Journal of Computer Sciences

Original Research Paper

Cross-Language Semantic Similarity of Arabic-English Short Phrases and Sentences

Salha Alzahrani

College of Computers and Information Technology (CIT), Taif University, Saudi Arabia

Article history
Received: 08-01-2016
Revised: 04-03-2016
Accepted: 11-03-2016

Email: 364715358@qq.com

Abstract: Measuring cross-language semantic similarity between short texts is a task that is challenging in terms of human understanding. This paper addresses this problem by carrying out a study of Arabic–English semantic similarity in short phrases and sentences. Human-rated benchmark dataset was carefully constructed for this research. Dictionary and machine translation techniques were employed to determine the relatedness between the cross-lingual texts from a monolingual perspective. Three algorithms were developed to rate the semantic similarity and these were applied to the human-rated benchmark. An averaged maximum-translation similarity algorithm was proposed using the term sets produced by the dictionary-based technique. Noun-verb and term vectors obtained by the Machine Translation (MT) technique were also suggested to compute the semantic similarity. The results were compared with the human ratings in our benchmark using Pearson correlation coefficient and these were triangulated with the best, worst and mean for all human participants. MT-based term vector semantic similarity algorithm obtained the highest correlation (r = 0.8657) followed by averaged maximum-translation similarity algorithm (r = 0.7206). Further statistical analysis showed no significant difference between both algorithms and the humans' judgement.

Keywords: Semantic Similarity, Cross-Language, Machine Translation, Arabic, English

Introduction

Semantic similarity is a measure that shows the connection between two words in a text in terms of the idea conveyed. Semantic similarity in natural language engineering has experienced increasing demand of late in a wide range of applications, including linguistics, cognitive science, information retrieval, biomedical informatics and geo-informatics. Semantic relatedness is an extension of semantic similarity, for example cars and petrol can be seen as being more closely related than cars and bicycles, but the latter pair is certainly more similar (Resnik, 1999). Semantic similarity has been widely explored beyond the word unit to the sentence unit in a monolingual domain (Li *et al*., 2006; O'Shea *et al*., 2008; Bar *et al*., 2012; Jimenez *et al*., 2012; Rios, 2014).

Cross-language semantic similarity is more challenging than monolingual similarity because the semantic relations of terms are evaluated between two different languages. Research studies have found the necessity for cross-language semantic similarity to improve the performance in a number of applications, including Machine Translation (MT) (Zou *et al*., 2013), Cross-Language Information Retrieval (CLIR) (Zhou *et al*., 2012) and plagiarism detection across different languages (Barrón-Cedeño *et al*., 2013). There is certainly a need for research on semantic similarity of short texts in the cross-language domain. In this study, we propose two pre-processing models of semantic similarity for Arabic-English cross-language sentences. The first model includes dictionary-based translation, where an Arabic text is converted into terms, which are then translated into English. The similarity of this English translation is then measured against its English candidate text using the proposed maximum-translation similarity approach. The second model involves using MT followed by a semantic similarity measure of the two texts, based on the algorithms proposed by Lee (2011) and Li *et al.* (2006). Experimental works have been conducted on a human-rated benchmark created from a standard and a ground-truth dataset.

The remainder of this paper is structured as follows. Section 2 provides an overview of the literature on word-to-word, text-to-text and cross-language semantic similarity techniques, as applied to words or sentences. Section 3 is split into three subsections, each explaining the various proposed algorithms. The algorithms described in section 3.1 are used for the pre-processing and general framework; those in section 3.2 are for the dictionary-based technique, namely averaged maximum-translation similarity; and those in section 3.3 are for the MT-based techniques, namely noun-verb vector-based and term vector-based similarity algorithms. Section 4 presents the experimental design, including the tools and packages used in this study, the datasets involving short phrases from the human language understanding and the constructed benchmark dataset. Section 5 presents the results and discussion of findings and, finally, in section 6, conclusions and recommendations for future research are provided.

Related Research

Word Semantic Similarity Techniques

Semantic similarity, semantic distance, semantic relations, or more broadly semantic relatedness are all terms used interchangeably in the literature to describe the extent to which term *A* can be used to indicate or replace term *B*. Semantic features exploit terms with semantic relations-such as synonyms, antonyms, hyponyms and hypernyms (Solé-Ribalta, 2014; Liu *et al*., 2012; Luo *et al*., 2011)- or semantic dependencies (Li *et al*., 2006; Muftah, 2009).

HowNet is an online knowledge-based database that relates concepts and attributes of concepts. It is organized into a hierarchy in which each concept is described by a series of attributes called *sememes* (Dai *et al*., 2008). WordNet is a lexical database for English (Miller, 1995), which arranges words with the same meanings into groups called *synsets*. The words are then linked with more abstract concepts called *hypernyms* and more specific concepts called *hyponyms*. Some knowledge-based metrics are based on a single taxonomy or, more precisely, on the directed-acyclic graph, which demarcates the boundaries between two concepts in the taxonomy. These measures can be called mono-taxonomy metrics, as summarised in Table 1.

A mono-taxonomy metric was proposed to evaluate the Information Content (IC) of two concepts based solely on the HowNet taxonomy (Bin *et al*., 2012). Unlike the originally proposed IC measure which depends on WordNet and a corpus (Resnik, 1995), the IC metric in (Bin *et al*., 2012) was computed based on HowNet stating that concepts with many hyponyms convey less information than concepts located as the leaves, as follows:

$$IC(c) = 1 - \frac{\log(hypo(c)+1)}{\log(\max_{hn})} \tag{1}$$

where, *hypo* is the number of hyponyms of a given sememe and $\max_{hn}$ is the maximum number of sememes in the taxonomy. The similarity based on the modified IC measure was calculated as follows:

$$sim(c_1,c_2) = \max_{c\in(c_1,c_2)} IC(c) \tag{2}$$

Dai *et al.* (2008) proposed a semantic similarity measure between two concepts based on the semantic similarity of their primary sememes in their concept hierarchy. The primary sememe of a concept, e.g., *doctor*, is the top term that describes the concept in the tree, which is *human*, whereby other sememes in the tree such as *status* and *education* are modifiers of it. The similarity between two concepts in Dai *et al.* (2008) was computed using *li* metric (Li *et al*., 2006) and the number of common sememes between them, as follows:

$$\begin{gathered} sim(c_1,c_2) = \alpha.li(c_1,c_2) + \\ \beta\frac{\sum \max(li(c_1,c_2)}{|c_1|} + \gamma\frac{hypo(c_1,c_2)}{n} \end{gathered} \tag{3}$$

where, *n* is the total number of sememes for both concepts and *hypo* is the number of common sememes. Zhang *et al.* (2014a) proposed a word semantic similarity measure that combines features obtained from the HowNet taxonomy for both concepts. Zhang *et al.* (2014a) study combined four features from the tree that holds two concepts including the depth, width, density and overlap as they believed that the more features are considered, the more closeness to what humans perceive is obtained. The equation is expressed as follows:

$$\begin{gathered} sim(c_1,c_2) = \alpha.depth(c_1,c_2) + \beta.width(c_1,c_2) \\ +\gamma.density(c_1,c_2) + \lambda.overlap(c_1,c_2) \end{gathered} \tag{4}$$

where, α, β, γ and λ are scaling parameters.

There have been many other similarity measures proposed based on WordNet, including *path*, *lch* (Leacock and Chodorow, 1998), *wup* (Wu and Palmer, 1994), *res* (Resnik, 1995), *lin* (Lin, 1998), *jcn* (Jiang and Conrath, 1997). Meng *et al.* (2014) recent study suggested a new metric that combines information density and the *path* metric and Li et al. (2006) earlier study proposed a semantic similarity combining the shortest *path* between two words, w_1 and w_2 and the *depth* of their Least Common Subsumer (LCS) in the taxonomy containing both words. The new metric proposed by Meng *et al.* (2014) showed more accurate results and outperformed Li *et al.* (2006) in terms of the similarity coefficient because Meng *et al.* (2014) metric not only reflects the semantic density information but also the path information. The description and mathematical representation of several word semantic similarity metrics are shown in Table 1.

Table 1. List of Word Semantic Similarity Metrics

	Metric	Taxonomy	Description	Mathematical Equation
Mono-Taxonomy Word Similarity Metrics	HowNet-based IC metric (Bin *et al*., 2012)	HowNet	Shows the information content based on the number of hyponyms for a given sememe and the maximum number of sememes in a HowNet taxonomy.	$IC(c)=1-\frac{\log(hypo(c)+1)}{\log(\max_{hn})}$ $sim(c_1,c_2)=\max_{c\in(c_1,c_2)} IC(c)$
	HowNet-based modified *li* metric (Dai *et al*., 2008)	HowNet	Measures the semantic similarity of their primary sememes in their concept hierarchy computed using *li* metric ((Li *et al*., 2006) and the number of common sememes between them	$sim(c_1,c_2)=\alpha.li(c_1,c_2)$ $+\beta\frac{\sum\max(li(c_1,c_2)}{\lvert c_1\rvert}+\gamma\frac{hypo(c_1,c_2)}{n}$
	HowNet-based combined features (Zhang *et al*., 2014a)	HowNet	Combines features obtained from the HowNet taxonomy for both concepts	$sim(c_1,c_2)=\begin{matrix}\alpha.depth(c_1,c_2)+\beta.width(c_1,c_2)\\+\gamma.density(c_1,c_2)+\lambda.overlap(c_1,c_2)\end{matrix}$ where α, β, γ and λ are scaling parameters.
	path metric (Jiang and Conrath, 1997; Li *et al*., 2003)	WordNet	Measures the shortest path between two terms/concepts in the taxonomy	$path(c_1,c_2)=n$ where *n* is the number of edges that makes the shortest link between two concepts.
	lch metric (Leacock and Chodorow, 1998)	WordNet	measures the shortest path between two terms' synsets and the maximum depth from the root of the taxonomy.	$lch(c_1,c_2)=\log(\frac{path(c_1,c_2)}{2*maxdepth})$ where *maxdepth* is the longest distance between the root and any leaf in the taxonomy that contains both synsets.
	wup metric (Wu and Palmer, 1994)	WordNet	measures the depth of the terms' synsets in the taxonomy and the depth of their least common subsumer	$wup(c_1,c_2)=\frac{2\times depth(LCS(c_1,c_2))}{depth(c_1)+depth(c_2)}$
	Information content (IC) (Fernando and Stevenson, 2008)	WordNet	shows a measure that a concept can be found in a standard textual corpus.	$IC(c)=-\log(P(c))$ where *P*(*c*) is the probability that *c* can be found in the corpus. $ICSim(c_1,c_2)=\max_{c\in S(c_1,c_2)} IC(c)$ where S is the set of concepts that subsume both concepts.
	res metric (Resnik, 1995)	WordNet	computers a similarity score of two concept synsets based on the IC of their least common subsumer (LCS) in the taxonomy.	$res(c_1,c_2)=IC(LCS(c_1,c_2))$
	lin metric (Lin, 1998)	WordNet	based on *res* metric and IC of the words' synsets	$lin(c_1,c_2)=\frac{2*IC(LCS((c_1,c_2))}{IC(c_1)+IC(c_2)}$
	jcn metric (Jiang and Conrath, 1997)	WordNet	based on the IC of the LCS and that of the words' synsets	$jcn(c_1,c_2)=1$ $-\frac{IC(c_1)+IC(c_2)-2*IC(LCS(c_1,c_2))}{2}$
	meng metric (Meng *et al*., 2014)	WordNet	combines the IC metric with the path metric for both concepts.	$meng(c_1,c_2)=\frac{2*IC(LCS(c_1,c_2))^{(\frac{1-e^{-k*path(c_1,c_2)}}{e^{-k*path(c_1,c_2)}})}}{IC(c_1)+IC(c_2)}$ where k is adapted parameter between 0 and 1
	li metric (Li *et al*., 2006)	WordNet	combines the shortest *path* between two words w_1 and w_2 and the *depth* of the their LCS in the taxonomy that has both words	$li(w_1,w_2)=e^{-\alpha.path(w_1,w_2)}$ $\times\frac{e^{\beta.depth(LCS(w_1,w_2))}+e^{\beta.depth(LCS(w_1,w_2))}}{e^{\beta.depth(LCS(w_1,w_2))}-e^{\beta.depth(LCS(w_1,w_2))}}$ where α∈[0,1] and β∈[0,1], are scaling parameters of the contribution of the *path* and *depth* metrics in the formula.
Across Taxonomies	*lesk* metric (Banerjee and Pedersen, 2003)	multiple	incorporates information from the directions between the lexical chains of two word synsets.	-
	hso metric (Hirst and St Onge, 1998)	multiple	measures the relationship of two words' synsets based on the overlap of their dictionary glosses.	-

Cross-taxonomy metrics, on the other hand, use multiple knowledge-based taxonomies and may work across domains. Examples of these metrics include the *lesk* (Banerjee and Pedersen, 2003) and *hso* (Hirst and St Onge, 1998) metrics, which measure semantic relatedness rather than similarity (Corley and Mihalcea, 2005; Budanitsky and Hirst, 2006). Statistical methods for semantic similarity, such as Latent Semantic Analysis (LSA) (Landauer *et al*., 1998) and Pointwise Mutual Information (PMI) (Turney, 2001), have been derived from large text corpora.

Ontology-based semantic similarity measures have been proposed in recent research studies (Jian-Bo *et al.*, 2013; Sánchez *et al.*, 2012; Ye and Zhan-Lin, 2010). Ontologies are constructed for several domains to structure the concepts in a way that supports logical reasoning and semantic information. One of the ontology-based semantic similarity metrics was denoted as follows (Ye and Zhan-Lin, 2010):

$$sim(c_1,c_2)=\frac{C_1\cap C_2}{(C_1\cap C_2)+\alpha(C_1-C_2)+\beta(C_2-C_1)} \tag{5}$$

where, C_1 and C_2 are the set of abstract concepts for the terms c_1 and c_2, respectively. Jian-Bo *et al.* (2013) recommended the development of an ontology-based measurement that combines a graph-based approach with features extracted from the ontology containing both concepts. The semantic similarity of words was studied based on multiple dictionaries (Zhang *et al.*, 2014b). In addition, the degree of commonality between concepts belonging to multiple ontologies has been used to modify the IC semantic similarity of concept pairs across ontologies (Solé-Ribalta, 2014; Batet *et al.*, 2014). Further, multiple trees were constructed from taxonomic relations among entities in an ontology and a multi-tree concept semantic similarity measure was proposed based on the following: (i) Combined tree of features, (ii) updated weights for the nodes in the combined tree and (iii) the premise that the similarity of two concepts is basically the weight of the root in the combined tree that has both entities (Hajian and White, 2011).

Short-text Semantic Similarity Techniques

Because natural language understanding requires more than the semantic similarity of words, several research studies have investigated short-text semantic similarity based on the semantic relations of their words (Li *et al.*, 2006; O'Shea *et al.*, 2008; Lee, 2011; Corley and Mihalcea, 2005; Mihalcea *et al.*, 2006). A survey of studies that have semantically evaluated textual elements is presented below.

Two candidate texts, T_1 and T_2, can be represented by concept vectors and the similarity between them can be evaluated accordingly using Cosine, Jaccard, Dice, or any similarity coefficient (Alzahrani *et al.*, 2012). For example, the texts can be represented by a binary vector with two entries: 1 if the concept is in the joint word matrix and 0, otherwise, where the joint word matrix W consists of distinct words in both texts (Fernando and Stevenson, 2008). The similarity score was computed as the mathematical product of the binary vectors and the similarity matrix was as follows:

$$Sim(T_1,T_2)=\frac{\vec{T}_1W\vec{T}_2}{|\vec{T}_1||\vec{T}_2|} \tag{6}$$

where, $\vec{T}_1$ and $\vec{T}_2$ are the binary vectors of texts T_1 and T_2.

Apart from using binary vectors in the previous study, an earlier study suggested to use the Inverse Document Frequency (IDF) measure combined with a local similarity metric, implemented by any of the word similarity measures (Corley and Mihalcea, 2005; Mihalcea *et al.*, 2006). The semantic similarity of the two texts was derived, as in Equation 7 from the maximum similarity gained by a word w from T_1 and words in T_2, referred to as *maxSim*(w, T_2) and *idf*(w) obtained from the relation n_w/N, where n_w is the number of documents that contain the word w and N is the total number of documents in a large text corpus:

$$Sim(T_1,T_2)=\frac{1}{2}\left(\frac{\sum_{w\in T_1} maxSim(w,T_2)\times idf(w)}{\sum_{w\in T_1} idf(w)} + \frac{\sum_{w\in T_2} maxSim(w,T_1)\times idf(w)}{\sum_{w\in T_2} idf(w)}\right) \tag{7}$$

Lee (2011) reported a short-text similarity measure that was computed based on the nouns and verbs because it was believed that the semantic similarity should be obtained in a fast but accurate way. Lee's study implemented a Noun Vector (NV) containing a joint noun set from two candidate texts, T_1 and T_2 and a Verb Vector (VV) containing a joint verb set from T_1 and T_2. The value of an entry in the NV vector (and VV vector, respectively) was defined as the highest *wup* similarity (Wu and Palmer, 1994) found between the corresponding noun and other nouns in the NV vector (and the corresponding verb and other verbs in the VV vector, respectively). The similarity score between the two texts, the noun vector similarity S_N and the verb vector similarity S_V were integrated as follows:

$$Sim(T_1,T_2)=\delta\cdot S_N(T_1,T_2)+(1-\delta)\cdot S_V(T_1,T_2) \tag{8}$$

where, δ is a scaling parameter ∈ [0.5,1] and both vectors are computed as the cosine similarity between the noun vectors and the verb vectors from T_1 and T_2, respectively.

A study by Li *et al.* (2006) proposed a semantic similarity measure between sentences derived from a semantic similarity and an order similarity as follows:

$$Sim(T_1,T_2)=\delta\cdot S_s(T_1,T_2)+(1-\delta)\cdot S_r(T_1,T_2) \tag{9}$$

where, S_s is the semantic similarity metric and S_r is the order similarity metric. S_s is computed as the Cosine similarity of the two vectors:

$$S_s(T_1,T_2)=\frac{s_1\cdot s_2}{\|s_1\|\cdot\|s_2\|} \tag{10}$$

The values of an entry in the semantic vector s_1 for text T_1 and s_2 for text T_2 are defined below:

$$s_1\left(w_i\right) = li\left(w_i, \tilde{w}\right) \times IC\left(w_i\right) \times IC\left(\tilde{w}\right) \tag{11}$$

$$s_2\left(w_j\right) = li\left(w_j, \tilde{w}\right) \times IC\left(w_j\right) \times IC(\tilde{w}) \tag{12}$$

where, the *li* metric is the highest word semantic similarity between the word w_i and any word in the candidate text and *IC* is the information content measure. Further, the order similarity, S_r, means that different word orders may convey different meanings and should be counted into the semantic similarity. The word order vectors from T_1 and T_2 can be given as r_1 and r_2, respectively. The cosine similarity was obtained from the order vectors as shown below:

$$S_r(T_1, T_2) = 1 - \frac{\| r_1 - r_2 \|}{\| r_1 + r_2 \|} \tag{13}$$

The above short-text similarity methods have been implemented on mono-language texts and compared thoroughly using the English sentence pairs benchmark, as reported in many research studies (Lee, 2011; Islam and Inkpen, 2008). The performance of the reported methods is not directly comparable due to different evaluation metrics and datasets. However from these studies, we can see that methods that take more textual features into consideration, such as Li *et al.* (2006) combined metric of semantic and order similarities, achieved high correlation coefficient with the human rating but with the cost of slow performance. On the other hand, methods that reduced the textual features, such as Lee (2011) combined metric of noun and verb vector similarities, may have less correlation coefficient with the human intuition but with the advantage of faster computation.

Following to the aforementioned studies, an ongoing series of computational semantic evaluation called SemEval have embodied several systems and methods in 2012, 2013, 2014 and 2015. Several tasks for semantic relatedness evaluation and paraphrase detection on various datasets have been investigated (Agirre *et al*., 2013; Agirre *et al*., 2012; 2014) Rich semantic analysis on new datasets has been conducted such as using Twitter data (Xu *et al*., 2015) or by featuring interpretability (Agirre *et al*., 2015). Generally these studies employ variety of NLP tools including lemmatizers, POS taggers, word sense disambiguation and syntax features. To obtain the similarity score, the methods differ in featuring the semantic notions amongst words and phrases in the candidate texts. Many studies employed WordNet lexical database and its semantic similarity measures, while others used Wikipedia knowledge base. Several studies investigated the use of semantic role labelling, distributional thesaurus and dictionaries, Machine Translation (MT) and machine learning algorithms. None of these tasks tackles cross-language semantic similarity and it may be forthcoming in 2016. Although SemEval 2014's task 10 was entitled "multilingual" semantic textual similarity (Agirre *et al*., 2014), it was separated into English subtask and Spanish subtask whereby each subtask was evaluated via monolingual datasets.

Cross-Language Semantic Similarity Techniques

Owing to the substantial increase in text data available in multiple languages, there has been a lot of research recently investigating semantic similarity measures across languages (Zou *et al*., 2013; Dai *et al*., 2008; Stoyanova *et al*., 2013; Vulic and Moens, 2014; 2013; Dai and Huang, 2011). Dai and Huang's study (Dai and Huang, 2011), for example, tested the effectiveness of a word semantic similarity measure for applications in the cross-language domain. They computed the similarities between words using an algorithm based on the Chinese–English HowNet. Their results showed a strong positive correlation with the humans' judgements, suggesting it would be a robust measure for use in cross-language applications. Additional studies (Vulic and Moens, 2014; 2013) proposed approaches that identified similar words across languages. Two words in different languages are similar if they generate similar words as their top semantic word responses. Semantic word responding is a cognitive science term indicating the terms that humans associate with a certain cue word. A study conducted by Wu *et al.* (2010) explored how to generate semantic classes of verbs across languages using parallel corpora.

Methods for cross-language identification of semantic relations have been proposed recently. One example is Stoyanova *et al.* (2013), which combined word semantic similarity measurements with the morphology and semantic relations obtained from WordNet. An automatic classifier was trained on parallel and comparable English-Bulgarian texts to perform semantic relations labelling and reduce word sense ambiguities. Zou *et al.* (2013) proposed a method that captures both mono and cross-lingual semantic relations across different languages. The method they proposed stored the bilingual embeddings between Chinese and English from a large unlabelled corpus while utilizing MT to align words with the same meanings.

Complementary to explicit semantic analysis which uses Wikipedia as a knowledge base, cross-language explicit semantic analysis CL-ESA have gained popularity in recent years (Sorg and Cimiano, 2010; Anderka *et al*., 2009) for computing semantic relatedness between words from different languages. Generally CL-

ESA works by mapping both of the query *q* and the document collection *d* into a multilingual concept space. In (Anderka *et al*., 2009), the mathematical representation for CL-ESA was simplified as follows:

$$f_{CL-ESA}(q_j, d_i) = q_j^T \cdot G_{j,i} \cdot d_i \qquad (14)$$

where, q_j^T is the matrix transpose of q_j and $G_{j.i} = A_{D_j^*} \cdot A^T_{D_i^*}$ is the mathematical product representing term-document matrices from the query q_j and candidate indexed document from Wikipedia D^*.

A contribution by Navigli and Ponzetto (2012a) proposed a multilingual semantic similarity approach that used BabelNet; a knowledge-rich lexicon and semantic database which supports multiple languages (Navigli and Ponzetto, 2012a). The proposed approach works by intersecting the semantic graphs from different languages into one core graph and computing the semantic similarity score based on the core graph. Given that w_1 and w_2 are two words from two languages l_1 and l_2, respectively and G_{joint} is the core graph formed by the intersection between graphs generated from BabelNet between w_1 and w_2 in *L* different languages, then the semantic relatedness between these two words was computed by Navigli and Ponzetto (2012a) as follows:

$$Sim(w_1, w_2) = \max_{\substack{s_1 \in Senses(w_1),\\ s_2 \in Senses(w_2)}} score(G, s_1, s_2) \qquad (15)$$

where, s_1 and s_2 are the different senses for w_1 and w_2, respectively and *G* is the graph that holds each two senses. The similarity *score* was computed as follows:

$$score(s_1, s_2) = \max_{p \in paths(G, s_1, s_2)} \frac{1}{length(p)} \qquad (16)$$

where, *paths* is the set of all possible paths between s_1 and s_2 in sub graph *G* and *length* is the number of nodes in a path *p*. The method obtained competitive results compared with traditional monolingual and multilingual measures. Though it works on words, it can be expanded in various ways to compute the semantic similarity of cross-language texts beyond the words.

Proposed Cross-Language Semantic-Similarity

General Framework

The pre-processing algorithm was divided into two parts: One for the English text and one for the Arabic candidate, as shown in Fig. 1.

For the English text, the pre-processing steps included: (i) Tokenization whereby the text was divided into word tokens referred to as [W]; (ii) Part-Of-Speech (POS) disambiguation, or in other words, each token was designated a POS tag, namely noun, verb, adjective and adverb referred to as [N], [V], [AJ], [AV], respectively; and (iii) removal of the stop words such as prepositions and articles. (iv) Then. for each word token w_i in *A*, we found the set of lemmas λ_i, knowing the corresponding POS tag for that word (note that in many cases there is one lemma for a word token), as follows:

$$\begin{aligned} &\forall w_i \in A \rightarrow t_i = \{\lambda_{i,1}, \lambda_{i,2}, \lambda_{i,3}, \ldots, \lambda_{i,x}\}: \\ &POS_i \in \{N, V, AJ, AV\} \end{aligned} \qquad (17)$$

where, *x* is the number of different lemma forms that can be found for the word using WordNet. A *term set* was constructed from the English sentence *A* as the union of sets t_1, t_2,..,t_n, i.e. $T_1 = \bigcup_{i=1}^{n} t_i$, where *n* is the total unique terms after removing the stop words.

The same processing steps were applied to the Arabic text, *B*, with the addition of the translation. First, *B* was split into word tokens. POS tagging was applied and the most frequent Arabic words were removed. Each word token referred to as w_i was reduced to its lemma (Roth *et al*., 2008), as below:

$$\forall w_j \in B \rightarrow l_j : POS_j \in \{N, V, AJ, AV\} \qquad (18)$$

Knowing the lemma of each Arabic word as well as its POS tag, an Arabic-to-English dictionary translation was applied to obtain possible senses (i.e., meanings) for that word in English, as follows:

$$\forall l_j \rightarrow t_j = \{\phi_{j,1}, \phi_{i,2}, \phi_{j,3}, \ldots, \phi_{j,x}\} : POS_j \in \{N, V, AJ, AV\} \qquad (19)$$

As a final step in the pre-processing of the Arabic text, the *translation term set* was constructed from the Arabic sentence as $TT_2 = \bigcup_{j=1}^{m} t_j$, where *m* is the number of unique terms and t_j is the translation subset of lemma l_j.

Figure 2 shows the general framework for this study. After the input texts *A* and *B* were pre-processed, we employed three different algorithms.

Following to the dictionary translation technique, we proposed an averaged maximum-translation similarity algorithm between the *term set*, referred to as T_1 (obtained from the English text) and the *translated term set,* referred to as TT_2, to estimate the cross-lingual semantic similarity. The semantic similarity score between the terms was then correlated and averaged as proposed by Yerra and Ng (2005).

Following to the MT technique, we obtained an English version of T_2 and then a *term set,* denoted as T_2, was constructed in the same way for the English text. In this path, we used two vector-based semantic similarity algorithms proposed for mono-lingual sentences. One was based on the combined similarity between the noun and verb vectors obtained from both texts, which was proposed by Lee (2011) and the other was based on the semantic similarity of term vectors, which was suggested by Li *et al.* (2006).

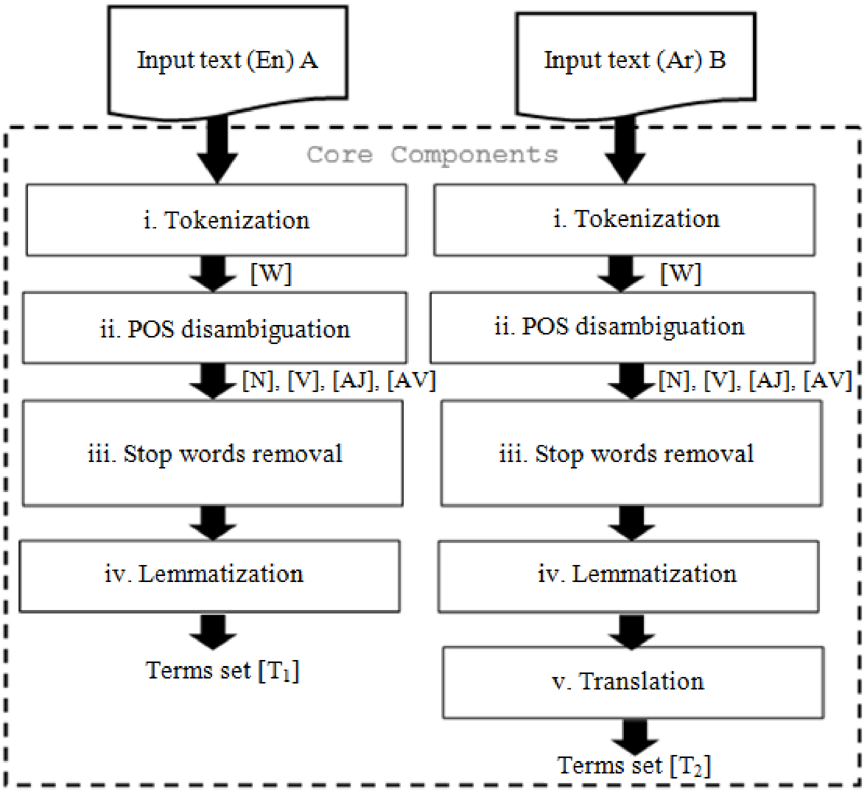


Fig. 1. Pre-processing steps of the input texts

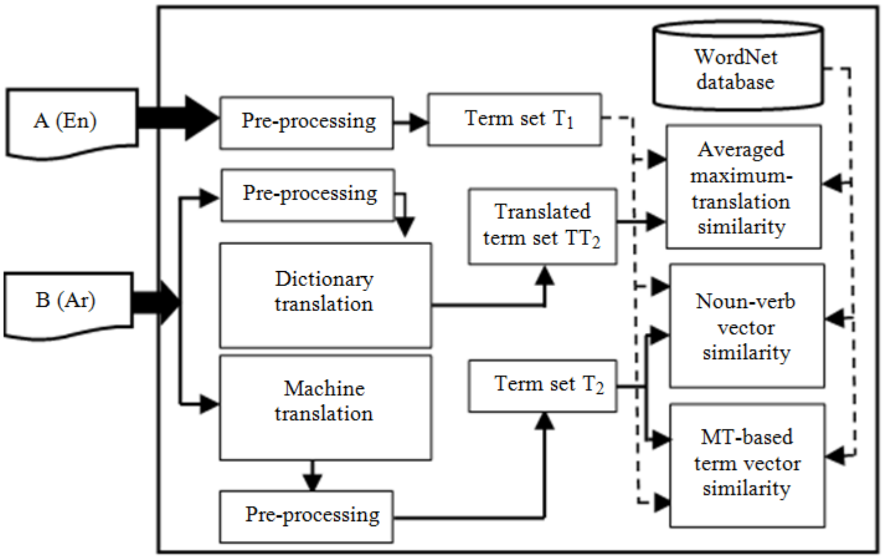


Fig. 2. General framework of the proposed algorithms

Dictionary-Based Technique

This method was proposed and implemented as a copy detection approach (Yerra and Ng, 2005). The algorithm uses two input vectors, namely T_1 and TT_2, constructed from *A* and *B*, respectively.

Algorithm 1: Averaged Maximum-Translation Similarity

Step 1: Each term in *B* was correlated with the terms in the English text *A* using the following function:

$$f(A,t_j) = 1 - \left(\prod_{i=1}^{n}(1 - MaxSim(t_i,t_j)\right): \\ t_i \in T_1, t_j \in TT_2, i = 1,2,...,n, j = 1,2,...,m \tag{20}$$

where, T_1 and TT_2 are the representative term vectors of *A* and *B*, respectively and Π is the product function. *MaxSim* refers to the maximum word semantic similarity obtained between the term t_i and the translated terms $t_j = \{\phi_{j,1},\phi_{i,2},\phi_{j,3},...,\phi_{j,x}\}$:

$$MaxSim(t_i,t_i) = \\ \max(wup(t_i,\phi_{j,1}), wup(t_i,\phi_{j,2}),..., wup(t_i,\phi_{j,x})) \tag{21}$$

where, *wup* metric (Wu and Palmer, 1994) is one of the knowledge-based semantic similarity measures between two terms c_1 and c_2 found useful in our previous work (Alzahrani *et al*., 2015), as follows:

$$wup(c_1,c_2) = \frac{2 \times depth(LCS(c_1,c_2))}{depth(c_1) + depth(c_2)} \tag{22}$$

Step 2: The degree of similarity between the candidate texts was computed as the averaged summation using the equation:

$$Sim(A,B) = \sum_{j=1}^{m} f(A,t_j) / m \tag{23}$$

Machine Translation-Based Techniques

Machine translation techniques have improved over recent years and have become, in some languages, almost as accurate as human translation for short phrases and sentences. When the Arabic candidate text is translated into English as a pre-processing step, the problem is shifted into a mono-lingual sentence semantic similarity problem. We decided to use the vector similarity methods proposed in earlier research studies and obtained promising results (Li *et al*., 2006; Lee, 2011).

Algorithm 2: Noun-Verb Vector Based Similarity

In this algorithm, we employed MT followed by the mono-lingual semantic similarity method (Lee, 2011). The algorithm was based on the combined similarity between the noun and verb vectors from the two texts. The following steps were implemented:

Step 1: *B* was translated into English using Google Translate API.

Step 2: The translated text was pre-processed in the same way as for the English text.

Step 3: The *term vector* was constructed from *B* as $T_2 = \{t_1,t_2,t_3,...,t_m\}$, where *m* is the total number of unique terms.

Step 4: A joint noun set from the two candidate texts, *A* and *B*, was constructed as the Noun Vector (NV), where the value of an entry was defined as the maximum word semantic similarity found between the corresponding noun and other nouns in the *NV* vector, as follows:

$$NV = | T_1 \cup T_2 | = | t_1,t_2,...,t_x |: x \in [1,n+m], POS = N \tag{24}$$

$$NV_{T_1} = | MaxSim(t_1,t_i), MaxSim(t_2,t_i),..., \\ MaxSim(t_x,t_i) |: t_i \in T_1, i = 1,2,...,n \tag{25}$$

$$NV_{T_2} = | \max sim(t_1,t_j), \max sim(t_2,t_j),..., \\ \max sim(t_x,t_j) |: t_j \in T_2, j = 1,2,...,m \tag{26}$$

where, *MaxSim* is the maximum semantic similarity as described in algorithm 1.

Step 5: Similarly, a joint verb set from *A* and *B* was constructed, namely *VV* and then the *verb vectors* containing the maximum *wup* similarity were obtained between each verb and other verbs in the *VV* vector, as below:

$$VV = | T_1 \cup T_2 | = | t_1,t_2,...,t_y |: y \in [1,n+m], POS = V \tag{27}$$

$$VV_{T_1} = | \max sim(t_1,t_i), \max sim(t_2,t_i),..., \\ \max sim(t_y,t_i) |: t_i \in T_1, i = 1,2,...,n \tag{28}$$

$$VV_{T_2} = | \max sim(t_1,t_j), \max sim(t_2,t_j),..., \\ \max sim(t_y,t_j) |: t_j \in T_2, j = 1,2,...,m \tag{29}$$

Step 6: The Cosine similarity values between the noun and verb vectors were computed as follows:

$$Sim_N(T_1,T_2) = Cos(NV_{T_1}, NV_{T_2}) = \frac{NV_{T_1} \cdot NV_{T_2}}{\| NV_{T_1} \| \cdot \| NV_{T_2} \|} \tag{30}$$

$$Sim_V(T_1,T_2) = Cos(VV_{T_1}, VV_{T_2}) = \frac{VV_{T_1} \cdot VV_{T_2}}{\| VV_{T_1} \| \cdot \| VV_{T_2} \|} \tag{31}$$

Step 7: The similarity score was computed based on the noun vector similarity *Sim*$_N$ and the verb vector similarity *Sim*$_V$:

$$Sim(A,B) = \delta \cdot Sim_N(T_1,T_2) + (1-\delta) \cdot Sim_V(T_1,T_2) \tag{32}$$

where, δ is a scaling parameter ∈ [0.5,1].

Algorithm 3: MT-based Term Vector Based Similarity

The algorithm was based on the following steps:

Step 1: *B* was translated into English using Google Translate.
Step 2: The translated text was pre-processed as in the previous algorithm and the *term vector T*$_2$ was constructed.
Step 3: A joint term set from the two candidate texts, *A* and *B*, was constructed and referred to as a *term vector TV*, where the value of an entry was defined as the maximum word semantic similarity found between the corresponding term and other terms in the candidate text, as follows:

$$TV = |T_1 \cup T_2| = |t_1, t_2, \ldots, t_x| : x \in [1, n+m], POS = any \tag{33}$$

$$TV_{T_1} = |MaxSim(t_1,t_i), \ldots, MaxSim(t_x,t_i)| : t_i \in T_1, i = 1,..,n \tag{34}$$

$$\begin{aligned} TV_{T_2} = |&MaxSim(t_1,t_j), \ldots, \\ &MaxSim(t_x,t_j)| : t_j \in T_2, j = 1, \ldots, m \end{aligned} \tag{35}$$

where, *li* similarity metric (Li *et al*., 2006) was used in this algorithm to find the *MaxSim* between any two terms.

Step 4: The Cosine similarity values between the term vectors were computed as follows:

$$Sim(A,B) = Cos(TV_{T_1}, TV_{T_2}) = \frac{TV_{T_1} \cdot TV_{T_2}}{\| TV_{T_1} \| \cdot \| TV_{T_2} \|} \tag{36}$$

Experimental Design

Tools and Packages

For the pre-processing of the English and Arabic input texts, we employed the Stanford NLP tools (Toutanova *et al*., 2003; Monroe *et al*., 2014). We also used the NLTK (Edward and Steven, 2002) for various tasks including the computation of WordNet-based semantic similarity metrics.

Datasets

To evaluate the proposed methods, we used sentence pairs annotated with ground-truth human similarity scores. Each pair consists of one element of an English sentence and one element of an Arabic sentence, which may be similar (or dissimilar) to the English sentence in some degree. For our initial investigation, selected sentences from books on natural language understanding with similarity scores close to humans' similarity intuition were used (Li *et al*., 2006) (Section 4.2.1). Moreover, cross-language similarity benchmark was constructed to evaluate the proposed techniques (Section 4.2.2).

Selected NLP Sentences

In our initial investigation, the sample of sentence pairs were used as follows: (i) The second sentence in each pair was translated into Arabic by a native speaker of Arabic, educated to graduate level and fluent in English as a second language; (ii) The translations were validated (and in some cases, modified) by two language experts, who speak both languages; (iii) A number of pairs from the sample proposed by Li *et al.* (2006) were excluded because they are too short and the remaining ten pairs were included.

Table 2 shows the original sentence pairs by Li *et al.* (2006) and the proposed translation for the second pair. We assumed that the validity of using the same similarity scores in the English pairs would hold for the Arabic-English pairs because of the short translations given for each sentence (which do not exceed five words for each pair). Besides, the similarity scores obtained by Li *et al.* (2006) have been proven to be fairly consistent with human intuition.

Table 2. Raw sentences of short lengths based on natural language understanding

Sentence pairs		En-Ar pairs	Sentence pairs		En-Ar pairs
1	I have a pen. Where do you live?	I have a pen. أين تسكن؟	6	I have a hammer. Take some nails.	I have a hammer. خذ بعض المسامير.
2	John is very nice. Is John very nice?	John is very nice. هل جون شخص لطيف؟	7	Canis familiaris are animals. Dogs are common pets.	Canis familiaris are animals. الكلاب حيوانات أليفة.
3	It is a dog. That must be your dog.	It is a dog. يجب أن يكون هذا هو كلبك.	8	I have a pen. Where is ink?	I have a pen. أين الحبر؟
4	Dogs are animals. They are common pets.	Dogs are animals. هي حيوانات أليفة.	9	It is a glass of cider. It is a full cup of apple juice.	It is a glass of cider. هذا كوب ممتليء بعصير التفاح.
5	It is a dog. It is a log.	It is a dog. إنه سجل.	10	I have a hammer. Take some apples.	I have a hammer. خذ بعض التفاح.

Original sentence pairs (Li *et al*., 2006) and translated 2nd-item pair into Arabic in this study.

Pilot Cross-Language Human Similarity Benchmark Dataset

In order to evaluate the cross-language semantic similarity in this study, human ratings were collected on a proposed dataset according to the existing designs of semantic similarity benchmarks. The rating participants were selected from among a population of Arabic mother tongue speakers of English as a second language. They were all educated to postgraduate level and all had an upper-intermediate to professional understanding level of English. They were either academics or postgraduate students in English universities.

A. Materials

A group of sixty-five English noun pairs, which have been proven to be fairly consistent in terms of human semantic similarity ratings, were proposed in the literature (Rubenstein and Goodenough, 1965). The definitions of these noun pairs, taken from Collins Cobuild dictionary, were semantically rated by thirty-two human participants in O'Shea *et al*. (2008). A subset consisting of thirty sentence pairs were selected in order to distribute the rated similarities evenly across the similarity ranges (Li *et al*., 2006).

In the present study, we proposed a benchmark dataset that made use of this standard dataset but with the second item in each pair replaced by its Arabic translation. The following procedure was used: (i) The second sentence was translated using three methods, namely MT, Human Translation (HT) and the Dictionary Definition (DD) of the original noun pair from a selected Arabic-Arabic dictionary; (ii) The original English sentence and the three translations were tabulated; (iii) To avoid researcher bias, three language experts, educated to PhD level, were asked to choose the most optimal Arabic translation for the English sentence; (iv) Using a majority vote procedure, the translation that indicated the most similar semantic content with no additional phrases was then tabulated with the original English sentence. This table was given to a further two participants to check whether any amendments were needed for each of the Arabic translations. Figure 3 shows a sample of the questionnaire that was distributed for participants to choose the optimal translation and Table 3 shows the majority voting results.

B. Procedure

To rate the similarity of constructed Arabic-English cross-language sentence pairs, seventeen participants were asked to complete a questionnaire, shown in Fig. 4. The participants were all native speakers of Arabic with upper-intermediate to professional proficiency in English as a second language. All of the participants were educated to graduate level or above. The procedure to obtain the human similarity scores is detailed as follows: (i) The order of the sentence pairs was randomized and given to new participants to avoid any evaluation bias; (ii) Following the same rating scale of standard semantic similarity datasets (Rubenstein and Goodenough, 1965), the participants were instructed to rate each sentence pair on a scale from 0.0 to 4.0. A rubric was provided to explain the evaluation scale, where 0 indicated that the two sentences are totally different/dissimilar in their meaning and 4 indicated that the sentences are identical.

Table 4 shows the cross-language sentence pairs used in this study with the mean similarity and the standard deviation for the human rating.

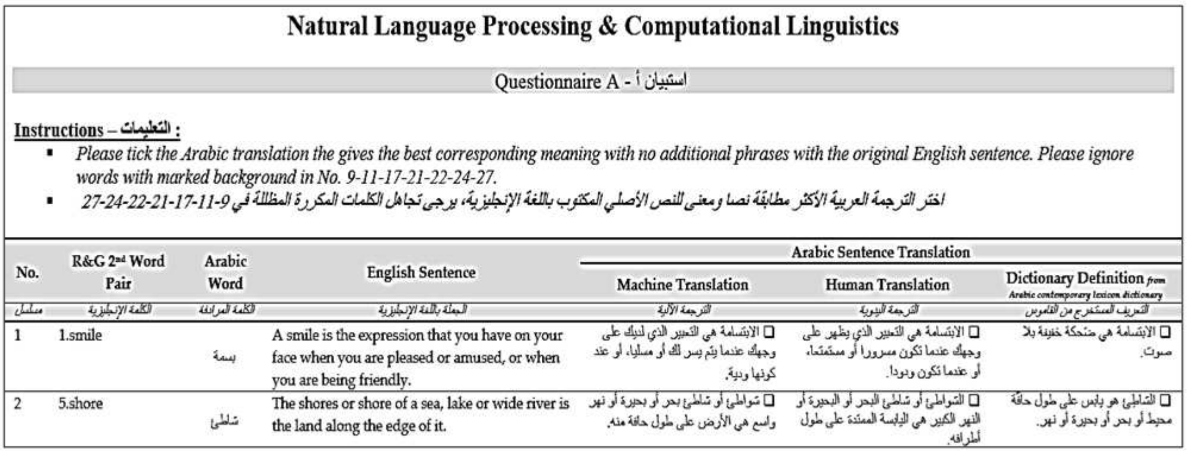
Table 3. Majority voting results: Machine Translation (MT), Human Translation (HT), Dictionary Definition (DD)

		Participant's Selection						Participant's Selection			
	R&G 2nd	---------------------------			Majority		R&G 2nd	-----------------------------			Majority
No.	Word Pair	(1)	(2)	(3)	Voting	No.	Word Pair	(1)	(2)	(3)	Voting
1	1.smile	HT	HT	HT	HT	15	50.string	MT	HT	HT	HT
2	5.shore	DD	HT	HT	HT	16	51.tumbler	HT	HT	HT	HT
3	9.fruit	HT	HT	DD	HT	17	52.smile	HT	HT	HT	HT
4	13.rooster	DD	HT	DD	DD	18	53.slave	HT	HT	HT	HT
5	17.forest	HT	HT	HT	HT	19	54.voyage	MT	MT	MT	MT
6	21.sage	DD	HT	DD	DD	20	55.signature	HT	HT	DD	HT
7	25.graveyard	HT	HT	HT	HT	21	56.shore	DD	HT	HT	HT
8	29.woodland	HT	HT	HT	HT	22	57.woodland	HT	HT	HT	HT
9	33.woodland	HT	HT	HT	HT	23	58.tool	HT	HT	DD	HT
10	37.oracle	HT	HT	HT	HT	24	59.rooster	DD	HT	DD	DD
11	41.sage	DD	HT	DD	DD	25	60.lad	HT	HT	DD	HT
12	47.stove	HT	HT	HT	HT	26	61.pillow	HT	HT	DD	HT
13	48.wizard	HT	HT	HT	HT	27	62. graveyard	HT	HT	HT	HT
14	49.mound	DD	HT	HT	HT	28	63.car	DD	HT	DD	DD

Table 4. Arabic-English cross-language standard benchmark

R&G Pair	En-Ar Pair	Arabic-English Cross-Language Sentences	Mean	STD
1. Cord: Smile	cord: بسمة	Cord is strong, thick string. الابتسامة هي التعبير الذي يظهر على وجهك عندما تكون مسرورا أو مستمتعا، أو عندما تكون ودودا.	0.12	0.08
5. Autograph: Shore	autograph: شاطئ	An autograph is the signature of someone famous which is specially written for a fan to keep. الشواطئ أو شاطئ البحر أو البحيرة أو النهر الكبير هي اليابسة الممتدة على طول أطرافه.	0.00	0
9. Asylum: Fruit	asylum: فاكهة	An Asylum is a psychiatric hospital. الفاكهة هي الشيء الذي ينبت على الشجرة أو الشجيرة ويحتوي على البذور أو قد يحتوي على اللب المغطى بالمادة التي يمكن أكلها.	0.06	0.06
13. Boy: Rooster	boy: ديك	A boy is a child who will grow up to be a man. دِّيكُ هو ذَكَرُ الدَّجاجِ.	0.56	0.15
17. Coast: Forest	coast: غابة	The coast is an area of land that is next to the sea. الغابة هي مساحة كبيرة تنمو فيها الأشجار بجوار بعضها البعض.	0.94	0.25
21. Boy: Sage	boy: حكيم	A boy is a child who will grow up to be a man. الحكيم هو مَنْ تصدر أعمالُه وأقوالُه عن رويّة سديدة ورأي سليم، صاحب حكمة، متقن للأمور.	0.59	0.23
25. Forest: Graveyard	forest: مقبرة	A forest is a large area where trees grow close together. المقبرة هي مساحة من الأرض تقع أحيانا بالقرب من الكنيسة حيث يتم دفن الموتى فيها.	0.88	0.19
29. Bird: Woodland	bird:غابة	A bird is a creature with feathers and wings, females lay eggs and most birds can fly. هي الأرض الممتلئة بالأشجار. الغابة	0.35	0.15
33. Hill: Woodland	hill: غابة	A hill is an area of land that is higher than the land that surrounds it. هي الأرض الممتلئة بالأشجار. الغابة	0.88	0.17
37. Magician: Oracle	magician: عرَّاف	A magician is a person who entertains people by doing magic tricks. العراف في العصور القديمة هو الكاهن أو الكاهنة الذي يعطي تنبؤات بأحداث المستقبل أو عن الحقائق.	2.06	0.32
41. Oracle: Sage	oracle: حكيم	In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth. الحكيم هو مَنْ تصدر أعمالُه وأقوالُه عن رويّة سديدة ورأي سليم، صاحب حكمة، متقن للأمور.	1.00	0.25
47. Furnace: Stove	furnace:موقد	A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam. الموقد هو قطعة من العتاد التي توفر الحرارة وتستخدم للطبخ أو تدفئة الغرفة.	2.88	0.19
48. Magician: Wizard	magician:ساحر	A magician is a person who entertains people by doing magic tricks. في الأساطير والقصص الخيالية، الساحر هو الرجل الذي يملك قوى سحرية.	3.05	0.29
49. Hill: Mound	hill: كومة	A hill is an area of land that is higher than the land that surrounds it. الكومة من الشيء هي مقدار كبير مستدير منه.	1.76	0.33
50. Cord: String	cord: سلسلة	Cord is strong, thick string. السلسلة هي حبل رفيع مصنوع من خيوط ملتوية وتستخدم لربط الأشياء مع بعضها أو لربط الطرود.	2.53	0.35
51. Glass: Tumbler	glass: قدح	Glass is a hard transparent substance that is used to make things such as windows and bottles. القدح هو إناء للشرب له أطراف مستقيمة.	1.41	0.25
52. Grin: Smile	grin: ابتسامة	A grin is a broad smile. الابتسامة هي التعبير الذي يظهر على وجهك عندما تكون مسرورا أو مستمتعا، أو عندما تكون ودودا.	2.00	0.29
53. Serf: Slave	serf: عبد	In former times, serfs were a class of people who had to work on a particular person's land and could not leave without that person's permission. العبد هو الشخص المملوك لشخص آخر وعليه أن يعمل لذلك الشخص.	3.29	0.28
54. Journey: Voyage	journey:رحلة	When you make a journey, you travel from one place to another. .رحلة هي رحلة طويلة على متن سفينة أو مركبة فضائية	2.29	0.25
55. Autograph: Signature	autograph: توقيع	An autograph is the signature of someone famous which is specially written for a fan to keep. التوقيع هو اسمك مكتوب بطريقة مميزة خاصة بك وغالبا في نهاية المستند ويشير إلى أنك كتبت المستند أو أنك موافق على ماجاء فيه.	2.53	0.25
56. coast: shore	coast: شاطئ	The coast is an area of land that is next to the sea. الشواطئ أو شاطئ البحر أو البحيرة أو النهر الكبير هي اليابسة الممتدة على طول أطرافه.	3.41	0.22
57. forest: Woodland	forest: غابة	A forest is a large area where trees grow close together. هي الأرض الممتلئة بالأشجار. الغابة	3.88	0.08
58. Implement: Tool	implement: أداة	An implement is a tool or other piece of equipment. .الأداة هي أي آلة أو قطعة بسيطة من المعدات التي تحملها في يديك وتستخدمها للقيام بنوع معين من العمل	3.24	0.20
59. Cock: Rooster	cock: ديك	A cock is an adult male chicken. الدِّيكُ هو ذَكَرُ الدَّجاجِ.	3.88	0.08
60. Boy: Lad	boy: فتى	A boy is a child who will grow up to be a man. .الفتى هو الرجل حديث السن أو الصبي	2.88	0.30
61. Cushion: Pillow	cushion: وسادة	A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable. الوسادة هي مخدة أو مسند مستطيل الشكل والذي يرتاح عليه رأسك عندما تكون في الفراش.	3.18	0.16
62. Cemetery: Graveyard	cemetery: مقبرة	A cemetery is a place where dead people's bodies or their ashes are buried. المقبرة هي مساحة من الأرض تقع أحيانا بالقرب من الكنيسة حيث يتم دفن الموتى فيها.	3.82	0.13
63. Automobile: Car	automobile: سيارة	An automobile is a car. السيارة هي عربة آليَّة سريعة السَّيْر، تُستخدم في نقل النَّاس أو البضائع، تسير بالبنزين ونحوه.	3.06	0.27
64. Midday: Noon	midday:ظهيرة	Midday is 12 o'clock in the middle of the day. الظهيرة هي الساعة 12 في منتصف النهار.	4.00	0
65. Gem: Jewel	gem: جوهرة	A gem is a jewel or stone that is used in jewellery. الجوهرة هي حجر كريم يستخدم لتزيين الأشياء القيمة التي نرتديها مثل الخواتم والقلائد.	3.65	0.12

The mean similarity of all human ratings is computed in the range [0,4] and the Standard Deviation (STD) is shown for each pair



Natural Language Processing & Computational Linguistics

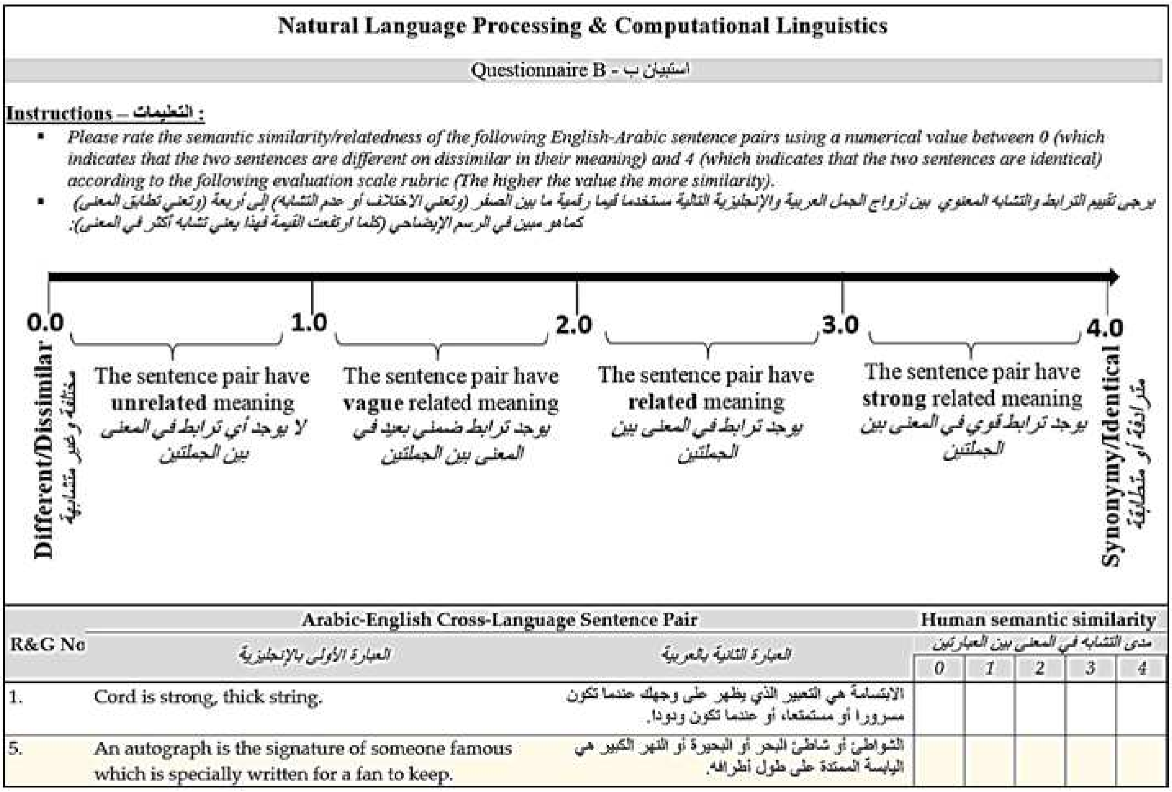
Questionnaire A - استبيان أ

Instructions – التعليمات :

- *Please tick the Arabic translation the gives the best corresponding meaning with no additional phrases with the original English sentence. Please ignore words with marked background in No. 9-11-17-21-22-24-27.*
- *اختر الترجمة العربية الأكثر مطابقة نصا ومعنى للنص الأصلي المكتوب باللغة الإنجليزية، يرجى تجاهل الكلمات المكررة المظللة في 9-11-17-21-22-24-27*

No.	R&G 2nd Word Pair	Arabic Word	English Sentence	Arabic Sentence Translation		
				Machine Translation	Human Translation	Dictionary Definition *from Arabic contemporary lexicon dictionary*
مسلسل	الكلمة الإنجليزية	الكلمة المرادفة	الجملة باللغة الإنجليزية	الترجمة الآلية	الترجمة اليدوية	التعريف المستخرج من القاموس
1	1.smile	بسمة	A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.	❑ الابتسامة هي التعبير الذي لديك على وجهك عندما يتم يسر لك أو مسليا، أو عند كونها ودية.	❑ الابتسامة هي التعبير الذي يظهر على وجهك عندما تكون مسرورا أو مستمتعا، أو عندما تكون ودودا.	❑ الابتسامة هي ضحكة خفيفة بلا صوت.
2	5.shore	شاطئ	The shores or shore of a sea, lake or wide river is the land along the edge of it.	❑ شواطئ أو شاطئ بحر أو بحيرة أو نهر واسع هي الأرض على طول حافة منه.	❑ الشواطئ أو شاطئ البحر أو البحيرة أو النهر الكبير هي اليابسة الممتدة على طول أطرافه.	❑ الشاطئ هو يابس على طول حافة محيط أو بحر أو بحيرة أو نهر.

Fig. 3. Questionnaire A distributed to participants to choose the optimal translation for each sentence



Natural Language Processing & Computational Linguistics

Questionnaire B - استبيان ب

Instructions – التعليمات :

- *Please rate the semantic similarity/relatedness of the following English-Arabic sentence pairs using a numerical value between 0 (which indicates that the two sentences are different on dissimilar in their meaning) and 4 (which indicates that the two sentences are identical) according to the following evaluation scale rubric (The higher the value the more similarity).*
- *يرجى تقييم الترابط والتشابه المعنوي بين أزواج الجمل العربية والإنجليزية التالية مستخدما قيما رقمية ما بين الصفر (وتعني الاختلاف أو عدم التشابه) إلى أربعة (وتعني تطابق المعنى) كما هو مبين في الرسم الإيضاحي (كلما ارتفعت القيمة فهذا يعني تشابه أكثر في المعنى):*

0.0
1.0
2.0
3.0
4.0
Different/Dissimilar
مختلفة وغير متشابهة
The sentence pair have **unrelated** meaning
لا يوجد أي ترابط في المعنى بين الجملتين
The sentence pair have **vague** related meaning
يوجد ترابط ضمني بعيد في المعنى بين الجملتين
The sentence pair have **related** meaning
يوجد ترابط في المعنى بين الجملتين
The sentence pair have **strong** related meaning
يوجد ترابط قوي في المعنى بين الجملتين
Synonymy/Identical
مترادفة أو متطابقة

R&G No	Arabic-English Cross-Language Sentence Pair		Human semantic similarity				
	العبارة الأولى بالإنجليزية	العبارة الثانية بالعربية	مدى التشابه في المعنى بين العبارتين				
			0	1	2	3	4
1.	Cord is strong, thick string.	الابتسامة هي التعبير الذي يظهر على وجهك عندما تكون مسرورا أو مستمتعا، أو عندما تكون ودودا.					
5.	An autograph is the signature of someone famous which is specially written for a fan to keep.	الشواطئ أو شاطئ البحر أو البحيرة أو النهر الكبير هي اليابسة الممتدة على طول أطرافه.					

Fig. 4. Questionnaire B distributed to participants to rate the sentence similarity in a predefined range

Evaluation and Statistical Analysis

Human similarity ratings were obtained as the means score in the range [0,4] and these were then scaled into the range [0,1] to be compared with the proposed semantic similarity algorithms. Statistics such as the mean, standard deviation and Pearson product-moment correlation coefficient r are commonly used for comparisons between human ratings and automated methods (Li *et al*., 2006). The results from the proposed algorithms were statistically compared with the constructed human-rated benchmark dataset using t-statistical hypothesis testing (Leech *et al*., 2008). We set a null hypothesis that "the semantic similarity evaluation by the machine and by the human perform equally (i.e., the true mean difference is zero)". A paired t-test was used to test the null hypothesis. To carry out the paired t-test on the benchmark dataset (k = 30), we calculated the difference of the results obtained by the algorithm and the mean human rating for each sentence pair as $d_i = x_i - y_i$, where $i = 1,2,\ldots,k$ and x_i refers to the *mean value of human rating* on the i^{th} pair and y_i refers to the *Sim* score obtained from the proposed algorithm on the i^{th} pair. The mean difference was computed as $\bar{d} = (\sum_{i=1}^{k} d_i) / k$ and the standard deviation of the mean differences across all sentence pairs was computed $\alpha = \sqrt{\sum_{i=1}^{k} (d_i - \bar{d})^2 / (k-1)}$

We used α to compute the standard error $SE(\bar{d}) = \alpha / \sqrt{k}$ and the *t*-statistic $T = \bar{d} / SE(\bar{d})$, which follows a normal distribution with *k*-1 degrees of freedom under the null hypothesis. Using *t*-distribution table, we compared *T* to the $t_{k\text{-}1}$ distribution to obtain the probability value, referred to as the *p*-value to reject/not reject the null hypothesis.

Results and Discussion

Results from Sentence Samples

Table 5 shows the results obtained from the proposed algorithms as compared with the human-like similarity obtained by Li *et al.* (2006) on a sample of sentences. The correlation coefficients, *r*, were 0.624, 0.793 and 0.928 obtained from the averaged maximum-translation, noun-verb vector and the MT-based term vector similarity algorithms, respectively. The first pair which has two sentences totally different in their words as well as their meaning (the pairs are shown in Table 2), all algorithms obtained zero similarity as to indicate the fact that they are completely dissimilar. Pairs 2 and 3 have sentences that share common words but their meaning is somehow different. We can see that the human-like similarity is 0.74 but our methods were computed based on the terms that share the same or very similar semantic meaning and, therefore, they obtained higher similarity results (0.89 and 1.0 using the term vector similarity method). Other results in the remainder pairs were almost consistent with the human understanding and they also showed that the MT-based term vector similarity algorithm obtained the highest correlation with the human-like similarity.

Results from the Human-Rated Benchmark Dataset

This section covers the experimental works that we carried out to validate the proposed models. As mentioned above, the ground-truth benchmark was created based on human ratings. Table 6 presents the human similarity scores for each sentence pairs and those obtained by the three algorithms, namely the averaged maximum-translation similarity algorithm, the noun-verb vector similarity algorithm and the MT-based term vector similarity algorithm. The correlation coefficients *r* obtained by these algorithms were 0.7206, 0.5512 and 0.8657, respectively.

As can be seen from the table, the averaged maximum-translation and noun-verb vector similarities obtained a reasonably good correlation with the human understanding if we take into consideration these sentences were processed from two different languages. MT-based term vector similarity achieved a remarkably better Pearson correlation coefficient with the human intuition significant at the 0.01 level.

However, as mentioned in Li *et al.* (2006), a further factor should be accounted in order to decide what is the best performance that can be achieved by the computerised similarity algorithms under this particular set of benchmarks and experimental conditions. Thus, an upper bound was determined in this study using leave-one-out resampling technique whereby we repeated the evaluation *n* times (*n* = number of the participants). Each time, we computed the Pearson correlation coefficient of the judgement of each individual participant against the group of all participants and then we took the mean as the upper bound. As shown in Fig. 5, the best human participant's correlation coefficient is 0.9445 and the worst is 0.5994 whereas the mean (upper performance) is 0.878. By considering the mean of all human participants as a typical higher performance rate can be attained, we found that our algorithm that used MT-based term vector similarity achieved a close estimation to this upper bound.

Table 5. Experimental results on raw sentences of short lengths

	Similarity	Dictionary-Based	Machine Translation-Based	
En-Ar pairs	*Human-like Simy*	*Averaged max-translation sim*	*Noun-verb vector sim*	*MT-based term vector sim*
Pair 1	0.00	0.000	0.000	0.000
Pair 2	0.98	0.500	0.600	0.890
Pair 3	0.74	1.000	1.000	1.000
Pair 4	0.74	0.940	0.990	0.870
Pair 5	0.62	0.000	0.940	0.700
Pair 6	0.51	0.700	0.610	0.430
Pair 7	0.36	0.380	0.590	0.390
Pair 8	0.13	0.000	0.000	0.000
Pair 9	0.68	0.660	0.960	0.550
Pair 10	0.12	0.400	0.450	0.280
Pearson Correlation *r*	1.00	0.624	0.793	0.928

Table 6. Experimental results on Arabic-English cross-language short texts benchmark using averaged max-translation, noun-verb vector and MT-based term vector similarity algorithms

			Similarity algorithms		
R&G No.	En-Ar sentence pair	Human similarity (Mean) [0,1]	*Averaged max-translation sim*	*Noun-verb vector sim*	*MT-based term vector sim*
1.	cord: بسمة	0.03	0.1000	0.0300	0.1500
5.	autograph: شاطئ	0.00	0.0900	0.0500	0.1200
9.	asylum: فاكهة	0.01	0.0600	0.0800	0.2900
13.	boy: ديك	0.15	0.4400	0.6100	0.2500
17.	coast: غابة	0.24	0.2600	0.6100	0.2600
21.	boy: حكيم	0.15	0.4500	0.3500	0.5700
25.	forest: مقبرة	0.22	0.5500	0.8800	0.4100
29.	bird:غابة	0.09	0.5100	0.6100	0.4700
33.	hill: غابة	0.22	0.3100	0.9700	0.5300
37.	magician: عرَّاف	0.51	0.4100	0.8500	0.4600
41.	oracle: حكيم	0.25	0.2600	0.9600	0.1600
47.	furnace:موقد	0.72	0.4300	0.8800	0.5400
48.	magician:ساحر	0.76	0.6000	0.5900	0.5700
49.	hill: كومة	0.44	0.7200	0.9400	0.5700
50.	cord: سلسلة	0.63	0.4800	0.9200	0.5100
51.	glass: قدح	0.35	0.5100	0.9300	0.5300
52.	grin: ابتسامة	0.50	0.6700	0.6100	0.3900
53.	serf: عبد	0.82	0.5600	0.9400	0.5600
54.	journey:رحلة	0.57	0.6500	0.5500	0.5500
55.	autograph: توقيع	0.63	0.5300	0.9600	0.6400
56.	coast: شاطئ	0.85	0.7300	0.6300	0.7600
57.	forest: غابة	0.97	0.4400	0.6500	0.9600
58.	implement: أداة	0.81	0.9500	0.6400	0.6200
59.	cock: ديك	0.97	0.9700	0.9900	0.9300
60.	boy: فتى	0.72	0.7700	0.6200	0.6900
61.	cushion: وسادة	0.79	0.5400	0.9700	0.7500
62.	cemetery: مقبرة	0.96	0.7000	0.9700	0.9800
63.	automobile: سيارة	0.76	1.0000	0.5700	0.7400
64.	midday:ظهيرة	1.00	0.5900	0.9400	0.8500
65.	gem: جوهرة	0.91	0.9900	0.9600	0.8500
Pearson correlation *r*		1.00	0.7206	0.5512	0.8657

Table 7. Statistical results obtained using *t*-test from human similarity evaluation versus proposed automatic similarity evaluation algorithms

Statistics	Human evaluation Vs. Algorithm 1	Human evaluation Vs. Algorithm 2	Human evaluation Vs. Algorithm 3
Hypothesis =	*Maximum-translation similarity algorithm performs equal to human evaluation.*	*Noun-verb vector similarity algorithm performs equal to human evaluation.*	*MT-based term vector similarity algorithm performs equal to human evaluation.*
Alpha level =	0.05	0.05	0.05
Mean differences =	-0.0080	-0.1743	-0.0210
Standard deviation =	0.2290	0.2935	0.1718
t-Statistic =	-0.1914	-3.2535	-0.6696
t-Critical Value =	±2.0452	±2.0452	±2.0452
p-Value =	0.8496	0.0029	0.5084
Decision =	Do not reject hypothesis	Reject hypothesis	Do not reject hypothesis
Confidence interval for paired difference			
Confidence level	0.95	0.95	0.95
Confidence interval	-0.0935 < µd < 0.07750	-0.2839 < µd < -0.06474	-0.0851 < µd < 0.04314

Statistical Results and Discussion

Further statistical analysis has been done using *t*-Test on the results obtained by the human participants versus each of the proposed automatic cross-language similarity algorithms. Table 7 shows the statistical results from the three algorithms as obtained from the benchmark dataset wherein the sample size = 30 and confidence level 0.95. It can be seen from the *p*-value that maximum-translation similarity algorithm and MT-based term vector similarity algorithm perform equivalently to the human evaluation, while noun-verb vector similarity algorithm is significantly different. This may be because the latter algorithm do not consider all of the terms found in the texts such as adjectives and adverbs and computed based on the nouns and verbs only.

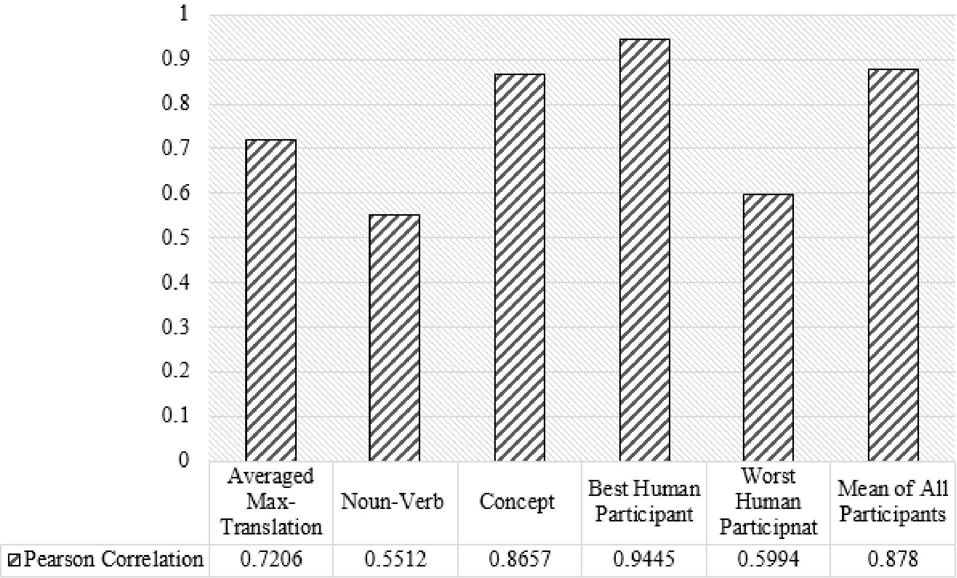


Fig. 5. Correlations from the proposed similarity algorithms, and best, worst and mean from all human participants

Accordingly, we can say that methods that conserved all of the semantic terms in terms of their meanings across two texts in different languages may work comparably and obtain positive results, regardless of the usage of dictionary translation or MT for finding the translation of these terms.

Conclusion and Future Research

This paper proposed and compared different methods for measuring the cross-language semantic similarity between short phrases and sentences. Three algorithms namely the averaged maximum-translation similarity algorithm, the noun-verb vector similarity algorithm and the MT-based term vector similarity algorithm have been investigated on Arabic-English texts. The influences made by this paper can be summarized in two points. First, a standard cross-language benchmark was constructed and verified based on a ground-truth dataset. Second, the proposed algorithms consider the impact of either dictionary translations, noun and verb vectors, or term vectors, in order to judge the relationship of two sentences derived from two different languages in terms of their meaning. These algorithms have been applied for the first time in the Arabic-English cross lingual setting as indicated by the literature review. Thus, our cross-language semantic similarity algorithms were developed and tested not only to capture common semantic similarity of two languages, but also to establish a comparison base for further research. To evaluate our cross-language similarity algorithms, we used a set of sentence pairs from computational linguistics. An initial experiment on this data illustrates that the proposed algorithms provides similarity scores that are fairly consistent with human understanding. Next, we compared the similarity results obtained by our algorithms with similarity scores rated by human participants in the benchmark by taking into consideration an upper bound similarity rate obtained by the participants. Statistical results showed that using MT or dictionary translation can both achieve a comparable behaviour if we employ good semantic similarity measurements. Further research will include the construction of a wider selection of sentence pairs annotated with human's ratings and explore these algorithms across different languages. An improvement to the algorithms can be made when we use word sense disambiguation. More sophisticated algorithms proposed recently such as BabelRelate (Navigli and Ponzetto, 2012b) and CL-ESA (Sorg and Cimiano, 2010; Anderka *et al*., 2009) will be explored in further studies which in turn would help to apply these techniques on sentences of medium to large lengths. Presently, comparison of our techniques with some of the other algorithms is difficult due to a lack of published work on measuring the semantic similarities in the Arabic-English cross-language domain.

Acknowledgment

The author thanks the participants in the survey and those who provided useful comments on the manuscript.

Ethics

No ehtical issues would arise after the publication of this manuscript.

References

Agirre, E., C. Banea, C. Cardie, D. Cer and M. Diab *et al.*, 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. Proceedings of the 8th International Workshop on Semantic Evaluation, Aug. 23-24, Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pp: 81-91. DOI: 10.3115/v1/S14-2010

Agirre, E., C. Banea, C. Cardie, D. Cer and M. Diab *et al.*, 2015. Semeval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. Proceedings of the 9th International Workshop on Semantic Evaluation, Jun. 4-5, Association for Computational Linguistics, Denver, Colorado, pp: 252-263. DOI: 10.18653/v1/S15-2045

Agirre, E., D. Cer, M. Diab and A. Gonzalez-Agirre, 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. Proceedings of the 6th International Workshop on Semantic Evaluation, Jun. 7-8, Montreal, Canada, pp: 385-393.

Agirre, E., D. Cer, M. Diab, A. Gonzalez-Agirre and W. Guo, 2013. Sem 2013 shared task: Semantic textual similarity. Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics, Jun. 13-14, Association for Computational Linguistics, Atlanta, Georgia, pp: 32-43.

Alzahrani, S.M., N. Salim and A. Abraham, 2012. Understanding plagiarism linguistic patterns, textual features and detection methods. IEEE Trans. Syst. Man Cybernet. Part C, 42: 133-149. DOI: 10.1109/TSMCC.2011.2134847

Alzahrani, S.M., N. Salim and V. Palade, 2015. Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. J. King Saud Univ. Comput. Inform. Sci., 27: 248-268. DOI: 10.1016/j.jksuci.2014.12.001

Anderka, M., N. Lipka and B. Stein, 2009. Evaluating cross-language explicit semantic analysis and cross querying. Proceedings of the 10th Cross-Language Evaluation forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments, Sept. 30-Oct. 2, Springer, Greece, pp: 50-57. DOI: 10.1007/978-3-642-15754-7_4

Banerjee, S. and T. Pedersen, 2003. Extended gloss overlaps as a measure of semantic relatedness. Proceedings of the 18 International Joint Conference on Artificial Intelligence, Aug. 9-15, Acapulco, Mexico, pp: 805-810.

Bar, D., C. Biemann, I. Gurevych and T. Zesch, 2012. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. Proceedings of the 6th International Workshop on Semantic Evaluation, Jun. 7-8, Association for Computational Linguistics, Montreal, Canada, pp: 435-440.

Barrón-Cedeño, A., P. Gupta and P. Rosso, 2013. Methods for cross-language plagiarism detection. Knowl. Based Syst., 50: 211-217. DOI: 10.1016/j.knosys.2013.06.018

Batet, M., S. Harispe, S. Ranwez, D. Sánchez and V. Ranwez, 2014. An information theoretic approach to improve semantic similarity assessments across multiple ontologies. Inform. Sci., 283: 197-210. DOI: 10.1016/j.ins.2014.06.039

Bin, Y., L. Xiao-Ran, L. Ning and Y. Yue-Song, 2012. Using information content to evaluate semantic similarity on hownet. Proceedings of the 8th International Conference on Computational Intelligence and Security, Nov. 17-18, IEEE Xplore Press, Guangzhou, pp: 142-145. DOI: 10.1109/CIS.2012.39

Budanitsky, A. and G. Hirst, 2006. Evaluating wordnet-based measures of lexical semantic relatedness. Comput. Linguist., 32: 13-47. DOI: 10.1162/coli.2006.32.1.13

Corley, C. and R. Mihalcea, 2005. Measuring the semantic similarity of texts. Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, (SEE' 05), Association for Computational Linguistics, Stroudsburg, PA, USA., pp: 13-18. DOI: 10.3115/1631862.1631865

Dai, L. and H. Huang, 2011. An English-Chinese cross-lingual word semantic similarity measure exploring attributes and relations. Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, Institute of Digital Enhancement of Cognitive Processing, Waseda University, pp: 467-476.

Dai, L., B. Liu, Y. Xia and S. Wu, 2008. Measuring semantic similarity between words using HowNet. Proceedings of the International Conference on Computer Science and Information Technology, Aug. 29-Sept. 2, IEEE Xplore Press, Singapore, pp: 601-605. DOI: 10.1109/ICCSIT.2008.101

Edward, L. and B. Steven, 2002. Nltk: The natural language toolkit. Proceedings of the Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, (PCL' 02), Association for Computational Linguistics, Philadelphia, Pennsylvania.

Fernando, S. and M. Stevenson, 2008. A semantic similarity approach to paraphrase detection. Proceedings of the 11th Annual Research Colloquium on Computational Linguistics (CCL' 08), Oxford University Computing Laboratory, Oxford, UK.

Hajian, B. and T. White, 2011. Measuring semantic similarity using a multi-tree model. Proceedings of the 9th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems, (PRS' 11), Barcelona, Spain, pp: 7-14.

Hirst, G. and D. St Onge, 1998. Lexical Chains as Representation of Context for the Detection and Correction Malapropisms. In: Wordnet: An Electronic Lexical Database, Fellbaum, C. (Ed.), MIT Press, Cambridge, ISBN-10: 026206197X, pp: 305-332.

Islam, A. and D. Inkpen, 2008. Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans. Knowl. Discovery Data, 10: 10-25.

Jian-Bo, G., Z. Bao-Wen and C. Xiao-Hua, 2013. Ontology-based semantic similarity: A new approach based on analysis of the concept intent. Proceedings of the International Conference on Machine Learning and Cybernetics, Jul. 14-17, IEEE Xplore Press, Tianjin, pp: 676-681. DOI: 10.1109/ICMLC.2013.6890375

Jiang, J.J. and D.W. Conrath, 1997. Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of the International Conference Research on Computational Linguistics, (RCL' 97).

Jimenez, S., C. Becerra and A. Gelbukh, 2012. Soft cardinality: A parameterized similarity function for text comparison. Proceedings of the 6th International Workshop on Semantic Evaluation, (WSE' 12), Association for Computational Linguistics, Montreal, Canada, pp: 449-453.

Landauer, T., P.W. Foltz and D. Laham, 1998. An introduction to latent semantic analysis. Discourse Processes, 25: 259-284. DOI: 10.1080/01638539809545028

Leacock, C. and M. Chodorow, 1998. Combining Local Context with Wordnet Similarity for Word Sense Identification. In: Wordnet: A Lexical Reference System and its Application, Fellbaum, C. (Ed.) MIT Press, Cambridge, MA., ISBN-10: 026206197X, pp: 265-283.

Lee, M.C., 2011. A novel sentence similarity measure for semantic-based expert systems. Expert Syst. Applic., 38: 6392-6399. DOI: 10.1016/j.eswa.2010.10.043

Leech, N.L., K.C. Barrett and G.A. Morgan, 2008. SPSS for Intermediate Statistics: Use and Interpretation. 3rd Edn., Lawrence Erlbaum Associates, New York, ISBN-10: 0805862676, pp: 270.

Li, Y., D. McLean, Z.A. Bandar, J.D. O'Shea and K. Crockett, 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowl. Data Eng., 18: 1138-1150. DOI: 10.1109/TKDE.2006.130

Li, Y., Z.A. Bandar and D. McLean, 2003. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowl. Data Eng., 15: 871-882. DOI: 10.1109/TKDE.2003.1209005

Lin, D., 1998. An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning, (CML' 95), Morgan Kaufmann Publishers Inc. San Francisco, pp: 296-304.

Liu, H., H. Bao and D. Xu, 2012. Concept vector for semantic similarity and relatedness based on wordnet structure. J. Syst. Software, 85: 370-381. DOI: 10.1016/j.jss.2011.08.029

Luo, Q., E. Chen and H. Xiong, 2011. A semantic term weighting scheme for text categorization. Expert Syst. Applic., 38: 12708-12716. DOI: 10.1016/j.eswa.2011.04.058

Meng, L., R. Huang and J. Gu, 2014. Measuring semantic similarity of word pairs using path and information content. Int. J. Future Generat. Commun. Network., 7: 183-194. DOI: 10.14257/ijfgcn.2014.7.3.17

Mihalcea, R., C. Corley and C. Strapparava, 2006. Corpus-based and knowledge-based measures of text semantic similarity. Proceedings of the 21th National Conference on Artificial Intelligence, (CAI' 06), AAAI Press, pp: 775-780.

Miller, G.A., 1995. Wordnet: A lexical database for English. Commun. ACM., 38: 39-41. DOI: 10.1145/219717.219748

Monroe, W., S. Green and D.C. Manning, 2014. Word segmentation of informal Arabic with domain adaptation. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Jun. 23-25, Association for Computational Linguistics, Baltimore, Maryland, USA, pp: 206-211. DOI: 10.3115/v1/p14-2034

Muftah, A.J.A., 2009. Document plagiarism detection algorithm using semantic networks. MSc Thesis, Universiti Teknologi Malaysia.

Navigli, R. and S.P. Ponzetto, 2012a. Babelrelate! A joint multilingual approach to computing semantic relatedness. Proceedings of the 26th AAAI Conference on Artificial Intelligence, (CAI' 12), Toronto, Ontario, Canada, pp: 108-114.

Navigli, R. and S. Ponzetto, 2012b. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intell., 193: 217-250. DOI: 10.1016/j.artint.2012.07.001

O'Shea, J., Z. Bandar, K. Crockett and D. McLean, 2008. A comparative study of two short text semantic similarity measures. Proceedings of the 2nd KES International Conference on Agent and Multi-Agent Systems: Technologies and Applications, Mar. 26-28, Springer, Korea, pp: 172-181. DOI: 10.1007/978-3-540-78582-8_18

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence, (CAI' 95), Morgan Kaufmann Publishers Inc., San Francisco, pp: 448-453.

Resnik, P., 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artificial Intell. Res., 11: 95-130. DOI: 10.1613/jair.514

Rios, M., 2014. Uow: Multi-task learning Gaussian process for semantic textual similarity. Proceedings of the 8th International Workshop on Semantic Evaluation, Aug. 23-24, Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pp: 779-784.

Roth, R., O. Rambow, N. Habash, M. Diab and C. Rudin, 2008. Arabic morphological tagging, diacritization and lemmatization using lexeme models and feature ranking. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, (TSP' 08), Association for Computational Linguistics, USA., pp: 117-120. DOI: 10.3115/1557690.1557721

Rubenstein, H. and J.B. Goodenough, 1965. Contextual correlates of synonymy. Commun. ACM, 8: 627-633. DOI: 10.1145/365628.365657

Sánchez, D., M. Batet, D. Isern and A. Valls 2012. Ontology-based semantic similarity: A new feature-based approach. Expert Syst. Applic., 39: 7718-7728. DOI: 10.1016/j.eswa.2012.01.082

Solé-Ribalta, A., 2014. Towards the estimation of feature-based semantic similarity using multiple ontologies. Knowl. Based Syst., 55: 101-113. DOI: 10.1016/j.knosys.2013.10.015

Sorg, P. and P. Cimiano, 2010. An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. Proceedings of the 14th international conference on Applications of Natural Language to Information Systems, Jun. 24-26, Springer, Germany, pp: 36-48. DOI: 10.1007/978-3-642-12550-8_4

Stoyanova, I., S. Koeva and S. Leseva, 2013. Wordnet-based cross-language identification of semantic relations. Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, (NLP' 13), Sofia, Bulgaria, pp: 119-113.

Toutanova, K., D. Klein, C.D. Manning and Y. Singer, 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, (HLT' 03), Association for Computational Linguistics, USA., pp: 173-180. DOI: 10.3115/1073445.1073478

Turney, P.D., 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the 12th European Conference on Machine Learning, Sept. 5-7, Springer, Germany, pp: 491-502. DOI: 10.1007/3-540-44795-4_42

Vulic, I. and M. Moens, 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Jun. 9-16, Atlanta, Georgia, USA, pp: 106-116.

Vulic, I. and M.F. Moens, 2014. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Oct. 25-29, Doha, Qatar, pp: 349-362.

Wu, S., J.D. Choi and M.S. Palmer, 2010. Detecting cross-lingual semantic similarity using parallel propbanks. Proceedings of the 9th Conference of the Association for Machine Translation in the Americas, (MTA' 10), Denver, Colorado.

Wu, Z. and M. Palmer, 1994. Verbs semantics and lexical selection. Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, (ACL' 94), Association for Computational Linguistics, Stroudsburg, PA, USA., pp: 133-138. DOI: 10.3115/981732.981751

Xu, W., C. Callison-Burch and D.B., Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (PIT). Proceedings of the 9th International Workshop on Semantic Evaluation, Jun. 4-5, Association for Computational Linguistics, Denver, Colorado, pp: 1-11. DOI: 10.18653/v1/S15-2001

Ye, Z. and Y. Zhan-Lin, 2010. Research on ontology-based semantic similarity computation. Proceedings of the International Conference on Machine Vision and Human-Machine Interface, Apr. 24-25, IEEE Xplore Press, Kaifeng, China, pp: 472-475. DOI: 10.1109/MVHI.2010.33

Yerra, R. and Y.K. Ng, 2005. A sentence-based copy detection approach for web documents. Proceedings of the 2nd International Conference on Fuzzy Systems and Knowledge Discovery, Aug. 27-29, Springer, China, pp: 557-570. DOI: 10.1007/11539506_70

Zhang, C., Y. Yang, X. Guo, Z. Du and N. Lin, 2014b. The improved algorithm of semantic similarity based on the multi-dictionary. J. Software, 9: 324-328. DOI: 10.4304/jsw.9.2.324-328

Zhang, P., Z. Zhang, W. Zhang and C. Wu, 2014a. Semantic similarity computation based on multi-feature combination using hownet. J. Software, 9: 2461-2466. DOI: 10.4304/jsw.9.9.2461-2466

Zhou, D., T. Brailsford, V. Wade and H. Ashman, 2012. Translation techniques in cross-language information retrieval. ACM Comput. Surv., 45: 1-44. DOI: 10.1145/2379776.2379777

Zou, W.Y., R. Socher, D. Cer and C.D. Manning, 2013. Bilingual word embeddings for phrase-based machine translation. Proceedings of the Empirical Methods in Natural Language Processing (NLP' 13), Seattle, USA.