465	**1.002**
	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 (BAE)	0.622	0.704	0.557	**1.000**	0.521	0.592	0.468	1.004

Note: ES-EN and DE-EN with best results for each category in boldface.

Table 6
Example of the type of cases detected by CL-KGA and XCNN. In this table, the cases detected by CL-KGA are not detected by XCNN and vice versa. The bold words highlight semantically related ones in the case of CL-KGA and frequent ones in the case of XCNN.

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

also confirms that knowledge graphs and continuous models capture different aspects of text and complement each other. In order to illustrate this fact, we selected from the English partition four cases of plagiarism generated with translated automatic obfuscation.[16] Two cases (referred as CL-KGA) were detected by CL-KGA and not by XCNN. The other two cases (referred as XCNN) represent the opposite situation. We can see them in Table 6. Thanks to the wide coverage of the BabelNet multilingual semantic network, our knowledge graph-based model ease the detection of cases with semantically related words. On the other hand, the XCNN model based on continuous representations covers knowledge graph shortcomings such as the out-of-vocabulary words and has the potential to take into account also their frequencies. We note that in this work we stemmed the input of the continuous representation models.

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 potentialities and limitations of Algorithm 1 for plagiarism detection.

After analysing the performance of the models, we are also interested in analysing whether or not the observed differences across the obtained results are statistically significant. In order to analyse this, we used bootstrap resampling[17] [11] to measure the plagdet of the models in ES-EN and DE-EN including also their confidence intervals. We show the results in Fig. 3. The KBSim and CL-KGA models do not show significant differences for ES-EN. Despite KBSim (XCNN) obtains on average a higher performance, these results show that CL-KGA or other KBSim models perform similarly. In contrast, KBSim (XCNN) and KBSim (S2Net) show significant differences for 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 its 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% of division suggests that the model is the most stable across languages for CL plagiarism detection.

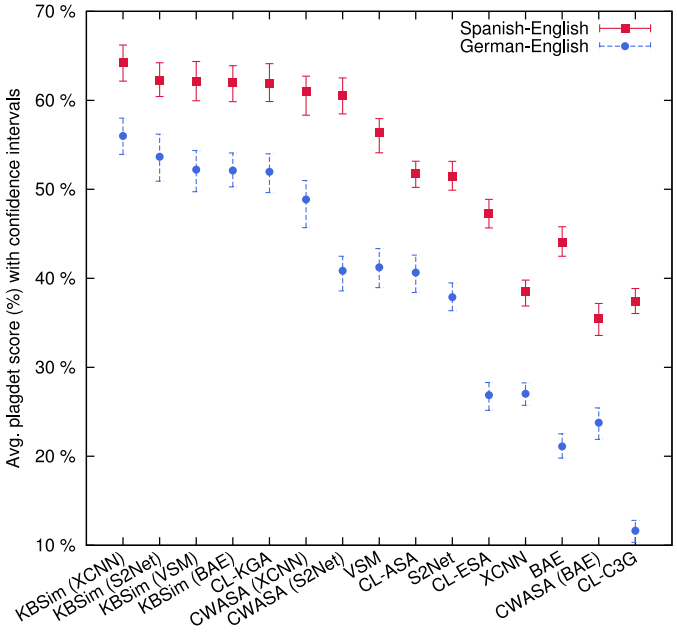

Fig. 3. Plagdet score (%) of the compared models with confidence intervals for the Spanish-English and German-English partitions. Non-overlapped intervals among models represent statistically significant differences.

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 considering the type of plagiarism case for CL plagiarism detection. As in Section 6.3.1, we divide the plagiarism cases in function of the type obfuscation employed to generate the case, and on the basis of the case length. We note the most relevant differences among the models with respect to the general plagiarism detection analysis of Section 6.4.

In Table 7, depending on 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 of Algorithm 1. In the statistics of Table 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

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

[17] Bootstrap methods obtained generally better results in parametric tests for small datasets — as the dataset in hand — or where sample distributions are non-normal. The statistical tests were calculated with an α of 0.05 and 1000 samplings.

Table 7
ES-EN and DE-EN performance analysis in terms of type of obfuscation, plagdet (Plag), precision (Prec), recall (Rec) and granularity (Gran).

Type of obfuscation	Model		Spanish-English				German-English			
			Plag	Prec	Rec	Gran	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)	XCNN	0.077	0.116	0.058	**1.000**	0.085	0.160	0.058	**1.000**
		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**
	(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 (VSM)	**0.143**	**0.166**	0.126	**1.000**	0.176	**0.229**	0.143	**1.000**
		KBSim (XCNN)	0.129	0.154	0.111	**1.000**	0.176	0.222	0.145	**1.000**
		KBSim (S2Net)	0.139	0.151	**0.129**	**1.000**	**0.196**	0.224	**0.174**	**1.000**
		KBSim (BAE)	0.132	0.152	0.116	**1.000**	0.183	0.226	0.155	**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)	XCNN	0.412	**0.791**	0.331	1.205	0.289	0.715	0.210	1.191
		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**
	(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 (VSM)	0.661	0.742	**0.596**	**1.000**	0.559	0.637	0.498	**1.002**
		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 (BAE)	0.663	0.751	0.594	**1.000**	0.556	0.630	0.500	1.005

Note: ES-EN and DE-EN with best results for each category in boldface.

by a single text fragment (see Section 3 to more information about the size of fragments, the division of documents with slide window and Algorithm 1). Therefore, Algorithm 1 fails to detect most of this type of cases: it needs offset overlaps of at least two detections in the five most similar fragments. Despite this fact, we observe that KBSim 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 in the document. Hence the high occurrence in the tables.

In Table 8 we can see the results depending on 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 excel 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 the versatility of the KBSim (XCNN) model for the task of CL plagiarism detection.

6.5. Study of the computational efficiency

The computational efficiency of the similarity models is a key aspect in order to decide which one choose for a cross-language plagiarism detection system. This decision affects to the accuracy and speed of the system and usually is based on its requirements and purpose. In Table 9 we show the time necessary to transform the texts to the space of the models (indexing), and the time for measuring the similarity once that transformation is done. We used an Intel-i5@2.8 Ghz with 16GB of RAM to perform these tests over the complete ES-EN partition. As we can see, there was more variability in the indexing time. The CL-KGA and KBSim knowledge graph-based models were slow due to the time required to search paths in the BabelNet multilingual semantic network. Note that it contains more than 9 million of concepts and more than 262 million of relations among them. More simple approaches such as VSM are recommended for fast document indexing. However, text indexing is usually part of the preprocessing step, being the indexing of the new documents needed only once. The XCNN, S2Net, and BAE models offered an acceptable indexing time and excelled at text similarity level. Thanks to the cosine similarity between low-dimensional vectors, these three approaches were the fastest ones. Finally, the efficiency of the CWASA model made its similarity calculation also fast. This, together with its good performance in the experiments of this work, highlights its potential for large scale systems.

7. Conclusions

In this paper, we studied hybrid models that combine knowledge graph and continuous representation methods for the task of cross-language plagiarism detection. We integrated the existing S2Net, BAE and XCNN models in the state-of-the-art KBSim model. In addition, we separately compared these multilingual continuous representations models with several state-of-the-art approaches. Finally, we introduced CWASA, a new continuous word alignment-based model for tasks of similarity analysis. Thanks to its combination, KBSim (XCNN) offered state-of-the-art performance and

Table 8
ES-EN and DE-EN performance analysis in terms of plagiarism case length, plagdet (Plag), precision (Prec), recall (Rec) and granularity (Gran).

Case length		Model	Spanish-English				German-English			
			Plag	Prec	Rec	Gran	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**
		VSM	0.399	0.416	0.391	1.016	0.320	0.386	0.300	1.077
		CL-ASA	**0.411**	**0.535**	0.375	1.106	0.339	**0.513**	0.299	1.168
		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)	XCNN	0.327	**0.655**	0.271	1.253	0.230	**0.619**	0.170	1.234
		S2Net	**0.411**	0.587	**0.368**	**1.145**	**0.322**	0.589	**0.269**	**1.212**
		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 (VSM)	0.407	0.414	**0.400**	**1.000**	0.364	0.384	0.347	**1.003**
		KBSim (XCNN)	**0.431**	**0.467**	**0.400**	**1.000**	**0.410**	**0.499**	**0.356**	1.019
		KBSim (S2Net)	0.406	0.413	**0.400**	**1.000**	0.365	0.386	0.348	**1.003**
		KBSim (BAE)	0.408	0.418	**0.400**	**1.000**	0.367	0.387	0.352	1.008
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)	XCNN	0.127	0.221	0.089	**1.000**	0.096	0.204	0.063	**1.000**
		S2Net	**0.176**	0.240	**0.139**	**1.000**	**0.135**	**0.217**	**0.098**	**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 (VSM)	0.223	0.222	**0.225**	**1.000**	0.214	0.232	0.198	**1.000**
		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 (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-ASA	0.006	0.005	0.009	**1.000**	0.006	0.005	0.009	**1.000**
		CL-ESA	0.009	0.006	0.015	**1.000**	0.005	0.003	0.008	**1.000**
		CL-C3G	0.005	0.004	0.006	**1.000**	0.004	0.003	0.005	**1.000**
	(b)	XCNN	0.006	0.006	0.006	**1.000**	**0.009**	**0.009**	0.009	**1.000**
		S2Net	**0.008**	**0.007**	**0.010**	**1.000**	0.008	0.006	**0.010**	**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 (VSM)	0.012	0.009	0.021	**1.000**	0.011	0.008	0.018	**1.000**
		KBSim (XCNN)	**0.015**	**0.011**	0.022	**1.000**	0.010	0.007	0.017	**1.000**
		KBSim (S2Net)	**0.015**	0.010	**0.025**	**1.000**	**0.013**	**0.010**	**0.023**	**1.000**
		KBSim (BAE)	0.013	0.009	0.021	**1.000**	0.012	0.008	0.019	**1.000**

Note: ES-EN and DE-EN with best results for each category in boldface.

Table 9
Comparison of time required to index and estimate similarity between texts. Results are estimated as the average for processing all the ES-EN partition. The results of the CL-ASA, CL-ESA, and CL-KGA models are from [14].

System		Text indexing (texts/second)	Text similarity (texts/second)
(a)	CL-KGA	11	1259
	VSM	2083	2291
	CL-ASA	1741	3627
	CL-ESA	282	1826
	CL-C3G	3127	2619
(b)	XCNN	390	8599
	S2Net	433	8599
	BAE	380	8598
(c)	CWASA (XCNN)	497	3824
	CWASA (S2Net)	500	3812
	CWASA (BAE)	510	3784
(d)	KBSim (VSM)	11	1286
	KBSim (XCNN)	11	1279
	KBSim (S2Net)	11	1283
	KBSim (BAE)	11	1278

the best stability on the Spanish-English and German-English partitions of the PAN-PC-11 dataset. The study of the model on the basis of the type of plagiarism case — translated obfuscation, translated manual obfuscation, as well as short, medium, and long cases — proved also its superiority. This confirms that, when combined, knowledge graphs and continuous models outperform the results obtained independently, capturing different aspects of text and complementing each other. This also proves the robustness 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, resulted to be the best alternative. The study of the computational efficiency of the models provided with average results for knowledge graph-based models. It also showed that continuous representation models are the fastest ones and that CWASA is the second in terms of similarity calculation. These facts prove the potential of CWASA in scenarios such as companies

where the continuous vectors have already been estimated and the similarity function needs to be fast.

As 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 on monolingual similarity with other continuous word representations such as the popular continuous skip-gram and continuous bag-of-words models [29,30]. Finally, we are interested in exploring postprocessing alternatives in order to detect plagiarism more accurately.

Acknowledgements

This research has been carried out in framework of the FPI-UPV pre-doctoral grant (№ de registro - 3505) awarded to Parth Gupta and in the framework of the national projects DIANA-APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01), and SomEMBED: SOcial Media language understanding - EMBEDing contexts (TIN2015-71147-C2-1-P). We would like to thank Martin Potthast, Daniel Ortiz-Martínez, and Luis A. Leiva for their support and comments during this research.

References

[1] S. Banerjee, T. Pedersen, Extended gloss overlaps as a measure of semantic relatedness, in: IJCAI, vol. 3, 2003, pp. 805–810.
[2] A. Barrón-Cedeño, On the mono- and cross-language detection of text re-use and plagiarism, Universitat Politènica de València, 2012 Ph.D. thesis.
[3] A. Barrón-Cedeño, P. Gupta, P. Rosso, Methods for cross-language plagiarism detection, Knowl. Based Syst. 50 (2013) 211–217.
[4] A. Barrón-Cedeño, P. Rosso, D. Pinto, A. Juan, On cross-lingual plagiarism analysis using a statistical model, in: Proc. of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, in: PAN'08, 2008.
[5] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[6] J. Bromley, J.W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, R. Shah, Signature verification using a "siamese" time delay neural network, IJPRAI 7 (4) (1993) 669–688, doi:10.1142/S0218001493000339.
[7] P. Clough, et al., Old and new challenges in automatic plagiarism detection, 2003. National Plagiarism Advisory Service, 2003; http://ir.shef.ac.uk/cloughie/index.html, Citeseer.
[8] R. Corezola Pereira, V. Moreira, R. Galante, A new approach for cross-language plagiarism analysis, in: M. Agosti, N. Ferro, C. Peters, M. de Rijke, A. Smeaton (Eds.), Multilingual and Multimodal Information Access Evaluation, Lecture Notes in Computer Science, vol. 6360, Springer Berlin Heidelberg, 2010, pp. 15–26.
[9] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, Indexing by latent semantic analysis, JASIS 41 (6) (1990) 391–407.
[10] S. Dumais, T.K. Landauer, M.L. Littman, Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing, in: AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval, 1997, pp. 18–24.
[11] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, CRC Press, 1994.
[12] C. Fellbaum, WordNet: An Electronic Lexical Database, Bradford Books, 1998.
[13] M. Franco-Salvador, P. Gupta, P. Rosso, Cross-language plagiarism detection using a multilingual semantic network, in: Proc. of the 35th European Conference on Information Retrieval (ECIR'13), in: LNCS(7814), Springer-Verlag, 2013, pp. 710–713.
[14] M. Franco-Salvador, P. Rosso, M. Montes y Gómez, A systematic study of knowledge graph analysis for cross-language plagiarism detection, Inf. Process. Manage. 52 (4) (2016) 550–570. http://dx.doi.org/10.1016/j.ipm.2015.12.004.
[15] M. Franco-Salvador, P. Rosso, R. Navigli, 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, 2014, pp. 414–423.
[16] E. Gabrilovich, S. Markovitch, Computing semantic relatedness using wikipedia-based explicit semantic analysis., in: IJCAI, vol. 7, 2007, pp. 1606–1611.
[17] P. Gupta, K. Bali, R.E. Banchs, M. Choudhury, P. Rosso, 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, 2014, pp. 677–686.
[18] P. Gupta, R.E. Banchs, P. Rosso, Continuous Space Models for Clir, Technical Report, Universitat Politécnica de Valéncia, 2015.
[19] P. Gupta, A. Barrón-Cedeño, P. Rosso, 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, in: LNCS(7488), Springer-Verlag, 2012, pp. 67–75.
[20] S. Hassan, R. Mihalcea, Semantic relatedness using salient semantic analysis, in: Proceedings of the 25th AAAI Conference on Artificial Intelligence, 2011, pp. 884–889.
[21] K.M. Hermann, P. Blunsom, Multilingual models for compositional distributed semantics, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 58–68.
[22] G. Hinton, R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[23] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, L. Heck, Learning deep structured semantic models for web search using clickthrough data, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, ACM, 2013, pp. 2333–2338.
[24] D.A. Jackson, K.M. Somers, H.H. Harvey, Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? Am. Nat. 133 (3) (1989) 436–453.
[25] S. Lauly, A. Boulanger, H. Larochelle, Learning multilingual word representations using a bag-of-words autoencoder, CoRR abs/1401.1803(2014a).
[26] S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V.C. Raykar, A. Saha, 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, 2014b, pp. 1853–1861.
[27] H.A. Maurer, F. Kappe, B. Zaka, Plagiarism-a survey., J. UCS 12 (8) (2006) 1050–1084.
[28] P. Mcnamee, J. Mayfield, Character n-gram tokenization for European language text retrieval, Inf. Retrieval 7 (1) (2004) 73–97.
[29] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proceedings of Workshop at International Conference on Learning Representations, 2013a, pp. 1–12.
[30] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems 26, 2013b, pp. 3111–3119.
[31] R. Navigli, S.P. Ponzetto, BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artif. Intell. 193 (2012) 217–250.
[32] F.J. Och, H. Ney, A systematic comparison of various statistical alignment models, Comput. Ling. 29(1) (2003) 19–51.
[33] D. Pinto, J. Civera, A. Barrón-Cedeno, A. Juan, P. Rosso, A statistical approach to crosslingual natural language tasks, J. Algorithms 64 (1) (2009) 51–60.
[34] J.C. Platt, K. Toutanova, W. tau Yih, Translingual document representations from discriminative projections, in: EMNLP, 2010, pp. 251–261.
[35] M. Potthast, A. Barrón-Cedeño, A. Eiselt, B. Stein, P. Rosso, Overview of the 2nd international competition on plagiarism detection., CLEF (Notebook Papers/LABs/Workshops), 2010.
[36] M. Potthast, A. Barrón-Cedeño, B. Stein, P. Rosso, Cross-language plagiarism detection, Lang. Resour. Eval. 45 (1) (2011) 45–62. Special Issue on Plagiarism and Authorship Analysis
[37] M. Potthast, A. Eiselt, A. Barrón-Cedeño, B. Stein, P. Rosso, Overview of the 3rd int. competition on plagiarism detection, CLEF (Notebook Papers/Labs/Workshop), 2011b.
[38] M. Potthast, M. Hagen, A. Beyer, M. Busse, M. Tippmann, P. Rosso, B. Stein, Overview of the 6th international competition on plagiarism detection, in: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15–18, 2014., 2014, pp. 845–876.
[39] M. Potthast, B. Stein, M. Anderka, A wikipedia-based multilingual retrieval model, in: Advances in Information Retrieval, Springer, 2008, pp. 522–530.
[40] M. Potthast, B. Stein, A. Barrón-Cedeño, P. Rosso, An evaluation framework for plagiarism detection, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010b, pp. 997–1005.
[41] R. Salakhutdinov, G. Hinton, Semantic hashing, Int. J. Approx. Reasoning 50 (7) (2009) 969–978, doi:10.1016/j.ijar.2008.11.006.
[42] Y. Shen, X. He, J. Gao, L. Deng, G. Mesnil, Learning semantic representations using convolutional neural networks for web search, in: Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, pp. 373–374.
[43] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, D. Varga, The jrc-acquis: A multilingual aligned parallel corpus with +20 languages, in: Proc. 5th International Conference on language resources and evaluation (LREC'06), 2006.
[44] W. Yih, K. Toutanova, J.C. Platt, C. Meek, 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, 2011, pp. 247–256.