Word Embedding-Based Approaches for Measuring Semantic Similarity of Arabic-English Sentences

El Moatez Billah Nagoudi[1](✉), Jérémy Ferrero[2,3], Didier Schwab[3], and Hadda Cherroun[1]

[1] Laboratoire d'Informatique et de Mathématique LIM,
Amar Telidji University, Laghouat, Algeria
{e.nagoudi,h.cherroun}@lagh-univ.dz

[2] Compilatio, 276 rue du Mont Blanc, 74540 Saint-Félix, France

[3] LIG-GETALP, Univ. Grenoble Alpes, Grenoble, France
{jeremy.ferrero,didier.schwab}@imag.fr

Abstract. Semantic Textual Similarity (STS) is an important component in many Natural Language Processing (NLP) applications, and plays an important role in diverse areas such as information retrieval, machine translation, information extraction and plagiarism detection. In this paper we propose two word embedding-based approaches devoted to measuring the semantic similarity between Arabic-English cross-language sentences. The main idea is to exploit Machine Translation (MT) and an improved word embedding representations in order to capture the syntactic and semantic properties of words. MT is used to translate English sentences into Arabic language in order to apply a classical monolingual comparison. Afterwards, two word embedding-based methods are developed to rate the semantic similarity. Additionally, Words Alignment (WA), Inverse Document Frequency (IDF) and Part-of-Speech (POS) weighting are applied on the examined sentences to support the identification of words that are most descriptive in each sentence. The performances of our approaches are evaluated on a cross-language dataset containing more than 2400 Arabic-English pairs of sentence. Moreover, the proposed methods are confirmed through the Pearson correlation between our similarity scores and human ratings.

Keywords: Semantic sentences similarity · Cross-language Arabic-English · Machine translation · Word embedding

1 Introduction

Semantic Textual Similarity (STS) is the task of measuring the degree of semantic equivalence between two textual units (texts, paragraphs or sentences) [1]. STS is a core field of Natural Language Processing (NLP) and plays an important role in several application areas, such as Information Retrieval (IR), Word

A. Lachkar et al. (Eds.): ICALP 2017, CCIS 782, pp. 19–33, 2018.
https://doi.org/10.1007/978-3-319-73500-9_2

Sense Disambiguation (WSD), Question Answering (QA), and Text Summarization (TS) among others. There are two known types of STS: monolingual and cross-language [3]. The first one estimates the degree to which the underlying semantics of two textual units written in the same language, are equivalent to each other, while the STS cross-language aims to quantify the degree to which two textual units are semantically related, independent of the languages they are written in [15].

Determining the similarity between sentences has been extensively reviewed in a monolingual domain [4,20,37,43]. While cross-language semantic similarity is relatively more difficult to identify since the relatedness of words are investigated between two different languages [15]. Thus, it is necessary to address this task to enhance the performance in several applications, such as Machine Translation (MT), Cross-Language Plagiarism Detection (CLPD) and Cross-Language Information Retrieval (CLIR).

In this paper we focus our investigation on measuring the semantic similarity between Arabic-English cross-language sentences using machine translation and word embedding representations. We also consider words alignment, term frequency weighting and Part-of-Speech tagging to improve the identification of words that are highly descriptive in each sentence.

The rest of this paper is organized as follows, the next section describes work related to STS cross-language detection and word embedding models. In Sect. 3, we present our proposed cross-language word embedding-based methods. Section 4 describes the experimental results of these systems. Finally, our conclusion and some future research directions are drawn in Sect. 5.

2 Related Work

In this section, we review the most relevant approaches for measuring cross-language semantic textual. Then, we study those dedicated to the Arabic-English semantic similarity. Finally, we recall some concepts related to word embedding.

2.1 Cross-Language Semantic Textual Similarity Detection

In the literature, many approaches are proposed for cross-language textual similarity detection. We can classify them according to the strategy they used to detect such similarity into five classes: Syntax-Based, Dictionary-Based, Parallel and Comparable Corpora-Based and MT-Based Models [10]. Figure 1 shows the taxonomy of different approaches for cross-language similarity detection. In the following, we will review the most commonly used methods.

Concerning the syntax-based models, the key idea lies in comparing multilingual texts without translation. For instance, Pouliquen et al. [16] have proposed a *"Length Model"* to estimate cross-language text similarity. It is mainly based on comparing the texts size. They observed the fact that the length of texts in different languages are closely linked by a factor, and there is a different factor for each language pair. McNamee and Mayfield [22] have introduced Cross-Language

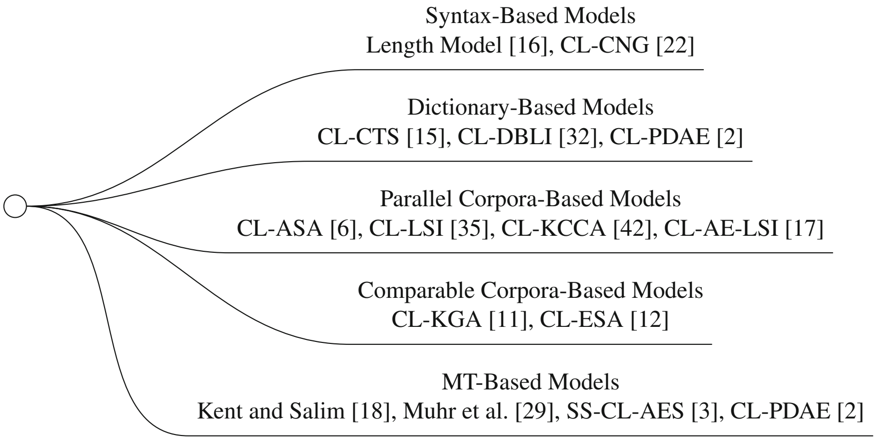
Syntax-Based Models
Length Model [16], CL-CNG [22]
Dictionary-Based Models
CL-CTS [15], CL-DBLI [32], CL-PDAE [2]
Parallel Corpora-Based Models
CL-ASA [6], CL-LSI [35], CL-KCCA [42], CL-AE-LSI [17]
Comparable Corpora-Based Models
CL-KGA [11], CL-ESA [12]
MT-Based Models
Kent and Salim [18], Muhr et al. [29], SS-CL-AES [3], CL-PDAE [2]

Fig. 1. Taxonomy of different approaches for cross-language similarity detection [10].

Character N-Gram (CL-CNG) model to compare two textual units by using their n-gram vectors representation. This technique achieves a good performance with languages that are close to each other, because of common root words.

In dictionary-based models, the semantic similarity is measured by constructing a vector space model of the textual units. For that, a vector of concepts is built for each textual unit using dictionaries or thesaurus. The similarity between the vectors of concepts can be measured using the Cosine similarity, Euclidean distance, or any other similarity measure. In [15] a Cross-Language Conceptual Thesaurus-Based Similarity model (CL-CTS) is proposed to measure the similarity between textual units written in different languages (Spanish, English and German). CL-CTS is based on the thesaurus concepts vectors presented in Eurovoc[1] where a Cosine similarity is computed between these vectors. In the same context, Pataki [32] have proposed a Cross-Language Dictionary-Based Language-Independent (CL-DBLI) model. CL-DBLI considers a translation synonym dictionary to extract the abstract concepts from words in textual units.

For comparable corpora-based models, Gabrilovich and Markovitch [12] have presented a Cross-Language Explicit Semantic Analysis (CL-ESA) model. CL-ESA is based on the Explicit Semantic Analysis (ESA), which represent the meaning of text by a vector of concepts derived from Wikipedia. In a cross-lingual context, Potthast et al. [36] use Wikipedia as comparable corpus to estimate the similarity of two documents by calculating the similarity of their two ESA representations. Another model called Cross-Language Knowledge Graph Analysis (CL-KGA), is introduced for the first time by Franco-Salvador et al. [11]. CL-KGA uses knowledge graphs built from multilingual semantic network (the authors use BabelNet [31]) to represent texts, and then compare them in a common lingual semantic graph space.

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

Regarding parallel corpora-based models, several approaches are proposed. For instance, Barrón-Cedeño et al. [6] have introduced a Cross-Language Alignment Similarity Analysis (CL-ASA) approach. CL-ASA estimates the similarity between two textual units using bilingual statistical dictionary extracted from parallel corpus. The same idea was used independently by Pinto et al. [34]. A Cross-Language Latent Semantic Indexing model (CL-LSI) is developed by Potthast et al. [35]. CL-LSI uses a parallel corpora with the common Latent Semantic strategy applied in IR systems for term-textual unit association. Another model named Cross-Language Kernel Canonical Correlation Analysis (CL-KCCA) model due to Vinokourov et al. [42], it analyzes the correspondences between two LSI spaces to measure the correlation of the respective projection values.

The main idea of the machine translation-based models consists in using MT tools to translate textual units into the same language (pivot language) in order to apply a monolingual comparison between them [5]. For this purpose, Kent and Salim [18] have used Google Translate API to translate texts, while Muhr et al. [29] replace each word of the original text by its most likely translations in the target language.

2.2 Arabic-English Cross-Language Semantic Similarity

In context of the Arabic-English cross-language semantic similarity, Hattab [17] has used Latent Semantic Indexing (LSI) to build cross-language Arabic-English semantic space (CL-AE-LSI), from which it checks the contextual similarity of two given texts, one in Arabic and the other in English.

Recently, Alzahrani [3] presented two models of Semantic Similarity for Arabic-English Cross-Language Sentences (SS-CL-AES). The first one used a dictionary-based translation, where an Arabic sentence is translated into English terms, then the semantic similarity is computed by using the maximum-translation similarity technique. In the second model, MT is applied on the Arabic sentence. After that, the algorithms proposed by Lee [19], and Liu et al. [21] are used to calculate the semantic similarity.

Alaa et al. [2] are interested in Cross-Language Plagiarism Detection of Arabic-English documents (CL-PDAE). In fact, after a candidate document retrieval step by key phrase extraction, they translate a source text by getting for a word all the available translations of all its available synonyms from WordNet [27], and then they use a combination of monolingual measures (Longest Common Subsequence (LCS), Cosine similarity and N-Gram) to detect similar phrases.

2.3 Word Embedding-Based Models

Recently, Word Embedding (WE) technique has received a lot of attention in the NLP community and has become a core building to many NLP applications. WE represents words as vectors in a continuous high-dimensional space. These representations allow capturing semantic and syntactic properties of the language [23]. In the literature, several techniques are proposed to build word embedding models.

For instance, Collobert and Weston [9] have presented a unified system based on a deep neural network, and jointly trained with many NLP tasks, such as: POS tagging, Semantic Role Labeling and Named Entity Recognition. Their model is stored in a matrix $M \in R^{d*|D|}$, where D represents the dictionary of all unique words in the training data, and each word in D is embedded into a *d-dimensional* vector. The sentences are represented using the embeddings of their forming words. A similar idea was independently proposed and used by Turian et al. [41].

Mnih and Hinton [28] have introduced another form to represent words in vector space, named Hierarchical Log-Bilinear Model (HLBL). Like almost all neural language models, the HLBL model is used to represent each word by a real-valued feature vector. HLBL concatenates the $(n-1)$ first embedding words $(w_1 \ldots w_{n-1})$ and learns a neural linear model to predicate the last word w_n.

In Mikolov et al. [26] a recurrent neural network (RNN) [24] is used to build a neural language model. The RNN model encode the context word by word and predict the next word. Afterwards, the weights of the trained network are considered as the word embedding vectors.

Based on the simplified neural language model of Bengio et al. [7], Mikolov et al. [23,25] presented two other techniques to build a words representations model. In their work, two models are proposed: the continuous bag-of-words (CBOW) model [23], and the skip-gram (SKIP-G) model [25]. The CBOW model, predicts a pivot word according to the context by using a window of contextual words around it. Given a sequence of words $S = w_1, w_2, ..., w_i$, the CBOW model learns to predict all words w_k from their surrounding words $(w_{k-l}, ..., w_{k-1}, w_{k+1}, ..., w_{k+l})$. The second model, SKIP-G, predicts surrounding words of the current pivot word w_k [25].

Pennington et al. [33] proposed a Global Vectors (GloVe) model to representing words in vector space. GloVe model builds a co-occurrence matrix M using the global statistics of word-word co-occurrence. Afterwards, the matrix M is used to estimate the probability of word w_i to appear in the context of another word w_j, this probability $P(i/j)$ represents the relationship between words.

In a comparative study conducted by Mikolov et al. [23] all the methods [9, 23,25,26,28,41] have been evaluated and compared, and they show that CBOW [23] and SKIP-G [25] models are significantly faster to train with better accuracy. For this reason, we have used the CBOW word representations for Arabic model, proposed by Zahran et al. [45]. In order to train this model, they have used a large collection from different sources counting more than 5.8 billion words[2].

In the Arabic CBOW model [45] each word w is represented by a vector v of *d-dimension*. The similarity between two words w_i and w_j (e.g. synonyms, singular, plural, feminization or closely related semantically) is obtained by comparing their vector representations v_i and v_j respectively [23]. This similarity can be evaluated using the Cosine similarity, Euclidean distance, Manhattan distance

[2] https://sites.google.com/site/mohazahran/data.

or any other similarity measure functions. For example, let "الجامعة" (*university*), "المساء" (*evening*) and "الكلية" (*faculty*) be three words. The similarity between them is measured by computing the cosine similarity between their vector as follows:

$$Sim(\text{المساء}, \text{الجامعة}) = Cos(v(\text{المساء}), v(\text{الجامعة})) = 0.13$$

$$Sim(\text{الكلية}, \text{الجامعة}) = Cos(v(\text{الكلية}), v(\text{الجامعة})) = 0.72$$

That means that, the words "الكلية" (*faculty*) and "الجامعة" (*university*) are semantically closer than "المساء" (*evening*) and "الجامعة" (*university*). In the following, we exploit this property to measure the semantic similarity at sentence level.

3 Proposed Methods

In this section, we present our two proposed methods for Arabic-English cross-language sentence similarity. These methods use Machine Translation-Based Model, followed by a monolingual semantic similarity analysis based on word embedding. They consist of three steps, including translation, preprocessing and similarity score attribution. First, MT is used to translate English sentences into Arabic. Afterwards, our two word embedding-based methods are employed to measure the semantic similarity of Arabic sentences. In the first one, we propose to use the words alignment technique proposed by Sultan et al. [39] with the words weighting methods of Nagoudi and Schwab [30], we call this method *Weighting Aligned Words* (W-AW). The second generate a Bag-of-Words for the aligned words to construct a vector representation of each sentence. Then the similarity is obtained by comparing the two sentence vectors, we name this method *Bag-of-Words Alignment* (BoW-A). Figure 2 gives an overview of the proposed methods.

Let $S_E = w_{e_1}, w_{e_2}, ..., w_{e_i}$ and $S_A = w_{a_1}, w_{a_2}, ..., w_{a_j}$ be an English and Arabic sentence, and their word vectors are $(v_{e_1}, v_{e_2}, ..., v_{e_i})$ and $(v_{a_1}, v_{a_2}, ..., v_{a_j})$ respectively. The semantic similarity between S_E and S_A is computed in three steps: translation, preprocessing and a monolingual similarity score attribution.

(1) **Translation:** in this step, we used Google Translate API[3] to translate English sentences into Arabic language, we denote the translated sentence $S_{E'}$. By this translation, the problem is reduced into a mono-lingual semantic similarity one.

(2) **Preprocessing:** in order to normalize the sentences for the similarity evaluation step, a set of preprocessing are performed:
   - Tokenization: input sentences are broken up into words;
   - Removing punctuation marks, diacritics, and non alphanumeric characters;
   - Normalizing أ ، إ ، آ to ا and ة to ه , as in the Arabic CBOW model [45];
   - Replacing final ى followed by ء by ئ.

[3] https://cloud.google.com/translate/.

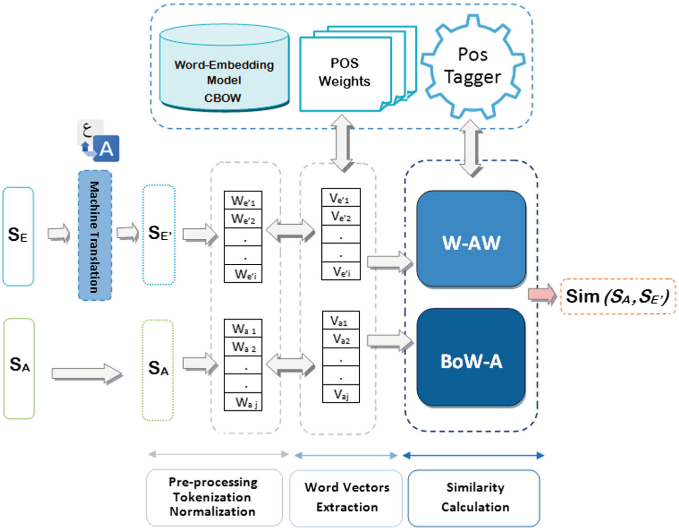

Fig. 2. Overview of the proposed methods

At this point we should mention that we will not remove the stop words because they can affect the similarity score, For example: S_E =“Joseph went to university” and $S_A =$ “يوسف لم يذهب إلى الجامعة” (Joseph does not go to university). If we remove the words لم, إلى and **to** as a stop words, both sentences become similar, whereas they have contradictory meanings.

(3) **Sentences similarity:** we propose two methods for measuring the semantic similarity between $S_{E'}$ and S_A: Weighting Aligned Words Method (W-AW) and Bag-of-Words Alignment Method (BoW-A). In the following, we develop our proposed methods, and we provide for each one how the semantic similarity is measured.

3.1 Weighting Aligned Words Method (W-AW)

A simple way to compare the translated sentence $S_{E'}$ and the Arabic one S_A is by summing their words vectors [30]. Then, the similarity is obtained by calculating the Cosine similarity $Cos(V_{E'}, V_A)$, where:

$$\begin{cases} V_{E'} &= \sum_{k=1}^{i} v_{e'_k} \\ V_A &= \sum_{k=1}^{j} v_{a_k} \end{cases}$$

For example, let S_E and S_A be two sentences:
$S_E =$ “Joseph went to college”.
$S_A =$ “يوسف يمضى مسرعا للجامعة” (*Joseph goes quickly to university*).

Step 1: Translation
In this step Google Translate API is used to translate the English sentence S_E into Arabic $S_{E'}$ = "ذهب يوسف إلى الكلية".
Step 2: Sum of the word vectors

$$\begin{aligned} V_{E'} &= V(\text{الكلية}) + V(\text{إلى}) + V(\text{يوسف}) + V(\text{ذهب}) \\ V_A &= V(\text{للجامعة}) + V(\text{مسرعا}) + V(\text{يمضي}) + V(\text{يوسف}) \end{aligned}$$

Step 3: Similarity Score
The similarity between $S_{E'}$ and S_A is obtained by calculating the cosine similarity between the sentence vectors $V_{E'}$ and V_A as follows:

$$Sim(S_E, S_A) = Sim(S_{E'}, S_A) = Cos(V_{E'}, V_A) = 0.71$$

In order to improve the similarity results, we have used the words alignment method presented by Sultan et al. [39], with the difference that we align the words based on their semantic similarity in the word embedding model, and not in a dictionary. We assume also that all words don't have the same importance for the meaning of the sentences. For that, we use three weighting functions (*idf*, *pos* and *idf-pos*) proposed by Nagoudi and Schwab in [30] for weighting the aligned words. Finally, the similarity between $S_{E'}$ and S_A is calculated as follows:

$$Sim(S_{E'}, S_A) = \frac{1}{2}\left(\frac{\sum_{w \in S_{E'}} WT(w) * BM(w, S_A)}{\sum_{w \in S_{E'}} WT(w)} + \frac{\sum_{w \in S_A} WT(w) * BM(w, S_{E'})}{\sum_{w \in S_A} WT(w)}\right) \quad (1)$$

where $WT(w)$ is the function which return the weight of the word w. WT uses three weighting methods: *idf*, *pos* and a mix of both. The $BM(w, S_k)$ function represent the *Best Match* score between w and all words in the sentence S_k. Therefore, BM function aligns words based on their semantic similarity included in the word embedding model. The function BM is defined as:

$$BM(w, S_k) = Max\{Cos(v, v_k),\ w_k \in S_k\} \quad (2)$$

For example, let us continue with the same example above, the similarity between $S_{E'}$ and S_A is obtained in four steps as follows:

Step 1: POS Tagging
Firstly, the POS tagger of Gahbiche-Braham et al. [13] is used to predict the part-of-speech tag of each word w_k in S_k.

$$\begin{cases} Pos\_tag(S_{E'}) = verb\ noun\_prop\ noun \\ Pos\_tag(S_A) = noun\_prop\ verb\ adj\ noun \end{cases}$$

Step 2: IDF & POS Weighting
For weighting the descriptive aligned words, we retrieve for each word w_k in the S_k its IDF weight $idf(w_k)$, we also use the POS weights proposed in [30].
Step 3: Words Alignment
In this step, we align words that have similar meaning in both sentences. For that, we compute the similarity between each word in $S_{E'}$ and the semantically closest word in S_A by using the BM function, e.g. $BM(\text{يمضي}, S_{E'}) = Max\{Cos(\text{يمضي}, v_k),\ w_k \in S_A\} = Cos(v(\text{يمضي}), v(\text{ذهب})) = 0.85$.

Step 4: Calculate the similarity
The similarity between $S_{E'}$ and S_A is obtained by using the formula (1), which gives us: $Sim(S_{E'}, S_A) = 0.82$.

3.2 Bag-of-Words Alignment Method (BoW-A)

Among the advantages of word embedding is that it allows to retrieve a list of words that are used in the same contexts with respect to a given word w [14]. We named this list of words the *k-closest* words to w. For example, Table 1 shows the 10-closest words of الجامعة and الكلية in the Arabic CBOW model.

Table 1. 10-closest words of الجامعة and الكلية.

BoW(الجامعة)	BoW(الكلية)
الجامعات, للجامعة, الكلية, بالجامعة العمادة, جامعة, الاكاديمية, جامعتنا حرم الجامعة, الجامعات	الاكاديمية, الجامعة, كلية الطب, كليتنا للجامعة, الجامعات, جامعة, للجامعة كليات الجامعة, العمادة

We used this property to evaluate the degree of semantic similarity between $S_{E'}$ and S_A, we first proceeded to construct a representation vector RV for each sentence. Let $RV_{E'}$ and RV_A be the representation vectors of $S_{E'}$ and S_A respectively, the size of each vector is the number of words in its corresponding sentence. The value of an entry in the representation vector, is determined as follows:

1. For each word w we retrieve its aligned word w' in the other sentence by using BM function defined by formula (2).
2. We use the embedding model to construct for both w and w' their Bag-of-Words BoW_w and $BoW_{w'}$. The BoW_w ($BoW_{w'}$) contains the *k-closest* words to w (w') in the embedding model.
3. We compute the Jaccard similarity between BoW_w and $BoW_{w'}$:

$$Jacc(BoW_w, BoW_{w'}) = \frac{BoW_w \cap BoW_{w'}}{BoW_w \cup BoW_{w'}}$$

4. The value of the entry RV[w] is set to $Jacc(BoW_w, BoW_{w'})$.
5. This process is applied for all words in both sentences to build $RV_{E'}$ and RV_A.
6. Finally, the similarity between $S_{E'}$ and S_A is obtained by:

$$Sim(S_{E'}, S_A) = \frac{1}{2}\left(\frac{\sum_{w \in S_{E'}} WT(w) * RV[w]}{\sum_{w \in S_{E'}} WT(w)} + \frac{\sum_{w \in S_A} WT(w) * RV[w']}{\sum_{w \in S_A} WT(w)}\right) \quad (3)$$

4 Experiments and Results

In order to evaluate our systems and monitor their performances, we have used four datasets drawn from the STS shared task SemEval-2017 (Task1: STS Cross-lingual Arabic-English)[4] [8], with a total of 2412 pairs of sentences. The sentence pairs have been manually labeled by five annotators, and the similarity score is the mean of the five annotators' judgments. This score is a float number between "0" (indicating that the meaning of sentences are completely independent) to "5" (indicating meaning equivalence). More information about the datasets used is listed in Table 2.

Table 2. Arabic-English evaluation sets.

Dataset	Source	Pairs
MSRvid	Microsoft research video description corpus	736
MSRpar	Microsoft research paraphrase corpus	1020
SMTeuroparl	WMT2008 development dataset	406
STS evaluation data	SNLI corpus	250

4.1 Experimental Results

We investigated the performance of both Weighting Aligned Words (W-AW) and Alignment Bag-of-Words (A-BoW) systems with three weighting functions: IDF, POS and mix of both. In addition, for the A-BoW method, we have used four different values of k to generate the 5-*closest*, 10-*closest*, 15-*closest* and 20-*closest* words. Afterwards, in order to evaluate the accuracy of each method, we calculate the Pearson correlation between our assigned semantic similarity scores and human judgments on the SemEval STS task datasets. Table 3 reports the results of the proposed methods.

These results indicate that when the IDF weighting method is used the mean correlation rate does not fall below 70% in all tested methods. When applying the POS and mixed weighting, the correlation rate of IDF weighting is outperformed in both methods A-AW and A-BoW with a mean of +2.35% and +3.91% respectively. Interestingly, increasing the parameter k to generate the k-*closest* words in the A-BoW method, leads each time to an enhancement in the correlation rate. For instance, the use of 15-*closest* words outperforms the 5-*closest* system by +2.01% of correlation in average. However, when k is raised to 20, the mean correlation rate gets a bit lower. This is due to the rise of the number of words with different meaning in the BoW.

From the above results, we can see that the estimated similarity provided by our approaches is fairly consistent with human judgments. However, the correlation is not good enough when two sentences share nearly the same words, but with a totally different meaning, for example:

[4] http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools.

Table 3. Our methods vs. human judgments

Method	MSRvid	MSRpar	SMTeuro.	STS Eval.	Mean
W-AW-IDF	0.6895	0.7019	0.7274	0.6951	0.7034
W-AW-POS	0.6924	0.7402	0.7478	0.7205	0.7252
W-AW-IDF-POS	0.7015	0.7385	0.7512	0.7375	**0.7321**
$k = 5$					
A-BoW-IDF	0.6863	0.7119	0.7174	0.6881	0.7009
A-BoW-POS	0.6933	0.7349	0.7364	0.7187	0.7218
A-BoW-IDF-POS	0.7074	0.7365	0.7482	0.7362	**0.7320**
$k = 10$					
A-BoW-IDF	0.6879	0.7131	0.7291	0.7114	0.7103
A-BoW-POS	0.7084	0.7437	0.7514	0.7305	0.7335
A-BoW-IDF-POS	0.7216	0.7418	0.7603	0.7565	**0.7450**
$k = 15$					
A-BoW-IDF	0.6954	0.7089	0.7284	0.7254	0.7145
A-BoW-POS	0.7124	0.7402	0.7578	0.7391	0.7398
A-BoW-IDF-POS	0.7575	0.7485	0.7672	0.7739	**0.7603**
$k = 20$					
A-BoW-IDF	0.6912	0.7055	0.7283	0.7254	0.7244
A-BoW-POS	0.7254	0.7382	0.7514	0.7351	0.7351
A-BoW-IDF-POS	0.7525	0.7477	0.7689	0.7613	**0.7576**

“يقرأ سعد كتابا عن عمر بن الخطاب” and *(Saad reads a book about Omar Ibn Al-Khattab)* “سعد يقرأ كتابا لعمر بن الخطاب” *(Saad reads a book for Omar Ibn Al-Khattab)*. In this example, the sentences share the same vectors, POS and IDF weights. This fact leads to a high correlation score, which is not the case. This issue is left for future work.

4.2 Comparison with SemEval-2017 Winners

We compared our optimal results with the three best systems proposed in SemEval-2017 Arabic-English cross-lingual evaluation task [8] (ECNU [40], BIT [44] and HCTI [38]) and the baseline system [8]. In this evaluation, ECNU obtained the best performance with a correlation score of 74.93%, followed by BIT and HCTI with 70.07% and 68.36% respectively. Table 4 shows a comparison of our best results with those obtained by the three systems were tested on the STS Evaluation Data[5].

The observed results indicate that our mixed weighted method with $k = 15$ is the best performing method with a correlation rate of 77.39%. The W-BoW-

[5] http://alt.qcri.org/semeval2017/task1/data/uploads/sts2017.eval.v1.1.zip.

Table 4. Comparison of the correlation results with three best systems in SemEval-2017.

Methods	STS Eval.
W-BoW-IDF-POS ($k = 15$)	**77.39%**
ECNU	74.93 %
W-AW-IDF-POS	73.75%
BIT	70.07 %
HCTI	68.36%
Cosine baseline	51.55 %

IDF-POS ($k = 15$) method yields a gain of +9.03%, +7.32% and +2.46% on the correlation rate compared with ECNU, BIT and HCTI respectively.

5 Conclusion and Future Work

In this paper, we have presented two methods for measuring the semantic relations between Arabic-English cross-language sentences using Machine Translation (MT) and word embedding representations. The main idea is based on the usage of semantic properties of words included in the word-embedding model. In order to make further progress in the analysis of the semantic sentence similarity, we have used a combination of words alignment, IDF and POS weighting to support the identification of words that are most descriptive in each sentence. Additionally, we evaluated our proposals on the four datasets of the STS shared task SemEval-2017. In the experiments we have shown how the Bag-of-words method clearly enhanced the correlation results. The performance of our proposed methods was confirmed through the Pearson correlation between our assigned semantic similarity scores and human judgments. In fact, we reached the best correlation rate compared to all the participating systems in STS Arabic-English cross-language subtask of SemEval-2017. As future work, we are going to combine these methods with those of other classical techniques in NLP field, including word sense disambiguation, linguistic resources and document fingerprint in order to make more improvement in the cross-language plagiarism detection.

References

1. Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., Wiebe, J.: Semeval-2016 task 1: semantic textual similarity, monolingual and cross-lingual evaluation. In: Proceedings of SemEval, pp. 497–511 (2016)
2. Alaa, Z., Tiun, S., Abdulameer, M.: Cross-language plagiarism of Arabic-English documents using linear logistic regression. J. Theor. Appl. Inf. Technol. **83** (2016)
3. Alzahrani, S.: Cross-language semantic similarity of Arabic-English short phrases and sentences. J. Comput. Sci. **12**, 1–18 (2016)

4. Bär, D., Biemann, C., Gurevych, I., Zesch, T.: UKP: computing semantic textual similarity by combining multiple content similarity measures. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics, pp. 435–440. Association for Computational Linguistics (2012)
5. Barrón-Cedeño, A., Gupta, P., Rosso, P.: Methods for cross-language plagiarism detection. Knowl. Based Syst. **50**, 211–217 (2013)
6. Barrón-Cedeno, A., Rosso, P., Pinto, D., Juan, A.: On cross-lingual plagiarism analysis using a statistical model. In: PAN, pp. 1–10 (2008)
7. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
8. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: Semeval-2017 task 1: semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, Canada, pp. 1–14. Association for Computational Linguistics, August 2017
9. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine learning, pp. 160–167. ACM (2008)
10. Ferrero, J., Agnes, F., Besacier, L., Schwab, D.: A multilingual, multi-style and multi-granularity dataset for cross-language textual similarity detection. In: 10th Edition of the Language Resources and Evaluation Conference (2016)
11. Franco-Salvador, M., Gupta, P., Rosso, P.: Cross-language plagiarism detection using a multilingual semantic network. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 710–713. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36973-5_66
12. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), Hyderabad, India, pp. 1606–1611. Morgan Kaufmann Publishers Inc., January 2007
13. Gahbiche-Braham, S., Bonneau-Maynard, H., Lavergne, T., Yvon, F.: Joint segmentation and POS tagging for Arabic using a CRF-based classifier. In: LREC, pp. 2107–2113 (2012)
14. Ganguly, D., Roy, D., Mitra, M., Jones, G.J.: Word embedding based generalized language model for information retrieval. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 795–798. ACM (2015)
15. Gupta, P., Barrón-Cedeño, A., Rosso, P.: Cross-language high similarity search using a conceptual thesaurus. In: Catarci, T., Forner, P., Hiemstra, D., Peñas, A., Santucci, G. (eds.) CLEF 2012. LNCS, vol. 7488, pp. 67–75. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33247-0_8
16. Happe, A., Pouliquen, B., Burgun, A., Cuggia, M., Le Beux, P.: Automatic concept extraction from spoken medical reports. Int. J. Med. Inform. **70**, 255–263 (2003)
17. Hattab, E.: Cross-language plagiarism detection method: Arabic vs. English. In: 2015 International Conference on Developments of E-Systems Engineering (DeSE), pp. 141–144. IEEE (2015)
18. Kent, C.K., Salim, N.: Web based cross language plagiarism detection. In: 2010 Second International Conference on Computational Intelligence, Modelling and Simulation (CIMSiM), pp. 199–204. IEEE (2010)
19. Lee, M.C.: A novel sentence similarity measure for semantic-based expert systems. Expert Syst. Appl. **38**, 6392–6399 (2011)

20. Li, Y., McLean, D., Bandar, Z.A., O'shea, J.D., Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowl. Data Eng. **18**, 1138–1150 (2006)
21. Liu, C., Chen, C., Han, J., Yu, P.S.: GPLAG: detection of software plagiarism by program dependence graph analysis. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 872–881. ACM (2006)
22. Mcnamee, P., Mayfield, J.: Character n-gram tokenization for European language text retrieval. Inf. Retr. **7**, 73–97 (2004)
23. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceeding of the International Conference on Learning Representations Workshop Track, ICLR 2013, pp. 1301–3781 (2013)
24. Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: Interspeech, vol. 2, p. 3 (2010)
25. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
26. Mikolov, T., Yih, W.-T., Zweig, G.: Linguistic regularities in continuous space word representations. In: HLT-NAACL, vol. 13, pp. 746–751 (2013)
27. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38**, 39–41 (1995)
28. Mnih, A., Hinton, G.E.: A scalable hierarchical distributed language model. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 21, pp. 1081–1088. Curran Associates Inc. (2009)
29. Muhr, M., Kern, R., Zechner, M., Granitzer, M.: External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system. In: Notebook Papers of CLEF 2010 LABs and Workshops (2010)
30. Nagoudi, E.M.B., Schwab, D.: Semantic similarity of arabic sentences with word embeddings. In: Proceedings of the Third Arabic Natural Language Processing Workshop, pp. 18–24. Association for Computational Linguistics (2017)
31. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. In: Proceedings of Artificial Intelligence, vol. 193, pp. 217–250 (2012)
32. Pataki, M.: A new approach for searching translated plagiarism (2012)
33. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: EMNLP, vol. 14, pp. 1532–1543 (2014)
34. Pinto, D., Civera, J., Barrón-Cedeno, A., Juan, A., Rosso, P.: A statistical approach to crosslingual natural language tasks. J. Algorithms **64**, 51–60 (2009)
35. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-language plagiarism detection. Lang. Resour. Eval. **45**, 45–62 (2011)
36. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-based multilingual retrieval model. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 522–530. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_51
37. Rios, M., Specia, L.: UoW: multi-task learning Gaussian process for semantic textual similarity. In: Proceedings of SemEval, pp. 779–784 (2014)
38. Shao, Y.: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017) (2017)
39. Sultan, M.A., Bethard, S., Sumner, T.: DLS@CU: sentence similarity from word alignment and semantic vector composition. In: Proceedings of the 9th International Workshop on Semantic Evaluation, pp. 148–153 (2015)

40. Tian, J., Zhou, Z., Lan, M., Wu, Y.: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017) (2017)
41. Turian, J., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 384–394. Association for Computational Linguistics (2010)
42. Vinokourov, A., Shawe-Taylor, J., Cristianini, N.: Inferring a semantic representation of text via cross-language correlation analysis. In: NIPS 2002: Advances in Neural Information Processing Systems, pp. 1473–1480 (2003)
43. Wali, W., Gargouri, B., Hamadou, A.B.: Enhancing the sentence similarity measure by semantic and syntactico-semantic knowledge. Vietnam J. Comput. Sci. **4**, 51–60 (2016)
44. Wu, H., Huang, H., Jian, P., Guo, Y., Su, C.: In: Proceedings of the 11th International Workshop on Semantic Evaluation (semeval 2017) (2017)
45. Zahran, M.A., Magooda, A., Mahgoub, A.Y., Raafat, H., Rashwan, M., Atyia, A.: Word representations in vector space and their applications for Arabic. In: Gelbukh, A. (ed.) CICLing 2015. LNCS, vol. 9041, pp. 430–443. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18111-0_32