Journal of Information & Knowledge Management
Vol. 19, No. 4 (2020) 2050033 (29 pages)

DOI: 10.1142/S0219649220500331

World Scientific
www.worldscientific.com

# Semantic Similarity for English and Arabic Texts: A Review

Marwah Alian
*Princess Sumaya University for Technology*
*Amman, Jordan*
*Hashemite University, Zarqa, Jordan*
*Marwah2001@yahoo.com*

Arafat Awajan
*Princess Sumaya University for Technology*
*Amman, Jordan*
*Awajan@psut.edu.jo*

Published 2 December 2020

**Abstract.** Semantic similarity is the task of measuring relations between sentences or words to determine the degree of similarity or resemblance. Several applications of natural language processing require semantic similarity measurement to achieve good results; these applications include plagiarism detection, text entailment, text summarisation, paraphrasing identification, and information extraction. Many researchers have proposed new methods to measure the semantic similarity of Arabic and English texts. In this research, these methods are reviewed and compared. Results show that the precision of the corpus-based approach exceeds 0.70. The precision of the descriptive feature-based technique is between 0.670 and 0.86, with a Pearson correlation coefficient of over 0.70. Meanwhile, the word embedding technique has a correlation of 0.67, and its accuracy is in the range 0.76–0.80. The best results are achieved by the feature-based approach.

*Keywords*: Semantic similarity; feature-based; word embeddings; statistical corpus-based; sentence similarity; word similarity, document similarity.

## 1. Introduction

Measuring the semantic similarity between two texts is challenging because it involves determining the relation between words in a text in accordance with the contained idea (Alzahrani, 2016). The similarity of short texts usually ranges from complete semantic equivalence to being exactly unrelated in meaning; the similarity score provides a notion of intermediate similarity because the two texts may share several aspects of meaning or have semantically important differences. The semantic similarity task is described as a black box in several natural language processing (NLP) applications, and it can be evaluated independently or as an internal part of an application (Agirrea *et al.*, 2015). This important task is used in many applications and areas, such as plagiarism detection, text entailment, text summarisation, machine translation, question–answer, text categorisation, paraphrasing, information retrieval, image retrieval from the web based on captions, web page retrieval,

information extraction, and machine translation (Nagoudi and Schwab, 2017; Li *et al.*, 2006). Most methods for depicting similarity among texts are based on approaches used for long texts. Therefore, the sentences are implemented and processed in a high-dimensional space, but they suffer from inefficiency and inadaptability in several domains (Li *et al.*, 2006). These methods do not provide good results, and many researchers rely on other approaches that are appropriate for sentence similarity.

Semantic similarity s a hot topic, and several methods have been developed and tested for English texts. However, the work on Arabic semantic similarity is limited in number and results. Not all existing techniques used for other languages have been tested for Arabic. In this research, we present a survey of studies conducted on Arabic and English. The rest of the paper is organised as follows. Section 2 categorises semantic similarity approaches, and Secs. 3 and 4 review and compare semantic similarity approaches for English and Arabic texts, respectively. Section 5 describes and compares the benchmarks that have been used in the evaluation of similarity methods. Section 6 provides a summary of the study, and Sec. 7 presents the conclusions.

## 2. Coverage of Semantic Similarity Approaches

Text similarity can be classified in many ways depending on the aspect of similarity, approach used, morphological level of each similarity technique, and text application. With regard to the aspect or object of similarity, previous work can be divided into three groups as follows: word, sentence, and document similarity. The similarity between sentences or documents depends mainly on the similarity between words as the smallest component of documents or sentences. The similarity between words is measured by the syntactic or semantic features of words (Gomaa, 2013).

Different approaches are used to measure the semantic similarity between texts. These approaches can be categorised into four main categories, namely, methods based on word co-occurrence, statistical corpus, descriptive features, and word embedding. The word co-occurrence-based approach is primarily used to depict the similarity between documents or find a related document for a given query. In this approach, source and destination documents are represented by their bag of words. For example, in search engines, documents and queries are presented as vectors, and the cosine similarity between the document and query vectors is measured to determine which documents are the most similar. However, this approach does not consider the word order or words with multiple meanings (Li *et al.*, 2006).

In the statistical corpus-based approach, a word-by-context or word-by-word matrix is formed based on the count of words in the corpus. Then, a vector is constructed for each sentence in the reduced dimension space, and the cosine similarity between these vectors is measured. This approach produces a correct match for a query with documents that have a similar meaning even when they do not have similar query words. However, this approach does not consider words with multiple meanings nor the word order (Landauer *et al.*, 1997; Kintsch *et al.*, 1998). In the descriptive feature-based approach, lexical database and corpus are utilised to

measure the semantic similarity between words in two sentences in order to perform a vector representation for each sentence, and the word order is used in the computation of the overall semantic similarity score between the two sentences; however, the research community does not use machine learning in this approach (Hatzivassiloglou *et al.*, 1999).

The word embedding-based approach considers the word context in its sentence representation depending on the content word vectors. However, this approach cannot learn separate embedding for multiple senses of a word (Kenter and de Rijke, 2015). Notably, the statistical corpus-based approach has a precision of over 0.70. Studies that used the descriptive feature-based technique obtained a precision value between 0.670 and 0.86 and a Pearson correlation of over 0.70. Meanwhile, the word embedding technique has a correlation of 0.67 and accuracy between 0.76 and 0.80.

According to Pronoza and Yagunova (2015), sentence similarity techniques can be classified into shallow, semantic, syntactic, and distributional approaches. The shallow approach depends on the surface forms of the words and uses the overlapping of words or phrases or other shallow features such as difference in sentence length, longest common subsequence, and proper name matching number (Cha, 2007). This approach does not utilise any semantic resources. On the contrary, the semantic approach depends on the semantic structure of the sentence and employs external semantic resources, such as WordNet, where synonymy or hyponymy relations are defined. In this approach, the similarity between words is defined as the shortest path between these words in WordNet or any other WordNet-based measures. This approach is better than the shallow approach, but its efficiency is based on the quality of the external resource used. The syntactic approach is based on the overlapping of dependency relations or distance between syntactic parse trees (Dagan *et al.*, 2005; Wan *et al.*, 2006). This approach can capture more linguistic features than other approaches can, but it could produce errors because it depends on a syntactic parser output. Meanwhile, the distributional approach depends on the idea that semantically similar words occur in a similar context and are represented by close vectors in the space. This approach is based on distributional or vector space models and involves two models: count based and predictive. The count-based model utilises latent semantic analysis (LSA), and the predictive model uses Word2vec or Glove.

It is proved by Achananuparp *et al.* (2008) that the distributional approach measures similarity better than syntactic and semantic approaches do. In accordance with the language of text similarity, the approach can consider one language, such as English and Arabic, or multiple languages where the similarity is measured between two texts in different languages, such as Arabic and English.

## 3. English Text Semantic Similarity

### 3.1. *Word level*

In English word level semantic similarity, Turney (2005) proposed a similarity approach that is based on Latent Relational Analysis. To evaluate his approach, he

uses a corpus of 5 × 1010 English words collected from the United States academic web site. This method achieves a recall of 0.536 and a precision of 0.559. Chen *et al.* (2006) proposed a double co-occurrence checking method which is evaluated using Rubenstein–Goodenough's (R&G) benchmark where the results show a correlation of 0.849 when the total snippets to be analyzed are equal to 600. Also, Aouicha and Taieb (2015) proposed a Gloss-based approach that depends on WordNet and Wiktionary to measure semantic similarity between words. This approach is evaluated using four datasets: RG65, MC30, AG203 and GeReSiD50. The results show that the Gloss-based approach achieves a correlation of a range from 0.620 to 0.824 with the datasets. Also, it provides a correlation of 0.563 compared to the measure X-Similarity which gives a correlation of 0.493.

Some researchers use feature-based approach for word similarity such as Taieb *et al.* (2013) where they use MTurk, WordSim353, Rubenstein and Goodenough (R&G), Miller and Charles (M&C), Yang and Powers dataset. Their approach obtains the best correlation of 0.654 with MTurk dataset and a correlation of 0.72 with WordSim353 dataset. When used with the application of solving word choice problems; it produces a gain equal to 0.153 compared to other approaches. Jiang *et al.* (2015) evaluate their approach on M&C's, R&G's benchmark, WordSimilarity-353 Test Collection4 (353-TC dataset) and their constructed benchmark. The results show that it has performed well on M&C benchmark with a correlation of 0.827 and 0.763 on R&G, and 0.827 on their own benchmark. But it does not perform well on the 353-TC dataset. A comparison between approaches used for similarity between words is shown in Table 1 and the correlation results of these approaches are presented in Fig. 1.

### 3.2. *Sentence level*

In the sentence similarity level, Sahami and Heilman (2006) apply Web-based kernel function on a set of documents retrieved from a web search engine while Islam and Inkpen (2005) have proposed an approach that uses corpus-based measure with the Longest Common Subsequence (LCS) string matching. In the evaluation, two datasets are used; the same dataset as Li *et al.* (2006) and the paraphrase dataset which contains 4076 training pairs and 1725 testing pairs. The results show that this approach achieves an accuracy of 0.726 with a recall of 0.891 and a precision of 0.74.

Some sentence similarity researches use feature-based approach such as Li *et al.* (2006) who have proposed an algorithm for measuring similarity between sentences where word order and semantic information are taken into account. Information from corpus statistics and lexical database are utilised to simulate Human common sense knowledge in a model and to adjust the algorithm to be applicable to different domains. This information is used to derive a semantic similarity then a word order similarity is considered to study its impact on the meaning of a sentence where the word order similarity is measured using the number of different words and the number of different order of word pairs. The combination of the semantic similarity

Table 1. English words similarity approaches.

| Technique | Year | Dataset | Results |
|---|---|---|---|
| Latent Relational Analysis (Turney, 2005) | 2005 | A corpus collected from the United States academic web site. | Achieves a recall of 0.536 and a precision of 0.559. |
| Co-occurrence Double checking (Chen *et al.*, 2006) | 2006 | R&G's benchmark | Obtains a correlation of 0.849 when the total analyzed snippets are 600. |
| Knowledge-based approach (Taieb *et al.*, 2013) | 2013 | MTurk, WordSim353 R&G, M&C, Yang and Powers dataset | Obtains the best correlation of 0.654 with MTurk dataset. |
| Gloss-based approach (Aouicha and Taieb, 2015) | 2015 | RG65, MC30, AG203 and GeReSiD50 | Achieves a correlation between 0.620 to 0.824 with the datasets |
| Feature-based approach (Jiang *et al.*, 2015) | 2015 | M&C, R&G's benchmark, WordSimilarity-353 Test Collection4 and a constructed benchmark | Performs well on M&C benchmark with a correlation of 0.827 and 0.763 on R&G, and 0.827 on their own benchmark. But it does not perform well on 353-TC dataset. |

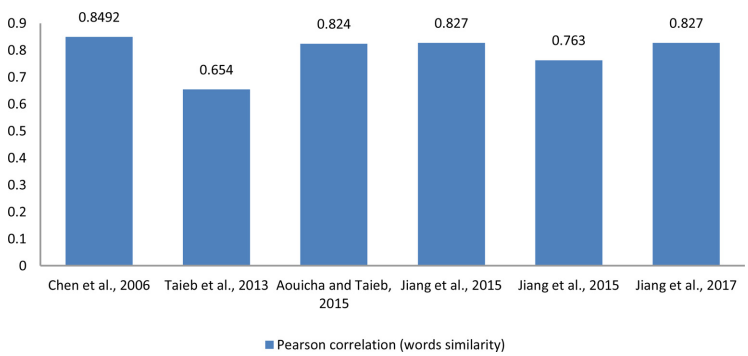


Fig. 1. Correlation results for English words similarity.

and the word order similarity construct the overall semantic similarity after adding a weight for order similarity less than the weight for semantic similarity. They use sentences from Brown corpus to evaluate the proposed approach which provides a Pearson correlation coefficient of 0.816 compared to human ratings.

Sultan *et al.* (2014) proposed an alignment pipeline which is evaluated on SemEval STS 2014 datasets and provides a correlation of 0.82 between system output and human annotations when it is evaluated on the dataset of image descriptions. Taieb *et al.* (2015) use Li *et al.* dataset and MicroSoft Paraphrase Corpus (MSPC) for evaluation where the similarity approach achieves a correlation of 0.79 when applied to Li *et al.* dataset with a recall of 0.76 and a precision of 0.79. Also, it outperforms other measures when applied to MSPC with a recall of 0.9937

and a precision of 0.670. Atish *et al.* (Pawar and Mago, 2018) used Pilot Short Text Semantic Similarity Benchmark dataset, and 65 noun pairs of R&G benchmark. Their approach obtains a good correlation of 0.8794 for sentences similarity compared to the mean of human similarity. Also, it provides a correlation of 0.8753 for words similarity.

The work of Wali *et al.* (2017) introduces an approach for measuring similarity between sentences based on three components: lexical similarity, semantic similarity and syntactico-semantic similarity. The lexical similarity uses common words while semantic similarity uses synonyms for words in each sentence and common semantic arguments which are used to measure syntactico-semantic similarity. Also, they utilise WordNet and VerbNet in measuring the semantic and syntactico-semantic similarity. They evaluate their method using Li *et al.* dataset (Li *et al.*, 2006) and MSPC. Also, it is compared to STS (Islam and Inkpen, 2005) and FM3S (Taieb *et al.*, 2015) measures where the proposed approach provides a high correlation and achieves a recall of 0.87 and a precision of 0.87 using Li *et al.* dataset.

Pilehvar *et al.* (2013) follow the Disambiguate and walk technique and evaluate it using SemEval-2012 datasets: RG-65, MSRpar and SMTnews dataset. This approach achieves a correlation of 0.866 for text similarity and provides a correlation of 0.841, 0.868 and 0.825 with R&G dataset as results with the Jaccard, Weighted Overlap and Cosine signature, respectively.

Other researchers use word embedding technique such as Kenter and de Rijke (2015) who use Microsoft Research Paraphrase Corpus dataset (MSRP) which consists of 5801 sentence pairs in total, 4076 for the training set and 1725 for the testing set. Their approach performs better than the neural network-based approach and the results, when the saliency-weighted semantic network features are added, show a better performance with accuracy of 0.76 and a recall of 0.906. Also, Pagliardini *et al.* (2018) have proposed a sentence embedding approach and used three datasets: the Toronto book corpus, Wikipedia sentences and tweets. This approach on supervised evaluations shows better performance than all other unsupervised methods with accuracy of 0.82 for Toronto book but it is less than the performance of SkipThought which has an accuracy of 0.83.8. While with the unsupervised evaluation Tasks, it outperforms other methods with an average correlation of 0.67.

Erkan *et al.* (2007) proposed a semi-supervised approach to extract protein interaction depending on sentence dependency parsing. In this approach, dependency parsing tree is constructed for each sentence and the similarity is defined using two kernel functions utilising cosine similarity and edit distance through the paths between protein names in the dependency parsing trees. These functions are used in two supervised algorithms, Support Vector Machines (SVM) and K-nearest-neighbor with their semi-supervised counterparts, transductive SVMs (TSVM) and harmonic functions. They experiment their approach using AIMED corpus and the Christine Brun (CB) corpus. The best F1-score (0.599) is achieved with TSVM edit distance kernel function while the best precision is achieved by SVM with edit distance kernel.

Also, SVM and TSVM have performed better than harmonic functions. With CB dataset, the edit distance function provides a better F1-score than cosine similarity and the best F1-score (0.852) is achieved by TSVM while the highest precision (0.878) is obtained by SVM with cosine similarity.

Le *et al.* (2018) construct a similarity model called ACV-tree based on a modified structured parsing tree. Building the ACV-tree begins with a morphological analysis of each sentence and a part of speech (POS) tagging of words in the sentence. Secondly, words are represented by weighted vectors using word2vec and attention weight for each word is defined by the term frequency-inverse document frequency (TF-IDF). Then relations between words in the sentence are found and words are linked to construct the tree. In this model, the similarity is measured by ACVT kernel that is based on tree kernel. The tree kernel measures similarity between structured trees by computing common substructures in both trees. The proposed model is tested on several datasets including datasets in Semantic Textual Similarity (STS) tasks between 2012 and 2015 except SMT dataset in 2013. The datasets contain pairs of sentences representing news, tweets, images captions, glosses and web forums. It achieves the best Pearson correlation result on 12 out of 19 datasets. For example, ACTV achieves a Pearson correlation of 0.58 on MSRpar dataset, 0.87 on OnWN, 0.79 on headlines dataset, 0.83 on MSRvid dataset, 0.50 on FNWN and 0.79 on answers-students.

He *et al.* (2015) proposed a semantic similarity model that investigates multiple aspects of the input sentences. The sentence is represented using a convolutional neural network to extract features through two sub-networks that work in parallel to process a sentence and they share the same weights. The extracted features are handled by a layer of similarity measurement that connects the sub-networks followed by a layer of similarity output score. In this work, they use structured similarity measure over local regions and multiple types of convolution and pooling. The proposed model is tested using paraphrasing identification task and SemEval task for semantic similarity. They use three datasets including Microsoft Research Paraphrase Corpus (MSRP), Microsoft Video Paraphrase Corpus (MSRVID) and the dataset of Sentences Involving Compositional Knowledge (SICK). The results of the proposed model outperform other work in the state of the art even though it does not use external resources. It achieves an accuracy of 0.786 on MSRP corpus and it provides a Pearson correlation of 0.909 and 0.869 on MSRVID and SICK datasets, respectively.

Sanborn and Skryzalin (2015) utilise recurrent and recursive neural networks with word embedding to provide a model for predicting the semantic similarity between sentence pairs. Their model is trained and tested using SemEval-2015 dataset which consists of 8331 sentence pairs with a semantic similarity score between 0 and 5. The best performance has been achieved by the model that consists of a recurrent neural network with 100-dimensional word embedding trained with 20% dropout and a learning rate of 0.01. This model achieves an F1-score of 0.812 for training and 0.338 for testing but it achieves a Pearson correlation of 0.560.

The authors refer to two reasons to explain the low value of Pearson correlation; the first one is that the authors have converted the task into a classification task by grouping the similarity scores into six categories while the second reason is that the explored models in this research have the challenge of learning tasks simultaneously (Sanborn and Skryzalin, 2015). Table 2 shows a comparison of similarity approaches used for English sentences, while the resulting Pearson correlation of different sentence similarity approaches is shown in Fig. 2.

### 3.3. *Document level*

Although researches that focus on semantic similarity for English documents are almost rare because of its complexity (Liua *et al.*, 2017), Chim and Deng (2007) proposed an approach based on Suffix tree document model and using OHSUMED benchmark corpus and RCV1. This approach has improved the similarity on the average F-measure of six document sets by 80% and it outperforms word Tf-IDF cosine similarity measure by 51% on average F-measure. Also, Huang (2008) applied K-means algorithm with different similarity measures; Euclidean distance, the averaged Kullback–Leibler (KL) divergence, cosine similarity, Jaccard and Pearson coefficient applied to documents clustering. In the evaluation, six datasets are used and the results show that all the measures used, except Euclidean distance, have close and comparable performance for the document clustering task. Pearson coefficient and averaged KL divergence produce more balanced clusters.

Fuzzy rough set-based approach has been proposed by Huang and Kuo (2010) where they investigate a sense–level document representation by applying fuzzy-rough hybrid approach which can reduce the problem of semantic ambiguity. They use semantic similarity measures in a system for cross-language retrieval. In this approach, several linguistic processing is performed based on the language in use. Then, WordNet is used to provide sense disambiguation for words in each document. The retrieved senses represent the main concepts carried in the document and will perform as the elements of the sense-level document representation. They use F-measure and Tversky's notion of similarity to define similarity measures that are based on the operations of fuzzy set. This approach is evaluated using a dataset that consists of 1600 Chinese documents from Taiwan- Panorama with their translated English documents. The proposed approach shows a better efficiency than LSI with a higher average precision of 0.216 for all datasets used in the evaluation.

Boling and Das (2014) use Latent semantic analysis approach (LSA) for semantic similarity between documents and a query. It is applied to ten documents collected from Medline Industries Inc. the results show that terms with similarity of 1.0 are used closely to the query term while terms with similarity less than or equal to 0.7 appear in documents that are not related to the query term.

Liua *et al.* (2017) study the semantic similarity on document level for academic articles by representing an article using several profile information grouped as topic events. The information gathered includes research domain, date, objectives,

Table 2. Similarity approaches for English sentences.

| Technique | Year | Dataset | Results | Class |
|---|---|---|---|---|
| Web-based kernel (Sahami and Heilman, 2006) | 2006 | Retrieved documents from a web search engine | Kernel function provides a high score for the correct answer to a question, even if they do not share common terms. | Distributional |
| Feature-based approach (Li *et al.*, 2006) | 2006 | Brown Corpus: 65 noun definition | Achieves a Pearson correlation coefficient of 0.816 | Semantic-syntactic |
| Supervised and semi-supervised algorithms with kernel functions (Erkan *et al.*, 2007) | 2007 | AIMED corpus and the Christine Brun (CB) corpus | On AIMED, the best F1-score (0.599) with TSVM – edit distance<br>On CB dataset, the best F1-score (0.852) with TSVM – edit distance | Syntactic |
| Corpus-based measure with LCS string matching (Islam and Inkpen, 2005) | 2008 | Two datasets: the same dataset as Li *et al.* (2006) and the paraphrase dataset | Achieves an accuracy of 0.726 with a recall of 0.891 and a precision of 0.74. It provides a correlation of 0.853 on Li *et al.* dataset. | Shallow |
| Disambiguate and walk (feature-based) (Pilehvar *et al.*, 2013) | 2013 | SemEval-2012 dataset | Achieves a correlation of 0.866 for text similarity | Semantic |
| Feature-based approach, Alignment pipeline (Sultan *et al.*, 2014) | 2014 | SemEval STS 2014 datasets | Provides a correlation of 0.82 | Distributional |
| Word embedding (Kenter and de Rijke, 2015) | 2015 | Microsoft Research Paraphrase Corpus dataset | Shows a better performance with accuracy of 0.76 and recall of 0.906 | Distributional |
| Feature-based approach (Taieb *et al.*, 2015) | 2015 | Li *et al.* dataset and Microsoft Paraphrase Corpus | Achieves a correlation of 0.79 with Li *et al.* dataset with a recall of 0.76 and a precision of 0.79. and a recall of 0.994 and a precision of 0.670 with MSPC | Semantic |

Table 2. (*Continued*)

| Technique | Year | Dataset | Results | Class |
|---|---|---|---|---|
| Convolutional neural network (He *et al.*, 2015) | 2015 | Microsoft Research Paraphrase Corpus (MSRP), Sentences Involving Compositional Knowledge (SICK) dataset, Microsoft Video Paraphrase Corpus (MSRVID) | Achieves a Pearson correlation of 0.909 on MSRVID data and a correlation of 0.869 on SICK data. also achieves an accuracy of 0.786 and F1-score of 0.847 on MSRP | Distributional |
| Recursive and recurrent neural networks (Sanborn and Skryzalin, 2015) | 2015 | SemEval-2015: 8331 sentence pairs | The best model achieves an F1-score of 0.812 for training and 0.338 for testing but it achieves a Pearson score of 0.560. | Distributional |
| Feature-based approach (Wali *et al.*, 2017) | 2017 | Li *et al.* dataset and MSPC corpus | Achieves a recall of 0.87 and a precision of 0.87. | Semantic-syntactic |
| Sentence embedding (Pagliardini *et al.*, 2018) | 2018 | Three datasets: the Toronto book corpus, Wikipedia sentences and tweets. | On supervised evaluations, it achieves an accuracy of 0.82 on Toronto book corpus while with the unsupervised Evaluation Tasks it provides an average correlation of 0.67. | Distributional |
| Feature-based approach (Pawar and Mago, 2018) | 2018 | Pilot Short Text Semantic Similarity Benchmark<br>65 noun pairs of R&G benchmark | Yields a good correlation of 0.879 for sentence similarity and provides a correlation of 0.875 for word similarity. | Semantic |
| ACV-tree kernel with word embedding (Le *et al.*, 2018) | 2018 | 2012: MSRpar, MSRvid, SMTeuroparl, OnWN, SMTnews<br>2013: headlines, OnWN, FNWN<br>2014: deft-forum,deft-news, headlines, images, OnWN, tweet-news,<br>2015: answers-forums, answers-students, belief, headlines, images | Achieves a Pearson correlation of 0.58 on MSRpar dataset, 0. 87 on OnWN, 0.79 on headlines dataset, 0.83 on MSRvid dataset, 0.50 on FNWN dataset, and 0.79 on answers-students. | Semantic- syntactic |

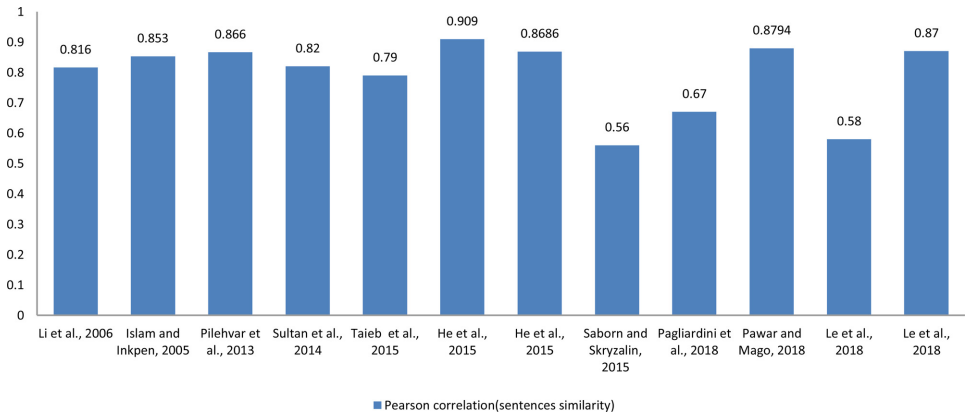


Fig. 2. Pearson correlation results for sentence similarity approaches.

keywords and methodology in order to describe the academic research. Then the similarity is computed between topic events using domain ontology. They construct a document semantic matching annotated corpus that has 1021 paper pairs generated from ACL Anthology Network corpus (Radev *et al.*, 2009). The results of this approach obtain a correlation of 0.559 and achieve an accuracy of 0.768 which is better than LSA based method when the threshold is higher than 0.250 which implies that domain ontology has a good impact on the document semantic similarity.

Madylova and Oguducu (2009) proposed an approach for measuring semantic similarity between documents based on parent and concept vectors where the similarity between documents is computed using cosine similarity. A parent vector is created for each term which appears in the document by finding the term in the taxonomy then follow the IS-A relations of this term to get its corresponding concepts up to the root in the taxonomy. Each term is weighted with a weighting scheme. Then all parent vectors for the terms of a document are merged to create a concept vector for the document. In order to reduce the vector dimensionality, the documents are represented by the 10 most frequent terms. Also, a disambiguation phase is provided to give the most appropriate sense. The documents are clustered based on their cosine similarity. They evaluate their approach using three datasets; the first dataset consists of 2382 documents while the second dataset has 481 documents and the third dataset consists of 1987 documents collected from the Turkish Internet Service Company website. Davies–Bouldin index is used as validity metric for clustering where low value means good clustering. The proposed method provides better results with Davies–Bouldin index value of 1.71 for the first dataset, 1.41 on the second dataset and 1.56 on the third one. Table 3 shows a comparison of the similarity approaches used for English documents. The results of the document similarity approaches that use F-measure to evaluate their work are presented in Fig. 3.

Table 3. English document similarity approaches.

| Technique | Year | Dataset | Results |
|---|---|---|---|
| Suffix tree document model (Chim and Deng, 2007) | 2007 | benchmark corpus OHSUMED and RCV1 | Improves the average F-measure of 6 document sets by 80% and outperforms Tf-IDF cosine similarity by 51% on average F-measure. |
| K-means algorithm with 5 measures (Huang, 2008) | 2008 | 20 news, CACM/CISI/ CRANFIELD/ MEDLINE, hitech, Reuters-21578, TREC5 and TREC6, WebACE, Web Knowledge Base. | Pearson coefficient and averaged KL divergence produce more balanced clusters. |
| Parent and concept vectors with cosine similarity (Madylova and Oguducu, 2009) | 2009 | Three datasets: dataset1 consists of 2382 documents; dataset2 has 481 documents and dataset3 consists of 1987 documents. | Davies-Bouldin index value of 1.71 for the first dataset, 1.41 on the second dataset and 1.56 on the third dataset. |
| Fuzzy set and rough set based approach (Huang and Kuo, 2010) | 2010 | 1,600 Chinese documents from Taiwan- Panorama with their translated English documents. | Achieves average precision of 0.216 for all used datasets and an average precision of 0.35 higher than that of LSI for all the datasets while the best F-measure is 0.783. |
| Latent semantic analysis (Boling and Das, 2014) | 2014 | Ten documents collected from Medline Industries, Inc. | Terms with similarity of 1.0 indicate that they are used closely to the query term while terms with similarity <= 0.7 appear in documents that are not related to the query term. |
| Topic Event and ontology-based method (Liua *et al.*, 2017) | 2017 | Annotated corpus with 1,021 paper pairs that are generated from ACL Anthology Network corpus. | Achieves a correlation of 0.559 and provides an accuracy of 0.768 while the best F1-score is 0.639 |

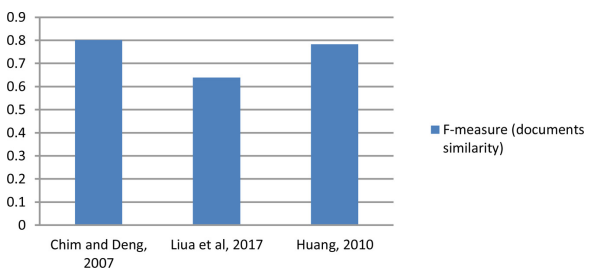


Fig. 3. F-measure results for document similarity approaches.

## 4. Semantic Similarity for Arabic Text

### 4.1. *Word level*

All existing word similarity measures are conducted for English texts while very rare measures have been developed specifically for Arabic and they also depend on

English measures. For example, the work of Almarsoomi *et al.* (2013) represents an approach for measuring semantic similarity between two Arabic words which originally depends on Li similarity measure (Li *et al.*, 2003) and the evaluation has been on a dataset, generated by the authors, consists of 70 pairs of words that vary between low, medium and high similarity. They obtain good results with a Pearson correlation of 0.894 compared to the human average rating. Also, Froud *et al.* (2012) propose another approach for measuring semantic similarity between Arabic words based on Latent Semantic Analysis and the evaluation has been on two datasets from Saudi Press Agency where the first one consists of 252 documents from different categories and the other consists of 257 documents from the same category. They mainly test the effect of using stemming and light stemming on the similarity between words and their results show that light stemming outperforms stemming approach.

The study of Nababteh and Deri (2017) investigates the semantic similarity for Arabic words using seven WordNet-based similarity measures; Wu and Palmer (1994), Path measure (Michelizzi, 2005), LCH measure (Leacock and Chodorow, 1998), Li measure (Li *et al.*, 2003), AWSS measure (Almarsoomi *et al.*, 2013), ResMeng measure (Resnik, 1995) (Meng *et al.*, 2012) and Zhou measure (Zhou, 2008). These measures are selected from path-based measures, information content measures, and hybrid measures. The experiment is applied using only 40 Arabic word pairs from the AWSS benchmark because not all the words in the benchmark have been found in AWN. The results show that Wu and Palmer measure has the best performance over other measures since it provides a correlation of 0.94 and a mean square error MSE value of 1.64 while the path measure has the lowest correlation value of 0.75 and the highest MSE value of 16.038. Table 4 shows a comparison of the approaches used for Arabic word similarity while Fig. 4 shows the

Table 4. Arabic words similarity approaches.

| Technique | Year | Dataset | Results |
|---|---|---|---|
| Li semantic similarity measure using Arabic knowledge base (Almarsoomi *et al.*, 2013) | 2013 | A dataset of 70 pairs of words selected randomly using high, medium, and low similarity. (AWSS) | The AWSS measure obtains a Pearson correlation of 0.894. |
| Latent Semantic Analysis (Froud *et al.*, 2012) | 2012 | Two datasets from Saudi Press Agency | Light Stemming approach outperforms Stemming approach. |
| Seven WordNet-based measures applied to Arabic words (Nababteh *et al.*, 2017) | 2017 | 40 words from AWSS dataset | The best correlation is achieved by Wu and Palmer measure (0.94). |

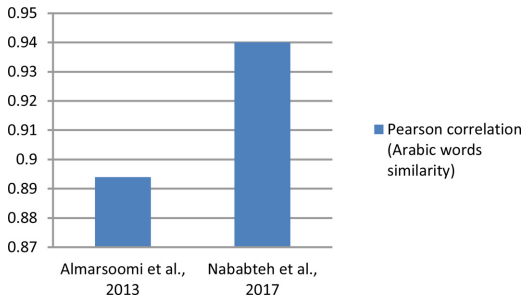


Fig. 4. Pearson correlation results for Arabic words similarity approaches.

results achieved by similarity approaches where the evaluation is carried out in terms of Pearson correlation.

### 4.2. *Sentence level*

At the level of Arabic sentence similarity, Alzahrani (2016) proposed three approaches for cross-lingual semantic similarity between Arabic and English sentences; the first approach is a dictionary-based translation with maximum similarity which correlates each Arabic term with the terms in English text then the similarity between words which is computed using Wu and Palmer metric (1994) while the similarity between sentences is the average sum of their content words similarity. The second approach is a machine translation-based approach where two vectors are constructed, one for noun and the other for verbs with the entry value as the maximum word semantic similarity found between the word and other words of the same POS in the other sentence. The similarity score is computed as the sum of nouns and verbs vectors cosine similarity. The third approach is a machine translation-based term vector similarity where the similarity score is computed by the cosine similarity between two terms vectors. The entry value in the term vector is the maximum semantic similarity between the corresponding word and other sentence words. The proposed approaches are evaluated using selected sentence pairs from the set proposed by Li *et al.* (2006) and a benchmark for testing cross-language similarity constructed by the author. The results show that machine translation-based term vector similarity algorithm outperforms other algorithms and achieves a high Pearson correlation rate of 0.8657.

Also, in the sentence level, Kadhem and AbdAlameer (2017) proposed a hybrid similarity approach using semantic similarity measure, cosine similarity, and N-gram approach. They evaluate their approach using a constructed SemanticNet network that stores Arabic keywords for computer science field and compare the results of the hybrid approach with each one of its component approaches. The results of the Hybrid approach outperform its component methods in terms of F1 score, precision and recall. Moreover, two researches have been conducted by Nagoudi and Schwab (2017) on the semantic similarity between Arabic sentences using word embedding

technique with a proposed approach for word alignment and different weighting functions for words vectors. The sentence vector is represented as the sum of its content words vectors. They have evaluated their approach using 750 pairs form Microsoft Research Video Description Corpus translated into Arabic while in the other research (Nagoudi *et al.*, 2018) they use four datasets from SemEval-2017 that have 2412 pairs of sentences and they apply cross-lingual similarity between Arabic and English sentences. The results of applying the similarity between sentence vectors without weighting obtain a correlation of 0.723 while using IDF weighting and POS weighting achieve a correlation of 0.782 and 0.796, respectively. The bag of words alignment approach with mixed weighting provides a correlation of 0.773 while weighting aligned words approach, which aligns the most similar words in the sentences, provides a correlation of 0.737.

Wali *et al.* (2017) proposed a feature-based approach that consists of three phases; the first phase is preprocessing where no stop words are removed, words are reduced to their stems and punctuation marks are removed. The second phase is the similarity scoring where lexical, semantic and syntactico-semantic similarity scores are considered based on the content of Lexical Markup Framework (LMF) standardised dictionaries (Francopoulo, 2013). The lexical score is computed using Jaccard coefficient based on the number of common words between the two sentences while the score of semantic similarity is measured as the cosine similarity between semantic vectors of the sentences. The content of the semantic vector is one if the word exists in both sentences and it is the maximum similarity score between two different words. After finding the set of synonyms for each word, the similarity score is measured using Jaccard coefficient based on the common synonyms between two words. The third phase of the proposed approach is the supervised learning in which the appropriate coefficients are determined for the similarity scores. They evaluate their method using a dataset which consists of 690 sentence pairs collected from Al-Wassit dictionary, Al-Muhit, Lissan Al-Arab and Tj Al-Arous. The proposed method achieves a Pearson correlation of 0.92. It is also resulted in a good Precision of 0.88 and a Recall of 0.83.

Mahmoud and Zrigui (2019) proposed a deep learning method to learn sentence embedding and measure the semantic relatedness between sentences. They investigate the efficiency of two different deep neural networks (i.e. CNN and LSTM) in extracting the appropriate features of sentences and capturing the words dependencies without relying on the syntactic and semantic structure of the language. The proposed approach consists of three phases; preprocessing, features extraction and similarity computation. The preprocessing phase aims at removing diacritics and eliminating the less useful words. In the feature extraction phase, words are represented as vectors using pre-trained embeddings. In the final phase, similarity between sentences is estimated. To conduct their experiments, they propose the construction of an automatic corpus-based on Open Source Arabic Corpus (OSAC). The results of LSTM model achieve a better rate of semantic similarity where the accuracy using LSTM model reaches 0.83 while CNN achieves an accuracy of 0.79. Table 5 provides a comparison between sentence similarity approaches used for

Table 5. Similarity approaches for Arabic sentences.

| Technique | Year | Dataset | Results | Class |
|---|---|---|---|---|
| A dictionary-based translation with maximum similarity, and Machine translation with feature-based similarity method. (Alzahrani, 2016) | 2016 | Selected sentences from the set proposed by Li *et al.* (2006) and a constructed benchmark for testing cross-language similarity. | Machine Translation with term vector semantic similarity obtains higher correlation of 0.866. | Syntactic-semantic |
| Hybrid similarity measure (Semantic similarity measure, Cosine similarity measure and N-gram) (Kadhem and AbdAlameer, 2017). | 2017 | A constructed SemanticNet that stores Arabic keywords for computer science field. | Precision over 0.90. | Shallow |
| Word embedding and feature-based (Nagoudi and Schwab, 2017) | 2017 | 750 pairs from Microsoft Research Video Description Corpus translated into Arabic. | The method of no weighting obtains a correlation rate of 0.723 while IDF-weighting and POS-weighting achieve a correlation of 0.782 and 0.797, respectively. | Distributional |
| Feature-based approach using semantic and syntactic-semantic knowledge (Wali *et al.*, 2017) | 2017 | 690 sentence pairs collected from Al-Wassit dictionary, Al-Muhit, Lissan Al-Arab and Tj Al-Arous. | Achieves a precision of 0.88 and a recall of 0.83. Also, it obtains a Pearson correlation of 0.92. | Syntactic-semantic |
| Weighting Aligned Words and Alignment Bag-of-Words with three weighting functions (Nagoudi *et al.*, 2018) | 2018 | Four datasets drawn from the STS shared task SemEval-2017. | The mixed weighted method with Alignment Bag-of-Words provides a correlation of 0.774 and the Weighting Aligned Words method obtains a correlation rate of 0.738. | Distributional |
| Pre-trained word embedding with CNN and LSTM. (Mahmoud and Zrigui, 2019) | 2019 | OSAC as source corpus to construct their paraphrased corpus. | LSTM model achieves an accuracy of 0.83 while CNN achieves an accuracy of 0.79. | Distributional |

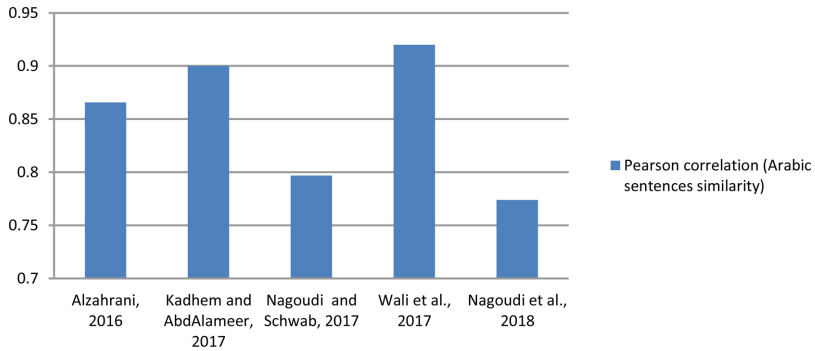


Fig. 5. Pearson correlation for Arabic sentences similarity approaches.

Arabic sentences while Fig. 5 shows the results of these approaches in terms of precision.

### 4.3. *Document level*

In the Arabic Document similarity level, Selamat and Ismail (2008) apply two neural networks models; Self-organising Map (SOM) and Growing Hierarchical Self-organising Map (GHSOM) for clustering documents. In this work, a preprocessing step is applied for documents where stop words are removed and word stems are extracted. The mapping techniques are evaluated using News documents collected from Al-jazeera news, Al-sharq Al-awsat, and others. Arabic documents are translated then document vectors are constructed to be used in the mapping techniques. The results show a precision of 0.87 for SOM technique and 0.93 for GHSOM mapping technique. Also, Froud *et al.* (2010) investigate different similarity measures with document clustering: Jaccard Coefficient, Cosine similarity, Pearson Correlation Coefficient, Averaged Kullback–Leibler Divergence and Euclidean Distance. They conduct their experiment on the Corpus of Contemporary Arabic (CCA) which has 12 several categories. Stemming is applied to words which have resulted in a smaller representation of documents and provided fast clustering. The results show that stemming with Jaccard measure performs better in generating more coherent clusters with a considerable purity score. Another research, conducted by Al-Ramahi and Mustafa (2012), utilises the Dice's similarity measure with bi-gram word based and document based for detecting Arabic document matching. To evaluate this approach, two datasets are used; the first one has 104 course descriptions and the other has 30 course descriptions selected from Jordanian Universities. The N-gram document matching achieves accuracy over 0.80.

Soori *et al.* (2013) use the concept of Lempel Ziv compression for Plagiarism detection. The experiment is applied to 150 documents and 330 paragraphs; 159 paragraphs extracted from the source documents and 171 paragraphs from Al-Khaleej corpus. The paragraphs are divided into chunks then Lempel Ziv is applied to detect the plagiarised documents. The proposed method detects 71.42% of the

plagiarised documents, 28.85% of the partially plagiarised documents and 100% of non-plagiarised documents.

Latent Semantic Analysis with N-gram approach is used in the work of Hussein (2016) and evaluated on 30 students' Tutor Marked Assignment (TMA) answer documents of the Kuwait Civilisation History Course. In this work, a morphological analysis is performed on words with POS tagging and stemming. Since the proposed approach depends on Natural language processing, it outperforms Plagiarism Checker X in the case of bigram and trigram.

In the work of Awajan (2016), Vector Space Model (VSM) is used with resources such as Arabic WordNet and Name Entities (Gazatters) to reduce the dimensionality of Arabic documents. The dataset consists of six different categories of texts from Middle East news, Word news, Business, sports, science, Arts and culture. The preprocessing step uses Al-Khalil morph-syntactic system and Stanford POS tagger. In this approach, the size of the text representation is reduced by 27% compared to the stem-based vector space model and reduced by 50 % compared to the traditional bag-of-words model.

Mahmoud and Zrigui (2019) proposed a deep learning-based technique to detect paraphrased documents that have the same meaning. They have used word2vec to extract relevant features then the sentence vectors are generated by averaging their content words vectors. Then, convolutional neural network (CNN) is used to learn more contextual information and to produce the semantic similarity score. The convolution layer in the CNN is used to extract features from documents and transform them to a proper form while the pooling layer uses max function to produce a reduced semantic vector then the comparator layer computes the cosine similarity between sentences vectors and convert the result into probability distribution. In order to conduct their experiment, they generate a paraphrased corpus from OSAC corpus by replacing a word by its synonym or most similar word with the same part of speech of the original word. The proposed method achieves a precision of 0.85 and a recall of 0.868. Table 6 provides a comparison between Arabic

Table 6. Arabic document similarity approaches.

| Technique | Year | Used dataset | Results |
|---|---|---|---|
| Self-organising Map (SOM) and Growing Hierarchical Self-organising Map (GHSOM) (Selamat and Ismail, 2008) | 2008 | News documents collected from Al-jazeera news, Al-sharq Al-awsat. | Precision of 0.87 for SOM and 0.93 for GHSOM. |
| Document clustering with: Jaccard Coefficient, Cosine similarity, Pearson Correlation Coefficient, Averaged Kullback-Leibler Divergence and Euclidean Distance. (Froud *et al.*, 2010) | 2010 | Corpus of Contemporary Arabic (CCA) | Purity of 0.64 with khoja stemmer.<br>Purity of 0.53 with Larkey stemmer. |

Table 6. (*Continued*)

| Technique | Year | Used dataset | Results |
|---|---|---|---|
| Dice's similarity measure with bi-gram (Al-Ramahi and Mustafa, 2012) | 2012 | Course descriptions | Accuracy level goes beyond 0.80. |
| N-gram (Lempel Ziv compression) (Soori *et al.*, 2013) | 2013 | Al-Khaleej corpus | 71.42% of plagiarised documents are detected. |
| Vector Space Model (VSM) (Awajan, 2016) | 2016 | Six different categories of texts: Middle East news, World news, Business, Sport, Science, Technology, Arts and Culture. | The size of text representation is reduced by 27% compared to stem-based vector and reduced by 50% compared to traditional bag-of-words model. |
| N-gram and Latent Semantic Analysis (Hussein, 2016) | 2016 | 30 students' Tutor Marked Assignment (TMA) answers | With N-gram equals 2 and 3 the max difference in estimating pairwise similarity is 4.78%. |
| Sentence embedding with Convolutional Neural Network (Mahmoud and Zrigui, 2019) | 2019 | (OSAC) Source and Paraphrased corpora (22,429 documents) | Achieves a precision of 0.85, a recall of 0.868 and an F-measure of 0.858 with sent2vec and CNN. |

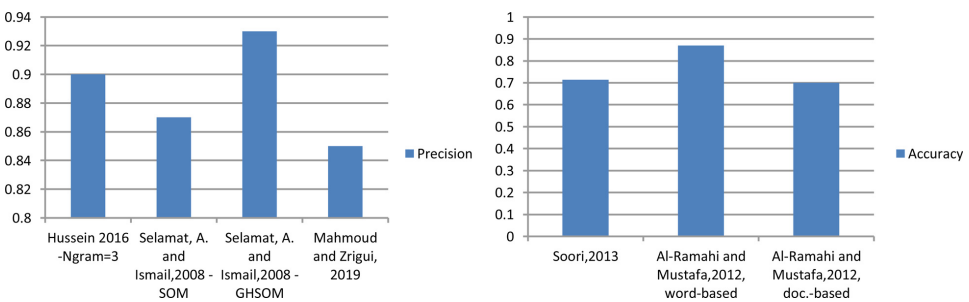


Fig. 6. Precision and accuracy of similarity approaches in Arabic documents.

document similarity approaches and Fig. 6 shows the precision and accuracy achieved by these approaches.

## 5. Benchmarks

Several benchmarks have been used in previous researches to evaluate their work. In this section, we describe and compare these benchmarks in terms of their size, scale of similarity, level of similarity, number of human annotators and the best results achieved by similarity approaches where Pearson correlation ($r$) or F-Measure ($F$) results are presented as shown in Table 7.

In the evaluation of words similarity methods, a number of popular datasets have been generated and made available to the research community, such as Rubenstein

Table 7. Benchmarks used in similarity approaches.

| Benchmark | Year | Similarity level | Size | Language | Scale | Human annotators | Best similarity technique in literature | Result |
|---|---|---|---|---|---|---|---|---|
| R&G | 1965 | word | 65 pairs | English | [0,4] | 51 | Co-occurrenceDouble checking (Chen *et al.*, 2006) | $r = 0.849$ |
| M&C | 1991 | word | 30 pairs | English | [0,4] | 38 | Feature-based approach (Jiang *et al.*, 2015) | $r = 0.827$ |
| WordSim353 | 2002 | word | 353 pairs | English | [0,10] | 13–16 | Knowledge-based approach (Taieb *et al.*, 2013) | $r = 0.72$ |
| MTurk287 | 2011 | word | 287 pairs | English | [1,5] | 23 | knowledge-based approach (Taieb *et al.*, 2013) | $r = 0.654$ |
| AWSS | 2012 | word | 70 pairs | Arabic | [0,4] | 60 | Knowledge based (Almarsoomi *et al.*, 2013) | $r = 0.894$ |
| Pilot Short Text Semantic Similarity (Li *et al.*, 2006) | 2006 | sentence | 65 pairs | English | [0,4] | 32 | Feature-based approach (Pawar and Mago, 2018) | $r = 0.879$ |
| Microsoft Research Paraphrase Corpus (MSRP) | 2005 | sentence | 5800 pairs | English | 0,1 | 3 | Convolutional neural network (He *et al.*, 2015) | $r = 0.909$ |
| SemEval-2012 | 2012 | sentence | 2234 training 3108 testing | English | N/A | N/A | Disambiguate and walk (feature-based) (Pilehvar *et al.*, 2013) | $r = 0.866$ |
| SemEval-2014 | 2014 | sentence | 3750 testing | English | N/A | N/A | Feature-based approach, Alignment pipeline (Sultan *et al.*, 2014) | $r = 0.82$ |
| SemEval-2015 | 2015 | sentence | 8331 pairs | English | [0,5] | 5 | Recurrent neural network (Sanborn and Skryzalin, 2015) | $F = 0.812 -$ training $F = 0.338 -$ testing |
| SemEval-2017 | 2017 | sentence | 2412 pairs | English/ Arabic | [0,5] | 5 | Weighted method with Alignment Bag-of-Words (Nagoudi *et al.*, 2018) | $r = 0.774$ |
| OHSUMED | 1994 | document | 50,216 | English | N/A | N/A | Suffix tree document model (Chim and Deng, 2007) | $F = 0.80$ |
| OSAC | 2010 | document | 22,429 | Arabic | N/A | N/A | Sentence embedding with Convolutional Neural Network (Mahmoud and Zrigui, 2019) | $F = 0.858$ |

and Goodenough dataset (RG-65), which consists of 65 word pairs, where each pair has a similarity score obtained by the judgments of 51 undergraduate students (Rubenstein and Goodenough, 1965). Another benchmark is the Miller and Charles 30 (MC-30) dataset which consists of 30 word pairs that have been originally taken from the Rubenstein and Goodenough dataset. It only includes word pairs that are presented in WordNet. Each pair of words has a similarity score of 0 to 4 representing the judgments of 38 human subjects (Miller, 1991). Also, WordSimilarity-353 is a collection of two sets of word pairs where the first set consists of 153 word pairs in which each pair has a similarity score assigned by 13 human subjects, while the second set consists of 200 word pairs with a similarity score assigned by 16 human subjects (Finkelstein *et al.*, 2002).

MTurk dataset differs from other datasets in the annotation process where human ratings have been obtained from Amazon Mechanical Turk workers. This dataset contains 287 word pairs where each pair has a similarity score provided by an average of 23 MTurk workers. The workers have evaluated pairs on a scale of 1–5 (Radinsky *et al.*, 2011). However, for Arabic words, Almarsoomi *et al.* (2012) have produced a benchmark called (AWSS) dataset consisting of 70 pairs of randomly selected Arabic words with high, medium, and low similarity. Each pair has been assigned to 60 Arabic native speakers to give a semantic similarity score between 0.0 and 4.0, and then the average of those ratings is calculated (Almarsoomi *et al.*, 2012).

Whereas two popular benchmarks have been used in the evaluation of sentence similarity methods; the first is Microsoft Research Paraphrase Corpus (MSRPC), which consists of 5800 sentence pairs extracted from web news with more detailed information on those sentences, such as the author and the source of the sentence. Each sentence pair has a binary tag assigned by three human annotators based on the semantic similarity of each pair (Dolan *et al.*, 2005; Alian and Awajan, 2020). The second is Pilot Short Text Semantic Similarity Benchmark provided in 2006 by Li *et al.* (2006) and consists of 65 sentence pairs collected from the Collins Cobuild Dictionary with some modifications. Each pair has a similarity score which represents an average of 32 human ratings (Li *et al.*, 2006). The Association for Computational Linguistics has also produced a number of datasets for several years, such as SemEval-2012, SemEval-2014, SemEval-2015 and SemEval-2017. Each of these datasets is collected from other corpuses and benchmarks such as Microsoft Research Video Description Corpus, Microsoft Research Paraphrase Corpus and others. However, SemEval-2015 consists of all pairs of sentences for similar tasks in 2012, 2013 and 2014. Thus, it contains 8331 pairs of sentences with a semantic similarity between 0 and 5 given as the mean of five annotators from Amazon Mechanical Turk (Sanborn and Skryzalin, 2015).

SemEval-2017 dataset has been used to evaluate the Cross-lingual Arabic-English Task. It consists of 2412 pairs of sentences collected from a number of resources. Sentence pairs have been assigned to five annotators to give a similarity score to each

pair, which is a float number between 0 and 5. The mean of the five annotators' judgments is then calculated to be the score of the sentence pair (CER *et al.*, 2017).

For document benchmarks, the OHSUMED dataset is used to evaluate documents similarity approaches as shown in the literature. It consists of 50,216 documents representing medical abstracts collected from the Medical Information Database (MEDLINE) that included 348,566 references from 270 medical journals in the period (1987–1991) (Hersh *et al.*, 1994). While in Arabic documents, Open Source Arabic Corpora (OSAC) is used by many researchers to evaluate their work. OSAC consists of 22,429 documents in 10 categories such as economics, history, sports, health, etc. (Mahmoud and Zrigui, 2019).

## 6. Summary

Text similarity for the Arabic or English language can be categorised into three levels, namely, document, sentence, and word similarity. The smallest part of any text is a word, and a group of words creates a sentence. Therefore, several studies based their sentence similarity methods on word similarity. The group of sentences forms a document. Several approaches for document similarity have been constructed based on word and sentence similarity.

The techniques used for semantic similarity between two texts can be classified into the following categories: corpus-based method (e.g. LSA and Hyperspace Analogues to Language), descriptive feature-based method that represents a text by using a set of syntactic and semantic features, and word embedding method that constructs a vector for a sentence by using its content word vectors. Paraphrasing and sense representation techniques are also included.

The level of analysis in previous work varied from lack of morphological analysis to partial morphological analysis, such as stemming, POS parsing, and dependency parsing. Figure 7 shows the techniques used in previous work in the last 10 years. The number of studies that used each technique is also shown in the figure. Among all the techniques, the feature-based approach has the largest number of studies.

After reviewing previous studies on semantic similarity for English and Arabic texts (five of them investigated English word similarity between 2005 and 2015), we

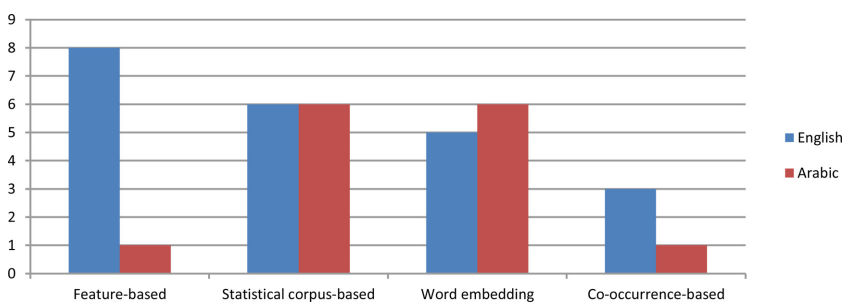


Fig. 7. Number of studies on similarity techniques.

found that 14 approaches have been proposed for English sentence similarity during 2006–2018, and six similarity approaches have been introduced for English documents between 2007 and 2017. However, the number of studies for Arabic is smaller than that for English. Three studies have investigated the similarity of Arabic words between 2013 and 2017, six sentence similarity approaches have been proposed from 2008 to 2019, and seven similarity approaches have been introduced for Arabic documents from 2008 to 2019. Moreover, we discovered that little work has been performed for the semantic similarity of Arabic sentences. The limited available studies depended on a small dataset of sentences when evaluating their approaches and did not consider stop words. No work has studied the effect of chunks on measuring semantic similarity for Arabic sentences. Moreover, words with multiple meanings were not considered in previous work on Arabic semantic similarity.

The statistical evaluation metrics used in the previous studies were precision, recall, accuracy, and Pearson correlation. Many techniques obtained a precision above 70%, and other techniques that use Pearson correlation achieved a correlation range of 0.73–0.89.

Table 8 shows that the best correlations were achieved by the feature-based approach in different aspects of similarity. For English text, the best correlations for word and document similarities (0.82 and 0.55, respectively) were obtained by the feature-based approach. For sentence similarity, the best results were provided by the feature-based and word embedding approaches with a correlation of 0.87. However, for Arabic text, the best results on measuring Arabic word similarity were produced by the Wu and Palmer measure with a correlation value of 0.94. The best correlation for sentence similarity was provided by the feature-based approach with a correlation of 0.92. For document similarity, the researchers used different measures for evaluation, but the best precision was achieved using multilevel neural networks (precision of 0.93). Determining the effectiveness of Arabic text similarity methods in other NLP applications is difficult because the majority of these methods use a specific set of documents or sentences collected for particular research purposes, and these sets are unavailable online for other researchers.

Table 9 shows a comparison of semantic similarity of Arabic and English texts in terms of the utilised benchmarks, evaluation measures, morphological features, preprocessing steps required for similarity approaches, and use of deep learning or neural networks in estimating semantic similarity.

Table 8. Best results for semantic similarity approaches.

| Similarity level | Language | Approach | Correlation |
|---|---|---|---|
| Word | English | Feature-based | 0.82 |
| Sentence | English | Feature-based and word embedding | 0.87 |
| Document | English | Feature-based | 0.55 |
| Word | Arabic | Wu and Palmer | 0.94 |
| Sentence | Arabic | Feature-based | 0.92 |

Table 9. Comparing semantic similarity of Arabic and English texts.

| Similarity level | Semantic similarity of Arabic text | | | Semantic similarity of English text | | |
|---|---|---|---|---|---|---|
| | Word | Sentence | Document | Word | Sentence | Document |
| Benchmarks | AWSS | SemEval-2017 | OSAC | M&C, R&G, WordSimilarity-353 | MSRP, SemEval 2012, 2014, 2015, and 2017 | OHSUMED, Reuters-21578, 20news |
| Morphological Features | Stemming | POS, Stemming | POS, Stemming | — | POS | POS |
| Preprocessing | — | Tokenisation, stop words removal | Tokenisation, stop words removal, removing diacritics | — | Tokenisation, stop words removal | Tokenisation, stop words removal |
| Evaluation measures | Pearson correlation | Pearson correlation, precision | Accuracy, precision, F-measure | Recall, precision, Pearson correlation | F-measure, recall, precision, Pearson correlation | F-measure, precision, Pearson correlation, accuracy |
| Usage of deep learning and neural networks | — | Limited | Limited | — | Satisfactory | Satisfactory |

## 7. Conclusion

Semantic similarity is a crucial research area that has been eliciting researchers' attention for years because it is a major task in many NLP applications. In this survey, three levels of similarity (word, sentence, and document similarity) were explained and discussed. The methods used in extant studies to investigate semantic similarity measurement at each level were classified into four groups, namely, co-occurrence-based, statistical corpus-based, descriptive feature-based, and word embedding techniques.

The results showed that in the case of English text similarity, the best results were achieved by the feature-based approach for word and document similarity; for sentence similarity, the hybrid approach that combines feature-based and word embedding techniques produced the best correlation. In the case of Arabic word similarity, the WordNet-based semantic similarity measure (Wu and Palmer) achieved the best results, whereas the best Arabic sentence similarity results were provided by the feature-based approach. With regard to Arabic document similarity, the best results were achieved by multilevel neural networks.

After analyzing the results of previous research, we conclude that the feature-based approach provides the best correlation value for all levels of semantic similarity. At the sentence similarity level, the results are improved when the feature-based technique is implemented with word embedding.

## References

Achananuparp, P, X Hu and SH Xiajiong (2008). The evaluation of sentence similarity measures. In *Data Warehousing and Knowledge Discovery*, JEI-Y Song (ed.), Lecture Notes in Computer Science, Vol. 5182, pp. 305–316. Berlin Heidelberg: Springer.

Agirrea, E, C Baneab, C Cardiec, D Cerd, M Diabe, A Gonzalez-Agirrea, W Guo, I Lopez-Gazpio, M Maritxalar, R Mihalceab, G Rigau, L Uria and J Wiebe (2015). SemEval-2015 Task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proc. 9th Int. Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, pp. 252–263.

Alian, M and A Awajan (2020). Paraphrasing identification techniques in English and Arabic texts. In *Proc. 11th Int. Conf. Information and Communication Systems*, Irbid, Jordan, pp. 155–160.

Almarsoomi, FA, JD O'Shea, Z Bandar and K Crockett (2013). AWSS: An algorithm for measuring Arabic word semantic similarity. In *2013 IEEE Int. Conf. Systems, Man and Cybernetics (SMC)*, pp. 504–509. Manchester, UK.

Almarsoomi, FO (2012). Arabic word semantic similarity. *ICALLL (WASET)*, Vol. 70, Dubai, UAE, pp. 87–95.

Al-Ramahi, MA and SH Mustafa (2012). N-gram-based techniques for Arabic text document matching; case study: Courses accreditation.*Abhath AL-Yarmouk: Basic Sciences & Engineering*, 21(1), 85–105.

Alzahrani, S (2016). Cross-language semantic similarity of Arabic-English short phrases and sentences.*Journal of Computer Sciences*, 12(1), 1–18.

Aouicha, MB and MAH Taieb (2015). G2WS: Gloss-based WordNet and Wiktionary semantic Similarity measure. In *2015 IEEE/ACS 12th Int. Conf. Computer Systems and Applications (AICCSA)*, pp. 1–7. Marrakech, Morocco.

Awajan, A (2016). Semantic similarity based approach for reducing Arabic texts dimensionality.*International Journal of Speech Technology*, 19(2), 191–201.

Boling, C and K Das (2014). Semantic similarity of documents using latent semantic analysis. In *Proc. National Conf. Undergraduate Research (NCUR) 2014*, pp. 1083–1092. University of Kentucky, Lexington, KY.

CER, DD-G (2017). SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *The 11th Int. Workshop on Semantic Evaluation (SemEval-2017)*, Association for Computational Linguistics, Vancouver, Canada, pp. 1–14.

Cha, S (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4), 300–307.

Chen, HH, M Lin and Y Wei (2006). Novel association measures using web search with double checking. In *21st Int. Conf. Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 1009–1016. Sydney, Australia.

Chim, H and X Deng (2007). A new suffix tree similarity measure for document clustering. In *The Int. World Wide Web Conf. Committee (IW3C2)*, pp. 121–130.

Dagan, I, O Glickman and B Magnini (2005). The PASCAL reconising textual entailment challenge. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment.* MLCW 2005. Lecture Notes in Computer Science, J Quiñonero-Candela, I Dagan, B Magnini, F d'Alché-Buc (eds.), Vol. 3944, pp. 177–190. Berlin, Heidelberg: Springer.

Dolan, BB (2005). *Microsoft Research Paraphrase Corpus.* Microsoft Research.

Erkan, G, A Ozgur and DR Radev (2007). Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proc. 2007 Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 228–237. Prague, Czech Republic.

Finkelstein, LG (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1), 116–131.

Francopoulo, G (2013). *LMF Lexical Markup Framework.*New York: Wiley.

Froud, H, R Benslimane, A Lachkar, A Lachkar and S AlaouiOuatik (2010). Stemming and similarity measures for Arabic documents clustering. In *5th Int. Symp. I/V Communications and Mobile Network*, Rabat, pp. 1–4.

Froud, H, A Lachkar and SA Ouatik (2012). Stemming versus Light Stemming for measuring the simitilarity between Arabic Words with Latent Semantic Analysis model. *2012 Colloquium in Information Science and Technology*, pp. 69–73. Fez, Morocco.

Gomaa, WF (2013). A survey of text similarity approaches. *Journal of Computer Applications*, 68(13), 0975–8887.

Hatzivassiloglou, V, JL Klavans and E Eskin (1999). Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *The Joint SIGDAT Conf. Empirical Methods in Natural Language Processing and Very Large Corpora.*

He, H, K Gimpel and J Lin (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proc. 2015 Conf. Empirical Methods in Natural Language Processing*, pp. 1576–1586. Lisbon, Portugal.

Hersh, WB (1994). Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *The 17th Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Dublin, Ireland, pp. 192–201.

Huang, HH and YH Kuo (2010). Cross-lingual document representation and semantic similarity measure: A fuzzy set and rough set based approach. *IEEE Transactions on Fuzzy Systems*, 18(6), 1098–1111.

Huang, M (2008). Similarity measures for text document clustering. In *Proc. New Zealand Computer Science Research Student Conf. 2008*, pp. 49–56. Christchurch, New Zealand.

Hussein, AS (2016). Visualizing document similarity using N-grams and latent semantic analysis. In *SAI Computing Conf.*, London, UK, pp. 269–279.

Islam, A and D Inkpen (2005). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, 2(2), 25.

Jiang, Y, X Zhang, Y Tang and R Nie (2015). Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. *Information Processing and Management*, 51, 215–234.

Kadhem, SM and AQ AbdAlameer (2017). Finding the similarity between two Arabic texts. *Iraqi Journal of Science*, 58(1), 152–162.

Kenter, T and M de Rijke (2015). Short text similarity with word embeddings. In *24th ACM Int. Conf. Information and Knowledge Management (CIKM '15)*, Melbourne, Australia, pp. 1411–1420.

Kintsch, W, TK Landauer and PW Foltz (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2–3), 285–307.

Landauer, TK, D Laham, B Rehder and ME Schreiner (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In *19th Annual Meeting of the Cognitive Science Society*, pp. 412–417. Stanford University, Palo Alto, California.

Le, Y, Z Wang, Z Quan, J He and B Yao (2018). ACV-tree: A new method for sentence similarity modeling. In *Proc. Twenty-Seventh Int. Joint Conf. Artificial Intelligence (IJCAI-18)*, pp. 4137–4143. Stockholm, Sweden.

Leacock, C and M Chodorow (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database*, 49(2), 265–283.

Li, Y, Z Bandar and D McLean (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4), 871–882.

Li, Y, ZA Bandar and D McLean (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4), 871–882.

Li, Y, D McLean, ZA Bandar, JD O'Shea and K Crockett (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8), 1138–1150.

Liua, M, B Langa and Z Gu (2017). Calculating semantic similarity between academic articles using topic event and ontology. arXiv:1711.11508, Cornell University.

Madylova, A and SG Oguducu (2009). A taxonomy based semantic similarity of documents using the cosine measure. In *24th Int. Symp. Computer and Information Sciences*, Guzelyurt, pp. 129–134.

Mahmoud, AZM (2019). Deep neural network models for paraphrased text classification in the Arabic language. In *Natural Language Processing and Information Systems. NLDB 2019.*, MF Métais (ed.), Lecture Notes in Computer Science Vol. 11608, pp. 3–16. New York: Springer.

Mahmoud, AZ (2019). Sentence embedding and convolutional neural network for semantic textual similarity detection in Arabic language. *Arabian Journal for Science and Engineering*, 44, 9263–927.

Meng, L, J Gu and Z Zhou (2012). A new model of information content based on concept's topology for measuring semantic similarity in WordNet. *International Journal of Grid and Distributed Computing*, 5(3), 81–94.

Michelizzi, J (2005). Semantic relatedness applied to all words sense disambiguation. Doctoral dissertation, University of Minnesota.

Miller, GA (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 1–28.

Nababteh, M and M Deri (2017). Experimental study of semantic similarity measures on Arabic WordNet. *IJCSNS International Journal of Computer Science and Network Security*, 17(2), 131–140.

Nagoudi, EMB, J Ferrero, D Schwab and H Cherroun (2018). Word embedding-based approaches for measuring semantic similarity of Arabic-English sentences. In *Arabic Language Processing: From Theory to Practice. ICALP 2017*. Communications in Computer and Information Science, BK Lachkar (ed.), Vol. 782, pp. 19–33. Cham: Springer.

Nagoudi, EMB and D Schwab (2017). Semantic similarity of Arabic sentences with word embeddings. In *Proc. Third Arabic Natural Language Processing Workshop (WANLP)*, Valencia, Spain, pp. 18–24.

Pagliardini, M, P Gupta and M Jaggi (2018). Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 528–540. New Orleans, Louisiana.

Pawar, A and V Mago (2018). Calculating the similarity between words and sentences using a lexical database and corpus statistics. Available at https://arxiv.org/abs/1802.05667.

Pilehvar, MT, D Jurgens and R Navigli (2013). Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 1341–1351.

Pronoza, E and E Yagunova (2015). Comparison of sentence similarity measures for Russian paraphrase identification. In *Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conf. (AINL-ISMW FRUCT)*, St. Petersburg, pp. 74–82.

Radev, DR, P Muthukrishnan and V Qazvinian (2009). The ACL anthology network corpus. In *Proc. 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries. Association for Computational Linguistics*, pp. 54–61. Suntec City, Singapore.

Radinsky, KA (2011). A word at a time: Computing word relatedness using temporal semantic analysis. In *The 20th Int. Conf. World Wide Web (WWW '11)*, pp. 337–346. NewYork, NY, USA: ACM.

Resnik, P (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proc. 14th Int. Joint Conf. Artificial Intelligence*, pp. 448–453. Montreal, Quebec, Canada.

Rubenstein, HG (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10), 627–633.

Sahami, M and T Heilman (2006). A web-based kernel function for measuring the similarity of short text snippets. In *Proc. 15th Int. World Wide Web Conf.*, pp. 377–386. Edinburgh, Scotland.

Sanborn, AA (2015). Deep learning for semantic similarity. In *CS224d: Deep Learning for Natural Language Processing*.Stanford, CA, USA: Stanford University.

Selamat, A and HH Ismail (2008). Finding English and translated Arabic documents similarities using GHSOM. In *Proc. Int. Conf. Computer and Communication Engineering 2008*, Kuala Lumpur, Malaysia, pp. 460–465.

Soori, H, M Prilepok, J Platos, E Berhan and V Snasel (2013). Text similarity based on data compression in Arabic. *Lecture Notes in Electrical Engineering*, 282, 211–220.

Sultan, Md A, S Bethard and T Sumner (2014). DLS@CU: Sentence similarity from word alignment. In *Proc. 8th Int. Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, pp. 241–246.

Taieb, MAH, M Ben Aouicha and A Ben Hamadou (2013). Computing semantic relatedness using Wikipedia features. *Knowledge-Based Systems*, 50, 260–278.

Taieb, MAH, MB Aouicha and Y Bourouis (2015). FM3S: Features-based measure of sentences semantic similarity. In *Int. Conf. Hybrid Artificial Intelligence Systems*, pp. 515–529. Bilbao, Spain.

Turney, P (2005). Measuring semantic similarity by latent relational analysis. In *Proc. Nineteenth Int. Joint Conf. Artificial Intelligence (IJCAI)*, pp. 1136–1141. Edinburgh, Scotland.

Wali, W, B Gargouri and A Ben Hamadou (2017). Enhancing the sentence similarity measure by semantic and syntactico-semantic knowledge. *Vietnam Journal of Computer Science*, 4, 51–60.

Wali, W, B Gargouri and A Ben Hamadou (2017). Sentence similarity computation based on WordNet and VerbNet. *Computación y Sistemas*, 21(4), 627–635.

Wan, S, M Dras, R Dale and C Paris (2006). Using dependency-based features to take the "para-farce" out of paraphrase. In *Proc. Australasian Language Technology Workshop*, pp. 131–138.

Wu, Z and M Palmer (1994). Verb semantics and lexical selection. In *Proc. Annual Meeting of the Association for Computational Linguistics*, pp. 133–138. Las Cruces, New Mexico, USA.

Zhou, ZW (2008). New model of semantic similarity measuring in wordnet, 2008. In *3rd Int. Conf. Intelligent System and Knowledge Engineering (ISKE)*, Vol. 1, IEEE, pp. 256–261.