Textual Similarity Measurement Approaches: A Survey

Amira H. Abo Elghit*[1], Aya M. Al-Zoghby**[2], and Taher Hamza*[3]

*[1,3]*Computer Science, Mansoura University, Mansoura, Egypt*
[1] amira-hamed@mans.edu.eg
[3]taher_hamza@yahoo.com
**[2]*Computer Science, Damietta University, Damietta, Egypt*
[2]aya_el_zoghby@du.edu.eg

Abstract: ***Survey research is appropriate and necessary to address certain research question types. This paper aims to provide a general overview of the textual similarity in the literature. Measuring textual similarity tends to have an increasingly important turn in related topics like text classification, recovery of specific information from data, clustering, topic retrieval, subject tracking, question answering, essay grading, summarization, and the nowadays trending Conversational Agents (CA), which is a program deals with humans through natural language conversation. Finding the similarity between terms is the essential portion of textual similarity, then used as a major phase for sentence-level, paragraph-level, and script-level similarities. In particular, we concern with textual similarity in Arabic. In the Arabic Language, applying Natural language Processing (NLP) tasks are very challenging indeed as it has many characteristics, which are considered as confrontations. However, many approaches for measuring textual similarity have been presented for Arabic text reviewed and compared in this paper.***

Keywords: ***Semantic Similarity, Arabic Language, WordNet, Lexical-Based Similarity, Textual Similarity, Hybrid-based Similarity, Word Embedding.***

1 INTRODUCTION

Every single second, millions of bytes are added all over the world. Therefore, the information stored on the web is enormous, indeed. As a result, searching tools as search engines are indexing billions of web pages, which is just a fraction of information that can be reachable on the Web. However, the searching process discloses a large volume of information changing in relevance and quality. Appraisal of information in terms of relevance and reliability is central since an inappropriate use of information can outcome in inappropriate decisions and grave penalties [1]. The ranking task is reordering the results retrieved from the search tool based on the relevancy between the search result and the original inquiry issued. It is a central task in many NLP topics like information retrieval, question answering, disambiguation, text summarization, plagiarism detection, paraphrase identification, and machine translation [2]. Finding the similarity between terms is the essential portion of textual similarity, then used as a major phase for sentence-level, paragraph-level, and script-level similarities. In the event of measuring the relevancy between documents, many papers tend to analyze the surface words occurrence between documents.

Text representation is a significant task used to overthrow the unregulated form of textual data into a more formal construction before any additional analysis of the text. The differences in the approaches that exist in the literature review for textual similarity depend on the text representation scheme used before text comparison. There are different text representation schemes suggested by researchers likes Term Frequency-Inverse Document Frequency (TF-IDF)[1], Latent Semantic Indexing (LSI)[2], and Graph-based Representation [3],[4],[5]. Due to these ways, the similarity measure to compare text units also differs because one similarity measure may not be convenient for all representation schemes. For example, the cosine similarity based on geometric distance is an appropriate textual similarity measure for text represented as a bag of terms. But, it is obscure whether cosine similarity will achieve passable results when text symbolizes as a graph-based representation [3]. While the majority existing textual similarity measures are developed and used for English texts, very rare measures have been developed especially, for the Arabic Language [6]. Thus, in this work, we discuss the effort done by researchers for the task of measuring similarity for many languages English, Spanish, Arabic, etc.

[1] https://monkeylearn.com/blog/what-is-tf-idf/

[2] https://blog.hubspot.com/marketing/what-is-latent-semantic-indexing-why-does-it-matter-for-your-seo-strategy

The following section is a background on the textual similarity concept and summarizes the most relevant associated work. Then the paper concludes.

2 TEXTUAL SIMILARITY CONCEPT AND LITERATURE REVIEW

Measuring textual similarity tends to have an increasingly important function in related topics like text classification, recovery of specific information from data, clustering, reveal the topic, subject tracking, question answering, essay grading, summarization, and the nowadays trending Conversational Agents (CA), which is a program that deals with humans through natural language conversation. Finding the similarity between terms is the essential portion of textual similarity, used as a major stage for sentence-level, paragraph-level, and script-level similarities [7]. The relevancy of words can be estimated in two manners: semantically and lexically. If two terms have a similar chain of characters, they are lexically analogous. Otherwise, if they have the identical context and significance, although they contain different characters, they are semantically analogous. Then a more recent approach (the hybrid similarity) has been used, which is an integration of different similarity measurements [8].

A. Text Similarity Approaches

According to [7],[8] We found that the text-similarity approaches are three categories and illustrated in Fig.1 and 2.

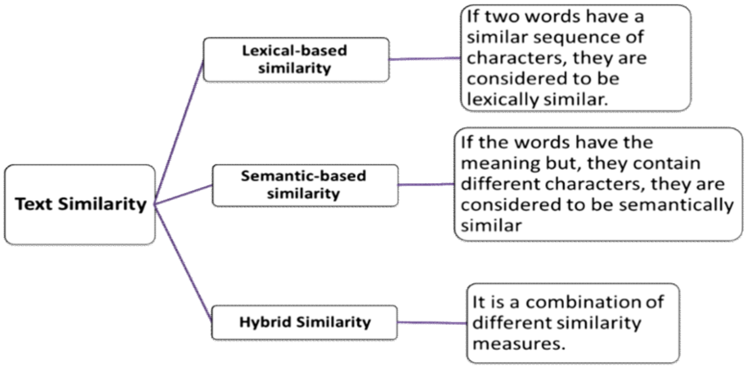
Text Similarity
Lexical-based similarity
If two words have a similar sequence of characters, they are considered to be lexically similar.
Semantic-based similarity
If the words have the meaning but, they contain different characters, they are considered to be semantically similar
Hybrid Similarity
It is a combination of different similarity measures.

Figure 1: Text Similarity Types.

1) *Lexical-Based Similarity (LBS):* These techniques depend on computing the distance among two chains to recognize the similarity among them. LBS measurements are categorized into two groups: character-based and term-based distance measurements. Character-based measurements were proposed to handle typographical errors. Even so, these measurements go wrong to captivate the similarity with term arrangement issues (like, “John Smith” versus “Smith, John”). Term-based similarity measurements try to recompense for this issue [7],[8]. In Table 1, we summarize the most prominent attempts to measure the lexical-based similarity and compare them according to the applied technique, the used dataset /sample in the experiment and the results released by each approach, chronologically arranged.

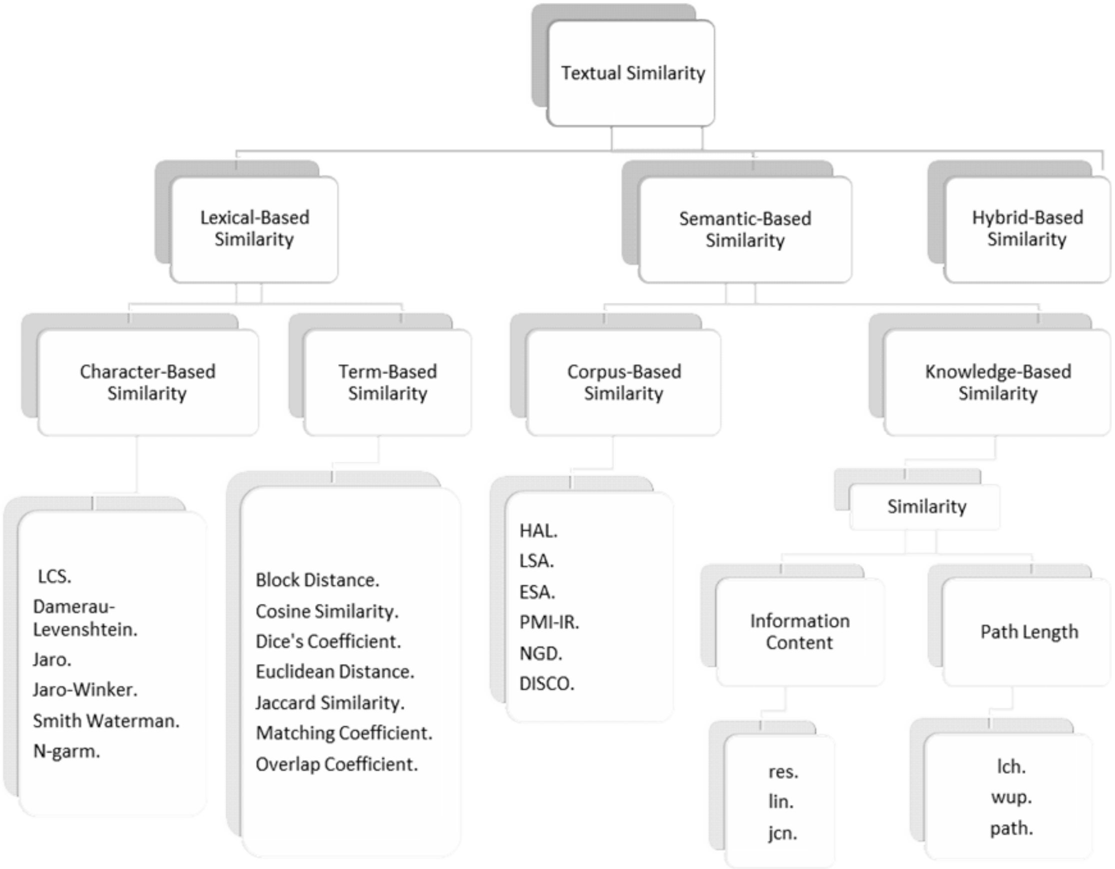

Figure 2: String Similarity Approaches according to [7], [8]

2) *Semantic-Based Similarity:* Semantic similarity defines the similarity among sequences that depend on their significance instead of using character-matching [8]. It is considered a potential part of Natural Language Processing (NLP) tasks such as word sense disambiguation, text summarization, entailment, machine translation, etc. [6]. In [9], the authors consider semantic similarity as a challenging task if you have two texts then the challenge is to measure how similar they are or to decide if they have a qualitative semantic relation between them. Generally, it is divided into two main ways to calculate the similarity among sequences: corpus-dependent and knowledge- dependent measures. A large corpus is used to define the similarity between words. However, Arabic is a weakly resourced since there is a lack of data, due to this research into corpus-based and computational linguistics in Arabic is affected. Otherwise, Knowledge-based similarity measurements can divide into two groups: measurements of semantic similarity/relevancy and measurements of semantic relatedness. Semantic similarity is a type of relatedness measurement between two terms in which a wide range of relations between connotations is covered [7].

In Table 2, we summarize the most prominent attempts to measure the semantic-based similarity and compare them according to the type of semantic similarity (Knowledge-based, Corpus-based), the applied technique, the used dataset /sample in the experiment and the results released by each approach, chronologically arranged.

3) *Hybrid-Based Similarity:* It is a combination of Lexical-Based Similarity measures and Semantic-Based Similarity measures. Most of the recent researches used this kind of Similarity measure. In Table 3, we summarize the most prominent attempts to measure the hybrid-based similarity and compare them according to the applied technique, the used dataset /sample in the experiment and the results released by each approach, chronologically arranged.

B. Literature Review

There is extensive literature that deals with the measurement of textual similarity addressed in the following. Mihalcea and et al. [10] developed a method that focus on measurement of the semantic similarity of short texts where Semantic similarity measurements: two were corpus-based (Point-wise Mutual Information, Latent Semantic Analysis) and six were knowledge-based measures (Leacock & Chodorow, Lesk, Wu, and Palmer, etc.) were examined. They noted that the maximum similarity was sought only within classes of words with the same part-of-speech.

Islam, A., & Inkpen, D. [11] provided a method that estimates the similarity of two texts from the integration of semantic and syntactic information. Since a corpus-based measure of semantic term similarity and adjusted version of the Longest Common Subsequence (LCS) string matching algorithm used. They concluded that the main advantage of this system was that it has a lower time complexity than the other system because it used only one corpus-based measure.

Abouenour et al. [12] proposed two directions for improvement: firstly, a semantic Query Expansion (QE) used in the purpose to have a senior level of recall when the Information Retrieval (IR) process retrieved passages; then a structure-based process used for passage re-ranking to have the expected answer at the top of the candidate passages list. In the first step, they used the content and the semantic relations within the Arabic WordNet (AWN) [3]ontology, and in the second step, they adopted the Passages Ranking (JIRS PR) module based on the Distance Density n-gram model. This system gives a higher similarity value to those passages containing more grouped structures. The highest performances were obtained when Java Information Retrieval System JIRS was used together with the semantic Query Expansion (QE).

Dai and Huang [13] presented a word semantic similarity based on the Chinese–English HowNet[4] ontology. The main aim of this work was to compute the similarity between terms by exploring their attributes and relations. For a given word pair, similarities between their attributes by combining distance, depth, and related information are computed. Then word similarity was estimated through a combination scheme.

Refaat et al. [14] presented a method to assess Arabic free text answer (essay) automatically based on Improved Latent Semantic Analysis (LSA) technique (the input text, unifying the form of letters, deleting the formatting, replacing synonyms, stemming and deleting "Stop Words") to produce a matrix that represents texts better than the traditional form of LSA matrix.

Navigli and Ponzetto [15] presented an automatic approach to the construction of BabelNet [5](a very large, wide-coverage multilingual semantic network). It is based on the integration of lexicographic and encyclopedic knowledge from WordNet and Wikipedia. To achieve the best translation performance, they relied on recent advances in machine translation by using Google from WordNet and Wikipedia.

Gomaa et al. [16] presented a different unsupervised approach to treat with students' answers using document similarity. It is divided into three stages: The first stage is measuring the similarity between model answer and student answer, using thirteen String-Based algorithms, 6 of them were Character-based, and the other 7 were Term-based measures. The Second stage was measuring the similarity using distributionally similar words using co-occurrences (DICSO[6]1 and DISCO2) Corpus-based similarity measurements. Finally, they are combined to accomplish a maximum correlation value.

Nitish, A. et al. [17] implemented two approaches to estimate how much the two sentences are similar. The first approach combined corpus-based semantic relatedness measure over the entire sentence with the knowledge-based semantic similarity degrees got for the words that have the same syntactic roles in both the sentences. Then, it fed all these scores as features to machine learning models to estimate the similarity score. The second approach used a bipartite based method over the WordNet and Lin measure, without any modification.

[3] https://www.researchgate.net/publication/305406094_Arabic_WordNet_New_Content_and_New_Applications

[4] http://godel.iis.sinica.edu.tw/taxonomy/taxonomy-edoc.htm

[5] https://babelnet.org/

[6] http://www.linguatools.de/disco/disco_en.html

Daniel Bar et al. [18] First, they computed text similarity scores between pairs of texts and their sources using Content similarity (longest common substring measure), Structural Similarity (N-gram model), and Stylistic Similarity (Sequential TTR*)*. Then, using these scores as features for two machine learning classifiers from the WEKA toolkit[7], such as a Naïve Bayes classifier and decision tree classifier.

Zou et al. [19] Introduced bilingual word embeddings[8]: semantic embeddings associated across 2 languages in the context of neural language models. A single semantic similarity feature induced with bilingual embeddings added near half a BLEU point to the results of the NIST08 Chinese-English machine translation task.

Kaleem et al. [20] presented a sentence similarity approach formed to mitigate the issue of free word order in the Urdu language. The main objective behind it was to alleviate the complex word order issue that comes with the Urdu language by matching all possible word order variations on a single scripted pattern to reduce the time and effort required to script an Urdu conversational agent.

F. Elghannam [21] proposed a new corpus-based method to measure the semantic similarity between short texts to ranking them. It uses the statistical lexical similarity between the vectors of similar words (second-order word vectors) extracted from the corpus instead of relying on only word distribution similarity calculations. To determine the degree of similarity, she measures the lexical similarity between their second-order word vectors.

J. Tian et al. [22] A universal model in the combination of traditional NLP methods and deep learning methods together to define semantic similarity proposed. First, they translate all sentences into English through the machine translation (MT) system, i.e., Google Translator.

S. Xiaolong et al. [23] proposed a new framework for computing semantic similarity. The model learned word segmentation automatically and the overall architecture LSTM network to extract semantic features and then used the attention model to enhance semantics since, LSTM model was used to extract the semantic features of the sentence based on the Siamese network, attention model used to weight the semantic output of each moment and the Policy network used to determine whether the sentences need to segment or not. The experiments showed that the model improved the accuracy by 95.7% compared to the previous baseline models.

Kim et al. [24] proposed an attention mechanism to capture the semantic correlation and to appropriately align the two-sentence elements using a densely connected co-attentive recurrent neural network (DRCN). They connected the recurrent and co-attentive features from the bottom up with no distortion. The result showed that the dense connections over-attentive features were more effective. DRCN showed the highest mean, which indicates that it did not only results in a competitive performance but also has a consistent performance.

Khafajeh et al. [25] proposed an automatic Arabic thesaurus. They used term frequency-inverse document frequency (TF-IDF) for index term weights, Similarity Thesaurus by using Vector Space Model with four similarity measurements (Cosine, Dice, Inner product, and Jaccard) and Association thesaurus by Applying Fuzzy model.

[7] https://www.cs.waikato.ac.nz/ml/weka/

[8] https://machinelearningmastery.com/what-are-word-embeddings/

TABLE I: Chronological Representation of the Most Important Related Researches Concern on Lexical-Based Similarity.

System	Year	Technique	Sample/Dataset	Aim	Results
S. H. Mustafa et al. [6]	2004	N-gram conflation scheme uses to transform a word into a chain of N-grams.	Arabic words (a sample of text consisting of 6,000 distinct textual words).	They examined the performance of the bigram and trigram techniques in the context of Arabic free text retrieval.	The results showed that the digram method gives overall highly effective values than the trigram method regarding the two measures used: conflation recall (CR) (Digram 0.62, Trigram 0.44) ratio and conflation precision (CP) ratio (Digram 0.66, Trigram 0.47).
Khafajeh et al. [25]	2010	Term frequency-inverse document frequency (TF-IDF) for index term weights, Similarity Thesaurus by using Vector Space Model with four similarity measurements (Cosine, Dice, Inner product, and Jaccard) and association thesaurus by applying Fuzzy model.	Arabic (59 queries ran against the 242 documents that were provided in the Saudi Arabian National Computer Conference).	An automatic Arabic thesaurus using term-term similarity used.	The experiment showed that the Jaccard and Dice similarity measurements are the same for the VSM model, while the Cosine and Inner similarity measurements are the same too, but they are a little bit better than Jaccard and Dice measures, and it gives nearly the same ranking for all the queries. Using stemmed words with similarity and association thesaurus in the Arabic language retrieving system is much better than using full words without using the thesaurus.
K. Shaalan et al. [26]	2012	Leven-shtein distance similarity measure, N-gram (tri-gram language model).	An Arabic term list that includes 9,000,000 fully inflected surface words.	They proposed a Spelling Checking tool for Arabic words, which involved in a trigram language model to approximate knowledge about permissible characters in Arabic.	They concluded that their system performed better than Hunspell in choosing the best solution, but it was still below the MS Spell Checker. Testing of this language model given the precision of 98.2% at a recall of 100%.
Al-Ramahi et al. [27]	2012	N-gram language modeling (bigram model), the weighting vector model, and term similarity measures (Cosine and Dice) were used.	Arabic course description among Jordanian universities.	They implemented a system that computes the similarity among course descriptions of the same subject from various universities or similar academic programs.	The results indicated that the word-based N-gram technique using Cosine Similarity provided better accuracy rates (80%) than the word-based technique and the whole document-based N-gram technique that use Dice's Coefficient.
A.Magooda et al. [2]	2016	TF-IDF module, language modeling module, and Wikipedia module.	Four datasets were used in this system. The training performed on data consists of 1030 search queries. Each query is accompanied by 30 search engine retrieved results.	They proposed a ranking system of three components integrated to produce a relevancy score.	The proposed system achieved the 3rd position in the Arabic subtask of SemEval 2016 task3 with 43.80 Mean Average Precision on test set compared to the baseline system (MAP = 28.88).

As shown in Table 1, some researches were concerned with string-based similarity. S. H. Mustafa and et al. [6] examined the performance of the bigram and trigram techniques in the context of Arabic free text retrieval since the N-grams [9]conflation scheme uses to transform a word into a chain of N-grams. The experiments were done on a corpus of thousands of distinct textual words drawn from several sources representing various disciplines.

K. Shaalan et al. [26] proposed a Spelling Checking tool for Arabic, which depend on a trigram language model to approximate knowledge about permissible characters to classify generated words as valid and invalid, a finite-state automaton that measures the edit distance between input words and candidate corrections, the Noisy Channel Model, and knowledge-based rules.

Al-Ramahi et al. [27] implemented a system that computes the similarity among course descriptions of the same subject from various universities or similar academic programs. Since 3 different bi-gram techniques used: the vector model to represent each document in a way that each bi-gram is associated with a weight that reflects the importance of the bi-gram in the document. Then, the cosine similarity is used to compute the similarity between the 2 vectors. The other two techniques were: word-based and whole document-based evaluation techniques. In both techniques, the Dice's similarity measure applied for calculating the similarity between any given pair of documents.

A. Magooda et al. [2] proposed a ranking system of three components: TF-IDF based module, Language model (LM) based module, and Wikipedia-based module. Then, the three relevancy values calculated are then converted into one relevancy score using weighted summation. Retrieved documents are then re-ordered based on the new weighted sum scores.

[9] https://www.tidytextmining.com/ngrams.html

TABLE II: CHRONOLOGICAL REPRESENTATION OF THE MOST IMPORTANT RELATED RESEARCHES CONCERN ON SEMANTIC-BASED SIMILARITY.

System	**Year**	**Type of Similarity**	**Technique**	**Sample/Dataset**	**Aim**	**Results**
Mihalcea et al. [10]	2006	Corpus- based	Semantic similarity measurements: six were corpus-based (Point-wise Mutual Information, Latent Semantic Analysis), and two were knowledge-based measures (Leacock & Chodorow, Lesk, Wu, and Palmer, etc.) were examined.	British National Corpus – a 100million word corpus of English.	They developed a method that focuses on measuring the semantic similarity of short texts. The maximum similarity was sought only within classes of words with the same part-of-speech.	The result showed that the semantic similarity approach performed better than other methods based on simple lexical matching, resulting in a reduction of error rate by up to 13% regarding the traditional vector-based similarity metric.
Abouenour et al. [12]	2010	Knowledge- based	Arabic WordNet (AWN) ontology, Distance Density N-gram model.	Two datasets were used: a set of 1500 from the Text Retrieval Conference (TREC) and 764 from Cross-Language Evaluation Forum (CLEF) questions manually translated to the Arabic language.	They presented an approach for improving the re-ranking of passages (PR) module for Arabic Question/Answering (Q/A) systems.	For the TREC questions, the accuracy passes from 8.16% to 18.99%, the MRR from 3.1 to 8.61 and the percentage of the answered questions from 16.78% to 24.48%. For the CLEF questions the accuracy is close to 22% instead of 12%, the MRR reaches 10.08 rather than 3.85.
Dai and Huang [13]	2011	Knowledge- based	HowNet ontology is used to explore attributes and semantic relations that connect sememes and concepts between terms.	An English-Chinese bilingual data set (denoted as EC62) was constructed.	They proposed a word semantic similarity based on which the Chinese–English HowNet ontology. The main aim of this work was to compute the similarity between terms by exploring their attributes and relations.	Experiments showed that the performance of the measure was close (i,e. correlation coefficient value of 0.9035) to human judgments and concluded that attribute and semantic relation contributed a great deal to word similarity measuring and HowNet is applicable in the task for English-Chinese cross-lingual scenarios.
Navigli and Ponzetto [15]	2012	Knowledge- based	BableNet, WordNet, encyclopedic, Wikipedia as knowledge bases, were used.	SemEval[10] 2010 CL-WSD dataset, which consists of 1,000 test instances used, and this approach supported all languages. (Machine Translation is applied to enrich the resource with lexical information).	A multilingual semantic similarity approach used Babel-Net; a knowledge-rich lexicon and semantic database proposed.	The experiments showed that the approach produces a large-scale lexical resource with high accuracy, and the better results achieved by the graph-based algorithm permits them to establish that exploiting the structure of the target resource boosts the performance on the mapping task.

[10] https://semeval.github.io/

Zou et al. [19]	2013	Corpus- based	MT Alignment weighting model to initialize Chinese word embeddings, Vector matching alignment model, Named Entity Recognition model, machine translation system.	Training was performed on the Chinese Gigaword corpus. The testing was used dataset contains 297 Chinese word pairs with similarity scores estimated by humans.	A scheme that captivates both mono and cross-lingual semantic relations among different languages proposed.	The new embeddings significantly out-performed baselines in word semantic similarity.
El Moatez Billah Nagoudi et al. [28]	2017	Corpus- based	Word Embedding model with IDF weighting and Part-of-Speech tagging weighting methods.	More than one Arabic corpus was used.	The major idea is to take advantage of vectors as term representations to captivate the semantic and syntactic features of words then ranking them based on the relevancy.	These results indicated that when no weighting method was used the correlation rate reached 72.33%. Both IDF-weighting and POS tagging methods performed better with the correlation to more than 78% (respectively 78.2% and 79.69%).
Yang S. Hitachi [29]	2017	Corpus- based	Word embedding (GloVE model) with a convolutional neural network (CNN).	SemEval 2017 STS Benchmark was used to support Multi-languages (English, Spanish, and Arabic).	They provided an approach that trained a CNN to transfer GloVe word vectors to calculate the semantic similarity score between two sentences.	They ranked third on primary track of SemEval STS 2017 task1where the difference in performance between this model and the best-performing systems that participated in this task were less than 0.1
M. Khaled and M.P.Ben Hamza. [30]	2018	Knowledge- based	Arabic WordNet (AWN), knowledge-based similarity measures.	The Arabic corpus consists of a sample of Arabic textual document that is taken from the Arabic test collection named as EveTAR.	They aimed to introduce a method of improvement of information retrieval in Arabic documents.	The results showed that efficient results were obtained by this method through comparison with other methods in terms of precision and recall. The mean average was about their algorithms, keyword searching, and query-expansion was 0.826364, and 0.576666457, respectively.

As shown in Table 2, Semantic-based similarity is divided into 2 types: corpus-dependent and knowledge-dependent measures. In Corpus-based similarity, a large corpus is used to define the similarity between words. Otherwise, Knowledge-based similarity used lexical resources such as WordNet, VerbNet, etc.

TABLE III: CHRONOLOGICAL REPRESENTATION OF THE MOST IMPORTANT RELATED RESEARCHES CONCERN ON HYBRID-BASED SIMILARITY.

System	Year	Technique	Sample/Dataset	Aim	Results
Islam, A., & Inkpen, D. [11]	2008	Normalized longest common subsequence measure and pointwise Mutual Information (PMI-IR)[11] method.	Two different English datasets were used: Li et al. dataset and Microsoft paraphrase corpus.	They provided a method that estimates the similarity of two texts from the integration of semantic and syntactic information.	The results evaluated by 2 diverse data sets showed that this STS method performs better than several competing methods regarding the accuracy of 72.42% when they used 0.6 as the similarity threshold score, and they recommend this threshold achieve an accuracy of 72.64%.
Refaat et al. [14]	2012	Latin Semantic Analysis technique (LSA) and Cosine similarity measure.	29 Arabic answering papers are collected from students' answers in the System Designing course in second grade in the Compute Teacher Preparation Department.	An automatic Arabic essay scoring system using LSA similarity measurement developed.	The correlation between the human assessor and the system is 0.91. The results indicated that a correlation varied from 0.78 to 0.87. Also, the computation time per answer has decreased with a percentage of 4%.
Gomaa et al. [16]	2012	String-based such as (Block Distance, N-gram modeling) and Corpus-based such as Latent Semantic Analysis (LSA), Point-wise Mutual Information - Information Retrieval (PMI-IR) and Distributional Similar words using Co-occurrences (DISCO) similarity measures were used.	Texas short English answer grading data set was used. It consists of ten assignments between four and seven questions each and two exams with ten questions each.	Using lexical-based and corpus-based similarity measurements to implement their short answer grading system.	The achieved correlation value of 0.504 was the best value achieved for the unsupervised approach Bag of Words (BOW) when compared to previous work.
Nitish, A., et al. [17]	2012	Explicit Semantic Analysis (ESA) as the corpus-based semantic Measure, Lin measure, modified WordNet, Machine learning models such as Linear Regression (LR), and Bagging models were used.	The training process used the MSRvid dataset and the testing performed on MSRpar and SMTeuro datasets.	An approach that integrates corpus-based semantic relatedness measurement over the whole sequence along with the knowledge-based semantic similarity scores presented.	The results showed significant improvement against ESA. Although, it can be apparent that the baseline results were even better than of the ESA in the cases of MSRpar and SMTeuro datasets.
Daniel Bar et al. [18]	2012	Content similarity (longest common substring measure), Structural Similarity (N-gram model), and Stylistic Similarity (Sequential TTR), machine learning classifiers (Naïve Bayes classifier and decision tree classifier).	Three standard evaluation datasets used: The Wikipedia Rewrite corpus, the METER corpus, and the Webis Crowd Paraphrase corpus.	A simple log-linear regression model depends on training data to integrate multiple textual similarity measurements was used.	They concluded that for novel datasets that were essential to address the dimensions explicitly in the annotation process, so that text reuse detection approaches can be evaluated precisely against the characteristics of different kinds of data.

[11] https://eranraviv.com/understanding-pointwise-mutual-information-in-statistics/

Kaleem et al. [20]	2014	Levenshtein Edit Distance Algorithm, Bipartite Matching model, Kuhn-Munkres algorithm used to find the maximum sum of a given matrix of weights.	Urdu language words.	Uses a hybrid approach that integrated a lexical sequence similarity measurement with pattern-matching methods.	The experiments show that this approach can be applying to any language with free word order such as Arabic, Hindi, and Bangladeshi can utilize it.
John Henderson et al. [31]	2015	Bag-of n-grams, Alignment method, Word Embedding, and Recurrent neural network architectures.	18,762 pairs of English tweets with a 70/25/5 split for train, development, and test sets.	Seven models of semantic similarity combined for paraphrase detection on Twitter.	This system was placed first in Semantic Similarity and second in Paraphrase Identification with scores of Pearson's r of 61.9%, F1 of 66.7%, and maxF1 of 72.4%.
F. Elghannam [21]	2016	Cosine similarity measure, Latent Semantic Analysis (LSA), and different classification algorithms (Naïve Bayes and Decision Table).	Two types of data are used a set of pairs of complete Arabic sentences and another set of short expressions.	The new corpus-based method proposed to measure the semantic similarity between short texts in to rank them.	The experiments show that the accuracy result in 97% in sentence test was obtained by the proposed method compared to 93% in the lexical similarity method.
G. Da San Martino et al. [32]	2016	Bag-of-words models e.g, n-grams, skip-grams, syntactic tree kernels, were used.	English Question /Answering pairs.	They studied the effect of various kinds of features for question ranking using initial rank provided by the search engine, which represents a strong baseline Google rank (GR).	This model outperformed Google Rank (GR) by 1.72 concerning mean average precision (MAP) 95%.
Hao Wu et al. [33]	2017	Word Embedding model, support vector machine (SVM) for regression, sequential minimal optimization (SMO).	SemEval 2017 STS Benchmark was used.	Three systems that are based on significance of information space (SIS) built on the significance hierarchical taxonomy in WordNet proposed.	Their team ranked second among thirty-one participating teams by the primary score of the Pearson correlation coefficient (PCC) mean of seven tracks and accomplish the best performance on the track one (AR-AR) dataset.
J. Tian et al. [22]	2017	Machine translation (MT) system, tree kernel model, N-gram, Word embeddings, and neural network architecture.	SemEval 2017 STS Benchmark was used.	A universal model in a combination of traditional NLP methods and deep learning methods together to define semantic similarity proposed.	The results showed that this combination not only enhanced the performance but also increases the robustness for modeling the similarity of multilingual sentences in terms of the mean Pearson correlation 73.16% in primary track at SemEval-2017 Task 1.
Basma Hassan et al. [34]	2017	BabelNet knowledge base, Word sense aligner, and Synset similarity measure.	SemEval 2017 STS Benchmark was used to support Multi-languages (English, Turkish, Spanish, and Arabic).	A sense-based language-independent textual similarity method is presented, in which a proposed alignment similarity method is combined with new usage of a semantic network (BabelNet).	The first run ranked 10th and the second-ranked 12th in the primary track in task 1, with correlation 0.619, and 0.617, respectively.

Bilal Ghanem et al. [35]	2018	Arabic WordNet, Corpus-based similarity measures, and Knowledge-based measures.	The ExAraPlagDet-2015 dataset that supports the Arabic language was used.	Corpus-based and knowledge-based approaches by utilizing an Arabic semantic resource (Arabic WordNet) were gathered.	This system develops a higher performance (F-score 89% vs. 84% accomplish by the other systems.
Nagoudi et al. [36]	2018	Word Embedding model, POS tagging method, Weighting Aligner for words, and Alignment Bag-of-Words with three weighting functions.	Four datasets drawn from the STS shared task SemEval-2017 with a total of 2412 pairs of sentences.	A combination of word embedding models, word alignment methods were used to estimate the semantic similarity between texts.	The mixed weighted method with Alignment Bag-of-Words provided a correlation rate of 77.39%, and the Weighting Aligned Words obtained a correlation rate of 73.75%
Kim et al. [24]	2019	GloVe embedding and variable GloVe embedding, densely connected co-attentive recurrent neural network (DRCN), and autoencoder for dimension reduction.	five popular and well-studied benchmark datasets used: SNLI and Multi NLI for natural language inference; Quora Question Pair for paraphrase identification; and TrecQA and SelQA for answer sentence selection in question answering.	They proposed an attention mechanism to obtain the semantic correlation and to appropriately align the two-sentence elements of using a densely connected co-attentive recurrent neural network (DRCN).	The result showed that the dense connections over-attentive features were more effective. DRCN showed the highest mean, which indicates that it did not only results in a competitive performance but also has a consistent performance.
S. Xiaolong et al. [23]	2020	Word embedding (word2vec), LSTM model, deep reinforcement learning for Siamese attention structure model (DRSASM).	The model used more than one dataset, such as the SNLI dataset, which was labeled with more than 500,000 sentence pairs and a Chinese car description dataset.	They proposed a new framework for computing semantic similarity. The model learned word segmentation automatically and the whole architecture LSTM network to extract semantic features and then used the attention model to improve semantics.	The experiments showed that the model improved the accuracy by 95.7% compared to the previous baseline models.

As shown in Table 3, the researches mentioned above combine String-based similarity and Semantic-based similarity to achieve the best results. For example, F. Elghannam [23] achieved accuracy result in 97% in sentence tests.

Ref. [37] Categorized the existing approaches that are concerned with measuring textual similarity between texts to three types depending on the text level to document similarity, sentence similarity, and word similarity. In the past, most of the researchers focused on the documents level similarity (two long texts or long text with small ones). Recently, the sentence level similarity has more interest, which led to provide training, test data in multi-languages, and deploy different approaches for sentence similarity. Generally, these approaches are divided into three categories, namely: vector space-based approaches in which the text is represented as a vector of features, using bag-of-words (BoW) then, compute the similarity between their vectors. Alignment-based approaches that assume that linguistics expressions that have similar meaning could be aligned. Moreover, machine learning-based approaches are based on supervised machine learning along with semantic similarity measures and features (Lexical, syntactic and semantic features) [38].

According to [37], the existing approaches that measure semantic similarity for Arabic texts either documents, sentences, or words divided into four types of techniques namely: word co-occurrence approaches that ignore word order of the sentence but, it successfully extracts keywords from documents, Statistical corpus-based approaches that use the Latent Semantic Analysis (LSA) as a language-understanding model, Descriptive features-based approaches, which depend on the semantic features that extract from dictionaries, or WordNet as a lexical resource. Finally, neural networks-based approaches with word embeddings. Regarding the previous taxonomy mentioned above, we review some of those approaches as following.

Nagoudi et al. [36] combined Word Embedding (CBOW model), word alignment method, IDF, and POS weighting features for extracting semantic and syntactic features from documents to capture the most relevant ones but, it was weak in representing data with higher dimensionality. It is confined to working on distant local contextual windows rather than counting global co-occurrences.

M. Al-Samdi et al. [38] proposed an approach for paraphrase identification (PI) and semantic text similarity (STS) analysis in Arabic news tweets. It employs a set of extracted features divided into Text overlap features (such as n-grams, stemmed n-grams and POS overlap features), Word Alignment features, and Semantic features (such as NER overlap features and topic modeling features) to detect the level of similarity between tweets pairs. They noted that the lexical overlap features play a notable role in improving the results of PI and STS analysis. Additionally, semantic features enhance the results of both tasks PI and STS. Word alignment features significantly enhance the results of PI, whereas results obtained by overlapping features based on NER and POS are acting as bad features when used alone with the lexical overlap features. The best-realized results in both tasks are when using the lexical overlap features with the word alignment and topic modeling features.

M. Zrigui and A. Mahmoud [39] presented a semantic approach that identifies whether an unseen document pair is paraphrased or not. It consists of two phases. At the feature extraction phase, they used global vectors representation combining global co-occurrence counting and a contextual skip-gram model. At the Paraphrase identification phase, a convolutional neural network is used to learn more contextual and semantic information between documents.

Konopik et al. [40] Introduced a system for estimation of semantic textual similarity in SemEval 2016. The core of this system consisted of exploiting distributional semantics to compare the similarity of sentence chunks. They used a broad range of machine learning algorithms in addition to several types of features (Lexical features include word base form overlap, word lemma overlap, chunk length difference, Syntactic features include POS tagging, Semantic features include GloVe, Word2Vec, and WorNet database.).

Almarwani et al. [41] addressed the problem of textual entailment in the Arabic Language. Their approach consisted of a combination between traditional features such as length of sentences and similarity scores (Jaccard and Dice), named entity recognition and Word Embedding (Word2Vec).

S. A. Al Awaida et al. [42] proposed an Automated Arabic essay grading model to achieve better accuracy by combining the F- score technique to extract features from student answers and model answers with Arabic WordNet (AWN) to find all relevant words from student answer for semantic similarity.

A. El-Hadi et al. [43] presented a new approach for semantic similarity measures based on the MapReduce framework and WordNet after the translation phase to compute the similarity between Arabic queries and documents. The experiments were running on a variable number of documents in the corpus stored in HDFS in an Arabic search engine.

M. Zrigui and A. Mahmoud [44] presented a method based on deep learning for paraphrase detection between documents. Since the word2vec model extracted the related features by anticipating each word with its neighbors. Then, the obtained vectors are averaged to generate a sentence vector representation (Sen2vec). Finally, a Convolutional neural network (CNN) is used to capture more contextual information and semantic similarity computation.

A. Omar and W. Hamoda [45] studied the effect of document length on measuring the semantic similarity in the text clustering of Arabic news by many experiments with different normalization techniques such as Byte length normalization, cosine normalization, Maximum normalization, Mean normalization, etc. to choose a reliable one for the previous purpose. The study proposed the integration of TF-IDF for ranking the words within all the documents. It deduced that the Byte length normalization method is the most appropriate for text clustering with TF-IDF values.

Based on what Wali et al. [46] discussed, we noted that most of the previous researches mentioned above estimated the semantic similarity based only on the word order or the syntactic dependency and the synonymy relationship between terms in sentences without taking into consideration the semantic arguments namely the semantic class and thematic role in computing the semantic similarity. Wali et al. [46] presented a hybrid method for measuring semantic similarity between sentences depending on supervised learning and three linguistics features (Lexical, Semantic and Syntactic-Semantic) extracted from learning corpus and Arabic dictionaries like LMF dictionary. This is a two-phase method: the learning phase, which consists of two processes: the pre-processing process that aimed to have an annotated corpus and the training process that is used to catch a hyperplane equation via the learning algorithm. The second phase was the testing phase that implemented the learning results from the first one to compute the similarity score and classify the sentences as similar or not similar.

Wali et al. [47] proposed the original idea because it has not been employed yet in former research in the literature. They presented a Hybrid Similarity measure that aggregated in linear function, three components (Lexical similarity using Lexsim, semantic similarity using Semsim that uses the synonymy words extracted from WordNet and syntactic-semantic similarity SynSenSim based on common semantic arguments such as thematic role and semantic class.) Moreover, the determination of the semantic arguments has been based on the VerbNet database.

Wali et al. [48] proposed a multilingual semantic approach based on similarity to calculate the similarity degree between the user's answer and the right one saved in the dataset. It supports three languages: English, French, and Arabic. Hybrid Similarity measure (Lexical, semantic, and syntactic-semantic) using knowledge bases like WordNet, LMF dictionary, etc. used. They concluded that the short sentences achieved the best measures of recall and precision. As the sentence gets longer, there will be more calculation, which reduces the system's performance.

We summarize them in Table 4 according to the applied technique (word co-occurrence, feature-based approach, Latent semantic analysis or Hybrid approach), the used dataset /sample in the experiment, the aim of each one, the similarity type (String-based, Corpus-based, Knowledge-based, Hybrid-based) and the results obtained by each approach.

TABLE IV: CHRONOLOGICAL REPRESENTATION OF THE MOST IMPORTANT RELATED RESEARCHES IN ARABIC.

System	**Year**	**Technique**	**Text Level**	**Sample/Dataset**	**Description**	**Results**	**Similarity Type**
Almarsoomi et al [49]	2013	Semantic similarity measure using Arabic WordNet knowledge base.	word	A dataset of 70 pairs of words was used.	A method to measure the semantic similarity between two Arabic words in the Arabic WordNet knowledge base was presented.	The Pearson correlation of this approach was 0.894 compared to the human average of 0.893 for the same data.	Semantic similarity (Knowledge-based)
Wali et al. [46]	2015	Hybrid Similarity measure (Lexical features like common words, semantic and syntactic-semantic features extracted from knowledge bases like LMF dictionary, etc.) and multiple classifiers.	Sentence	Set of 1380 sentences like the dictionary definitions and examples of definitions of words taken from Arabic dictionaries such as Lissan Al-Arab and Al-Wassit.	They presented a hybrid method for measuring semantic similarity between sentences based on supervised learning and three linguistics features extracted from an annotated corpus and Arabic dictionaries like the LMF dictionary.	The results showed a good performance that approximates human decisions. They noted that short sentences (<=10 words) presented the highest measures of recall and precision, but the longer ones reduced the system's performance.	Hybrid-based
Konopik et al. [40]	2016	Hybrid similarity measure: Lexical features include word base form overlap, word lemma overlap, chunk length difference. Syntactic features include POS tagging. Semantic features include GloVe, Word2Vec, and WordNet database.	Sentence	Global annotated evaluation dataset (Images, Headlines, Answer students). They have chosen not to fix the system for individual datasets but to fix it for the task as a whole.	Introduced a system for estimation semantic textual similarity in SemEval 2016.	The combination of relation types increased the score of similarity to 0.6484 in terms of F1-measure.	Hybrid-based
M. Al-Samdi et al. [38]	2017	Latent Semantic Indexing and feature-based: N-grams, POS overlap features Word Alignment and NER features.	Sentence	Dataset consists of Arabic tweets in the general domain of news collected from well-known tweeter accounts, namely Al-Arabiya and Al-Jazeera.	They proposed an approach for Paraphrase Identification (PI) and Semantic Text Similarity (STS) analysis of news Arabic tweets. It uses a set of extracted features based on lexical, syntactic, and semantic features.	The best-achieved results in both tasks are when using the lexical overlap features with the word alignment and semantic features. The overall improvement over the baseline method MaxEnt is 6% in PI (F1 = 0.872) and around 9% in STS (P = 0.912).	Hybrid-based

Almarwani et al. [41]	2017	Hybrid Technique consisted of: Traditional features included length of sentences, similarity scores (Jaccard and Dice), NER, and Word embedding (Word2Vec). Then supervised Classifiers used for prediction (SVM, LR, Random Forest).	Sentence	Different datasets were used, as Arabic Gigaword, Arabic Treebank, Arabic Wikipedia, and annotated data (ArbTE), including 600 modern standard Arabic pairs.	They addressed the problem of textual entailment in the Arabic Language.	Their approach yielded a peak performance on the ArbTE standard dataset, reaching 76.2% accuracy.	Hybrid-based
Wali et al. [47]	2017	Hybrid Similarity measure (Lexical, semantic, and syntactic-semantic) using knowledge bases (WordNet and VerbNet).	Sentence	Li et al. [6] dataset and Microsoft Paraphrase Corpus.	They summed up several methods to calculate the similarity between the sentence.	Their approach yielded competitive results compared to other methods tested on Li's benchmark. The precision reached a peak with Θ =0.9 showed a 0.742 as a value for the training and test dataset.	Hybrid-based
Nagoudi et al. [36]	2018	Word Embedding (CBOW model), word alignment method, IDF, and POS weighting were combined.	Document	External Arabic Plagiarism Corpus used.	They combined the mentioned features for extracting semantic and syntactic features from documents to detect the plagiarism between documents.	This method led to 0.8593 and 0.8781 of accuracy and recall, respectively.	Hybrid-based
M.Zrigui and A.Mahmoud [39]	2019	Word Embedding (Word2Vec and Sen2Vec models) with Convolutional neural network (CNN) to estimate the degree of similarity.	Document	An Arabic paraphrased corpus was developed based on the skip-gram model, and the experiments were carried out on the Open Source Arabic Corpus (OSAC) dataset.	They presented a method based on deep learning for paraphrase detection between documents.	The proposed system achieved good results in terms of precision 85% and recall 86.8% than previous studies.	Hybrid-based
Wali et al. [48]	2019	Hybrid Similarity measure (Lexical, semantic, and syntactic-semantic) using knowledge bases like WordNet, LMF dictionary, etc.	Sentence	Multilingual ontology-based question-answering training for patients with Alzheimer's disease.	They proposed a multilingual semantic approach based on similarity to estimate the similarity score between the user's answer and the right one saved in the dataset.	The performance of their approach confirmed through experiments on 20 patients, which promised capability for incorporating in Autobiographical Training and other applications, such as automatic summarization and data clustering.	Hybrid-based

A. El-Hadi et al. [43]	2019	Word co-occurrence, WordNet as a Knowledgebase and, Cosine similarity measure.	Document	The experiments were running on a variable number of documents in the corpus stored in HDFS in an Arabic search engine.	They presented a new approach for semantic similarity measures based on the MapReduce framework and WordNet after the translation phase.	The results showed that this approach gave better results than others in terms of precision= 60%, recall=50%, and F1=54.5% when it depends on Leacock and Chodorow approach and Cosine similarity measure.	Hybrid-based
S.A Al Awaida et al. [42]	2019	F-score is a selection technique, Arabic WordNet, and Cosine similarity measure.	Document	Arabic Essay dataset that has been created from a computer, science, and social school lectures from Allu'lu'a modern school in Madaba-Jorden.	They proposed an Automated Arabic essay grading model to achieve better accuracy.	The results showed that the model coupled with using Arabic WordNet (AWN) according to mean absolute error (MAE) value was 0.117, and the Pearson correlation was between 0.5 to 1. So, it produced a better result compared to the case without AWN.	Hybrid-based
S.Abdulateef et al. [50]	2020	Word embedding, Bag of words with weighted principal component analysis (WPCA) function, and K-means algorithm for clustering.	Sentence	Essex Arabic Summaries Corpus	They aimed to reduce the redundancy issue by extracting key sentences.	Recall-Oriented Understudy for Gisting Evaluation (ROUGE) used as an evaluation measure based on, their dataset and the method has achieved an F-score of 0.644.	Hybrid-based
M.Zrigui and A.Mahmoud [45]	2020	Word embedding and Convolutional Neural Network.	Document	A collection of documents randomly used from Open-Source Arabic Corpora (OSAP) source corpus.	They presented a semantic approach that identifies whether an unseen document pair paraphrased or not.	The model achieved promising results in terms of precision 82% and recall 80%.	Semantic Similarity (Corpus-based)
A.Omar and W. Hamoda [45]	2020	Word co-occurrence by term-frequency inverse document frequency (TF-IDF), Vector space clustering models (K-means algorithm), and Length normalization methods.	Document	A corpus of 693 stories representing the different categories and different lengths collected from Al-Ahram (Egypt), Al-Sharq Al-Awsat (KSA), Al-Bayan (UAE) and Al-Ghad (Jordan).	They studied the effect of document length on measuring the semantic similarity in the text clustering of Arabic news with different normalization techniques.	This work concluded that: the use of a single normalization method was not sufficient in addressing the issue of document length variations. this model could be used in retrieval systems in Arabic in terms of finding the most related documents which are based on semantic similarity, not document length.	Semantic Similarity (Corpus-based)

As can be seen, from Table 4, we observed that measuring semantic similarity for documents with word embeddings technique achieves better results. While for sentence semantic similarity, a hybrid technique that joins Latent Semantic Analysis (LSA) with word co-occurrence achieves better results. For word similarity, the feature-based approach provides the best results. In Table 5, we represent the most important textual similarity measuring tools provided in recent years.

TABLE V: CHRONOLOGICAL REPRESENTATION OF THE MOST IMPORTANT SIMILARITY MEASUREMENT TOOLS.

Tool Name	**Is it used String similarity?**	**Is it used Semantic Similarity?**	**Arabic Support**
LingPipe (String Comparison Tool)**12**	Yes	-	Yes
Harry 13	Yes	-	Yes
Tools 4 noobs 14	Yes	-	Yes
Neutered Paranoid Meerkat string similarity**15**	Yes	-	Yes
Smell: A flexible tool for text similarity **16**	Yes	Yes	Yes
SimString **17** Ferret Copy Detection software**17** Aplag**17**	Yes	Yes	Multi-lingual tools.
FREJ means "Fuzzy Regular Expressions for Java" **18**	Yes	-	No
SimMetrics **19**	Yes	-	No
SecondString **20**	Yes	-	No
Sword Algorithm **21**	Yes	-	Multi-lingual tool
java-string-similarity tool **22**	Yes	-	Yes
SEMILAR: A Semantic Similarity Toolkit (2013) **23**	Yes	Yes	Yes, Multi-lingual
SML-TOOLKIT**24**	-	Yes	Yes
DISCO 25	-	Yes	Yes, Multi-lingual
Retina API **26**	-	Yes	Yes, Multi-lingual
Turnitin **27**	-	Yes	Yes, Multi-lingual
QARNET **28**	-	Yes	Arabic English

[12] http://alias-i.com/lingpipe/demos/tutorial/stringCompare/read-me.html

[13] http://www.mlsec.org/harry

[14] https://www.tools4noobs.com/

[15] https://www.npmjs.com/package/string-similarity?activeTab=readme

[16] https://www.researchgate.net/publication/321685579_SimAll_A_flexible_tool_for_text_similarity

[17] http://www.chokkan.org/software/simstring/

[18] http://frej.sourceforge.net/

[19] https://sourceforge.net/projects/simmetrics/

[20] https://www.sourceforge.net/projects/secondstring/

[21] https://icann.sword-group.com/algorithm/

[22] https://github.com/tdebatty/java-string-similarity

[23] http://www.semanticsimilarity.org/

[24] https://omictools.com/semantic-measures-library-toolkit-tool

[25] http://www.linguatools.de/disco/disco_en.html

[26] http://www.cortical.io/product_retina_api.html

[27] http://turnitin.com/ar/

[28] http://portal.sinteza.singidunum.ac.rs/Media/files/2016/173-178.pdf

In Table 5, we classify these tools based on the type of textual similarity they provide and if they support the Arabic Language or not.

3 Conclusion

In this paper, we presented a chronological overview of textual similarity measuring in a variety of languages. Generally, we found out that the textual similarity is divided into three categories: Lexical-based similarity, Semantic-based similarity, and Hybrid similarity. Then we shed light on semantic analysis in the Arabic Language, which we can divide into four types: Word co-occurrence approach, Latent Semantic Analysis approach, feature-based approach, and hybrid-based approach. Word co-occurrence approach that ignores the term order of the sentence, and it does not take into consideration the meaning of a term according to its context. But it successfully extracts keywords from documents. The Latent Semantic Analysis (LSA) seems like a complete model of language understanding, and it is a successful approach in information extraction, especially for documents, but it ignores word order and function words. Also, this approach is based on Singular Value Decomposition (SVD), which is computationally expensive, and it is difficult to update as new documents appear. The third type is the features-based approaches in which a word in a short text represented using semantic features based on dictionaries or WordNet which means, a high-quality resource is needed, and this is not always available. Finally, the Hybrid-based approach that used neural networks and word embeddings which have two limitations for short texts; the first one is that word embedding does not consider term order and the second one is that it is unable to capture polysemy; it cannot learn separate embeddings for multiple senses of a word. In the future, we need further researches to enhancing the accuracy of Arabic similarity measurements as achieved in English Languages.

References

[1] S. Brand-Gruwel and M. Stadtler, "Solving Information-based Problems: Searching, Selecting and Evaluating Information," *Journal of the European Association for Research on Learning and Instruction (EARLI)*, vol. 21, pp. 175–179, 2011.

[2] A. Magooda *et al.*, "RDI_Team at SemEval-2016 Task 3: RDI Unsupervised Framework for Text Ranking," *SemEval 2016 - 10th Int. Work. Semant. Eval. Proc.*, pp. 822–827, San Diego, California, June, 2016.

[3] P. Nakov *et al.*, "SemEval-2017 Task 3: Community Question Answering," *SemEval 2017 - 11th Int. Work. Semant. Eval. Proc.*, pp. 27–48, Vancouver, Canada, August, 2018.

[4] A. M. Al-Zoghby and K. Shaalan, "Ontological Optimization for Latent Semantic Indexing of Arabic Corpus," *Procedia Comput. Sci.*, vol. 142, pp. 206–213, 2018.

[5] S. S.Sonawane and P. A. Kulkarni, "Graph based Representation and Analysis of Text Document: A Survey of Techniques," *International Journal of Computer Applications(IJCA)*, vol. 96, no. 19, pp. 1–8, 2014.

[6] S. H. Mustafa and Q. A. Al-Radaideh, "Using N-grams for Arabic text searching," *The Journal of the Association for Information Science and Technology (JASIST)*, vol. 55, no. 11, pp. 1002–1007, 2004.

[7] W. H.Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," *International Journal of Computer Applications(IJCA)*, vol. 68, no. 13, pp. 13–18, 2013.

[8] S. S. Aljameel, J. D. O'Shea, K. A. Crockett, and A. Latham, "Survey of string similarity approaches and the challenging faced by the Arabic language," *Proc. 2016 11th Int. Conf. Comput. Eng. Syst. ICCES 2016*, pp. 241–247, Cairo, Egypt, December, 2017.

[9] M. Lintean and V. Rus, "Measuring semantic similarity in short texts through greedy pairing and word semantics," *Proc. 25th Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS-25*, pp. 244–249, Marco Island, US, May, 2012.

[10] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," *Proc. Natl. Conf. Artif. Intell.*, vol. 1, pp. 775–780, United States, 2006.

[11] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Trans. Knowl. Discov. Data*, vol. 2, no. 2, pp. 1–25, 2008.

[12] L. Abouenour, K. Bouzouba, and P. Rosso, "An evaluated semantic query expansion and structure-based approach for enhancing Arabic question / answering," *Int. J.*, vol. 3, no. 3, 2010.

[13] L. Dai and H. Huang, "An English-Chinese Cross-lingual Word Semantic Similarity Measure Exploring Attributes and Relations," *25th Pacific Asia Conf. Lang. Inf. Comput.*, pp. 467–476, Singapore, December, 2011.

[14] M. M. Refaat, a a Ewees, and M. M. Eisa, "Automated Assessment of Students ' Arabic Free-Text Answers," *International Journal of Intelligent Computing and Information Sciences (IJICIS)*, vol. 12, no. 1, pp. 213–222, 2012.

[15] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *journal of Artificial Intelligence (AIJ)*, vol. 193, pp. 217–250, 2012.

[16] W. H. Gomaa and A. A. Fahmy, "Short Answer Grading Using String Similarity And Corpus-Based Similarity,"

International Journal of Advanced Computer Science and Applications (IJACSA), vol. 3, no. 11, pp. 115–121, 2012.

[17] N. Aggarwal, K. Asooja, and P. Buitelaar, "DERI&amp;UPM: Pushing Corpus Based Relatedness to Similarity: Shared Task System Description," *First Jt. Conf. Lex. Comput. Semant. (*SEM), pages 643–647,* pp. 643–647, Montr. Canada, June 7-8, 2012.

[18] D. Bär, T. Zesch, and I. Gurevych, "Text Reuse Detection Using a Composition of Text Similarity Measures Erkennung von Textwiederverwendung durch Komposition von Textähnlichkeitsmaßen," *Proc. of COLING ,* pp. 167–184, Mumbai, India, December, 2012.

[19] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning, "Bilingual Word Embeddings for Phrase-Based Machine Translation," *Proc. ofthe 2013 Conf. Empir. Methods Nat. Lang. Process.,* pp. 1393–1398, Seattle, WA, USA, October, 2013.

[20] M. Kaleem, J. D. O'Shea, and K. A. Crockett, "Word order variation and string similarity algorithm to reduce pattern scripting in pattern matching conversational agents," *2014 14th UK Work. Comput. Intell. UKCI 2014 - Proc.*, Bradford, UK, May, 2014.

[21] F. Elghannam, "Automatic Measurement of Semantic Similarity among Arabic Short Texts," *Commun. Appl. Electron.*, vol. 6, no. 2, pp. 16–21, 2016.

[22] J. Tian, Z. Zhou, M. Lan, and Y. Wu, "ECNU at SemEval-2017 Task 1: Leverage Kernel-based Traditional NLP features and Neural Networks to Build a Universal Model for Multilingual and Cross-lingual Semantic Textual Similarity," *SemEval 2017 - 11th Int. Work. Semant. Eval. Proc.*, pp. 191–197, Vancouver, Canada, August, 2018.

[23] G. Chen, X. Shi, M. Chen, and L. Zhou, "Text similarity semantic calculation based on deep reinforcement learning," *International Journal of Security and Networks(IJSN)*, vol. 15, no. 1, pp. 59–66, 2020.

[24] S. Kim, I. Kang, and N. Kwak, "Recurrent and Co-attentive Information." *Proc. of the AAAI Conf. Artif. Intell. 33*, pp. 6586-6593, Honolulu, Hawaii, USA, January 27 – February 1, 2019.

[25] H. Khafajeh *et al.*, "Automatic Query Expansion for Arabic Text Retrieval Based on Association and," *Information Retrieval Journal.*, no. October, 2002.

[26] K. Shaalan, M. Attia, P. Pecina, Y. Samih, and J. van Genabith, "Arabic Word Generation and Modelling for Spell Checking," *Proc. Eight Int. Conf. Lang. Resour. Eval.*, pp. 719–725, Istanbul, Turkey, May, 2012.

[27] M. A. Al-Ramahi and S. H. Mustafa, "N-Gram-Based Techniques for Arabic Text Document Matching; Case Study: Courses Accreditation," *Basic Sci. Eng.*, vol. 21, no. 1, pp. 85–105, 2012, [Online]. Available: http://journals.yu.edu.jo/aybse/Issues/Vol21No1_2013/07.pdf.

[28] E. M. B. Nagoudi and D. Schwab, "Semantic Similarity of Arabic Sentences with Word Embeddings," *Proc. of Third Arab. Natur. Lang. Process. Work.,* pp. 18–24, Valencia, Spain, April, 2017.

[29] Y. S. Hitachi, "HCTI at SemEval-2017 Task 1 : Use convolutional neural network to evaluate Semantic Textual Similarity," *SemEval 2017 - 11th Int. Work. Semant. Eval. Proc.*, pp. 130–133, Vancouver, Canada, August, 2017.

[30] M. K. A. Al-Maghasbeh and M. P. Bin Hamzah, "Arabic Information Retrieval Using Semantic Analysis of Documents," *International Journal of Computer Science and Network Security(IJCSNS)*, vol. 18, no. 5, pp. 53–58, 2018.

[31] G. Zarrella, J. Henderson, E. M. Merkhofer, and L. Strickhart, "MITRE: Seven Systems for Semantic Similarity in Tweets," *SemEval 2015 - 9th Int. Work. Semant. Eval. Proc.*, pp. 12–17, Denver, Colorado, June, 2015

[32] G. Da San Martino, A. Barrón Cedeño, S. Romeo, A. Uva, and A. Moschitti, "Learning to Re-Rank Questions in Community Question Answering Using Advanced Features," *CIKM '16: Proc. the 25th ACM Int. Conf. Inf. Knowl. Manag.,* pp. 1997–2000, Indianapolis, Indiana, USA, October, 2016.

[33] H. Wu and H. Huang, "BIT at SemEval-2017 Task 1 : Using Semantic Information Space to Evaluate Semantic Textual Similarity," *SemEval 2017 - 11th Int. Work. Semant. Eval. Proc.*, pp. 77–84, Vancouver, Canada, August, 2017.

[34] B. Hassan, S. AbdelRahman, R. Bahgat, and I. Farag, "FCICU at SemEval-2017 Task 1: Sense-Based Language Independent Semantic Textual Similarity Approach," *SemEval 2017 - 11th Int. Work. Semant. Eval. Proc.*, pp. 125–129, Vancouver, Canada, August, 2018.

[35] B. Ghanem, L. Arafeh, P. Rosso, and F. Sánchez-Vega, "HYPLAG: Hybrid arabic text plagiarism detection system," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10859 LNCS, no. October 2016, pp. 315–323, 2018.

[36] E. M. B. Nagoudi, A. Khorsi, H. Cherroun, and D. Schwab, "2L-APD: A two-level plagiarism detection system for Arabic documents," *journal Cybernetics and Information Technologies*, vol. 18, no. 1, pp. 124–138, 2018.

[37] M. Alian and A. Awajan, "Arabic Semantic Similarity Approaches - Review," *ACIT 2018 - 19th Int. Arab Conf. Inf. Technol.*, khalde, Lebanon, November, 2019.

[38] M. AL-Smadi, Z. Jaradat, M. AL-Ayyoub, and Y. Jararweh, "Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features," *International Journal*

Information Processing & Management(IJIPM), vol. 53, no. 3, pp. 640–652, 2017.

[39] A. Mahmoud and M. Zrigui, “Sentence Embedding and Convolutional Neural Network for Semantic Textual Similarity Detection in Arabic Language,” *Arabian Journal for Science and Engineering,* vol. 44, no. 11, pp. 9263–9274, 2019.

[40] M. Konopík, O. Pražák, D. Steinberger, and T. Brychcín, “UWB at SemEval-2016 Task 2: Interpretable semantic textual similarity with distributional semantics for chunks,” *SemEval 2016 - 10th Int. Work. Semant. Eval. Proc.*, pp. 803–808, San Diego, California, June, 2016.

[41] N. Almarwani and M. Diab, “Arabic Textual Entailment with Word Embeddings,” *Proc. of Third Arab. Natur. Lang. Process. Work.*, pp. 185–190, Valencia, Spain, January, 2017.

[42] S.A. Al Awaida, B. Al-Shargabi1 and Th. Al-Rousan, “Automated Arabic Essay Grading System Based ON F-Score And Arabic Wordnet” *Journal of Computer Engineering & Information Technology (JJCIT),* Vol. 05, No. 03, December 2019.

[43] A. El Hadi, Y. Madani, R. El Ayachi, and M. Erritali, “A new semantic similarity approach for improving the results of an Arabic search engine,” *Procedia Comput. Sci.*, vol. 151, pp. 1170–1175, 2019.

[44] A. Mahmoud and M. Zrigui, “Distributional semantic model based on convolutional neural network for Arabic textual similarity,” *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, vol. 14, no. 1, pp. 35–50, 2020.

[45] A. Omar and W. I. Hamouda, “Document length variation in the vector space clustering of news in arabic: A comparison of methods,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, no. 2, pp. 75–80, 2020.

[46] W. Wali, B. Gargouri, A.B. hamadou, "Supervised Learning to Measure the Semantic Similarity Between Arabic Sentences," *Comput. Collect. Intell. Lect. Notes Comput. Sci.,* pp 158-167 vol 9329. Springer, Cham, 2015.

[47] W. Wali, B. Gargouri, and A. Ben Hamadou, “Sentence Similarity Computation based on WordNet and VerbNet,” *Comput. y Sist.*, vol. 21, no. 4, pp. 627–635, 2017.

[48] W. Wali, F. Ghorbel, B. Gragouri, F. Hamdi, E. Metais, "A Multilingual Semantic Similarity-Based Approach for Question-Answering Systems," *KSEM 2019. Int. Conf. Knowl. Sci. Eng. Manag. Lect. Notes Comput Sci.*, pp 604-614, vol 11775. Springer, Cham, 2019.

[49] F. A. Almarsoomi, J. D. O’Shea, Z. Bandar, and K. Crockett, “AWSS: An algorithm for measuring Arabic word semantic similarity,” *Proc. - 2013 IEEE Int. Conf. Syst. Man, Cybern. SMC 2013*, pp. 504–509, Washington, United States, 2013.

[50] S. Abdulateef, N. A. Khan, B. Chen, and X. Shang, “Multidocument Arabic text summarization based on clustering and word2vec to reduce redundancy,” *Inf.*, vol. 11, no. 2, 2020.

BIOGRAPHY

Amira Hamed is a Demonstrator at the Department of Computer Sciences, Faculty of Computers and Information Science, Mansoura University, Egypt. She has obtained her bachelor's degree with an excellent grade in 2016. She was appointed as a Demonstrator on 6/12/2016 in the department of Computer Science. She registered for the M.Sc. with the department of Computer Science in 2017. Her current research areas include Natural Language Processing and its applications.

Assoc. Prof. Dr. Aya M. Al-Zoghby

She is an Associate Professor with the Department of Computer Sciences. She obtained her bachelor's degree with an excellent grade in 2001. She obtained a master's degree in 2008. She was appointed as a Demonstrator on 28/2/2002 in the Department of Computer Science, as Assistant Professor on 23/9/2013, and as an Associate Professor on 1/7/2019.

She worked at Mansoura University- Egypt, 2001 – 2019, Damietta University, Egypt, 2020. She is specialized in the following research domains and related: Natural Language Processing, Semantic Web, Information Retrieval.

Assoc. Prof. Dr. Taher Hamza
He is an Associate Professor with the Department of Computer Sciences, Faculty of Computers and Information Science, Mansoura University, Egypt. He received his BA with an excellent grade with honors in 1975. He received a master's degree in 1979. He was appointed as a Demonstrator on 9/29/1975 with the department of Computer Science, as Assistant Professor on 05/26/1986, and as a full-time Associate Professor on 07/30/2013. He is specialized in the AI domain and related.

ARABIC ABSTRACT

منهجيات قياس التشابه النصي: بحث استقصائي

اميرة حامد أبوالغيط*1, آية محمد الزغبي**2 , طاهر حمزة3*

قسم علوم الحاسب، كلية الحاسبات و المعلومات، جامعة المنصورة، مصر
[1] amira-hamed@mans.edu.eg
[3]taher_hamza@yahoo.com
***قسم علوم الحاسب، كلية الحاسبات و المعلومات، جامعة دمياط، مصر*
[2]aya_el_zoghby@du.edu.eg

ملخص :

تُعد البحوث الاستقصائية أحد أنواع البحوث العلمية الضرورية والمناسبة لتناول أنواع معينة من الأسئلة البحثية **.**في هذه الورقة نهدف إلى تقديم نظرة عامة عن التشابه النصي من الأدبيات والدراسات السابقة **.**أصبح قياس التشابه النصي ذو أهمية متزايدة في المواضع المتعلقة به كتصنيف النصوص واستخراج معلومات محددة من البيانات وتجميعها. واسترجاع المواضيع وتتبعها وكذلك الإجابة على الأسئلة وتقييم المقالات والتلخيص وفي الوقت الحاضر أنظمة المحادثات وهي عبارة عن برامج تتواصل مباشرة مع البشر عن طريق استخدام اللغات الطبيعية المختلفة. يعد أهم جزء في التشابه النصي هو إيجاد التشابه بين المصطلحات ومن ثم استخدامه في إيجاد التشابه بين الجمل ومن ثم الفقرات ثم النصوص الكاملة. في هذا النص تناولنا على وجه الخصوص التشابه النصي في اللغة العربية؛ يمثل تطبيق مهمات معالجة اللغات الطبيعية في اللغة العربية عملية في غاية الصعوبة كونها تحوي العديد من الخصائص. على الرغم من ذلك قد قُدمت العديد من الطرق لقياس التشابه النصي في النصوص العربية والتي سيتم تناول بعضها في هذه الورقة البحثية.

الكلمات المفتاحية:
التشابه الدلالي، اللغة العربية، وورد نت، التشابه الحرفي، التشابه النصي، التشابه الهجين، تضمين الكلمات.