ISSN 2007-9737

Semantic Textual Similarity: Overview and Comparative Study between Arabic and English

Samira Boudaa*, Tarik Boudaa, Anass El Haddadi

Abdelmalek Essaadi University,
National School of Applied Sciences,
LSA Laboratory, Al Hoceima,
Morocco

samira.boudaa@etu.uae.ac.ma, t.boudaa@uae.ac.ma, a.elhaddadi@uae.ac.ma

Abstract. Semantic Textual Similarity is crucial for various end-user applications of Natural Language Processing, including Search Engines, Chatbots, Machine Translation Systems, Plagiarism Detection, and Text Summarization. While substantial research has been conducted on this topic for widely spoken languages such as English, there exists a need for comprehensive surveys focusing on less-studied languages, such as Arabic. This work is a comprehensive resource for researchers working on Semantic Textual Similarity especially for the Arabic language. Our survey synthesizes the current state of research in Semantic Textual Similarity in Arabic, providing valuable insights into this field's unique challenges and opportunities. We review state-of-the-art approaches, datasets, and methodologies proposed for Arabic Semantic Textual Similarity. The paper highlights the differences between Arabic and English, which necessitate tailored approaches to Semantic Textual Similarity. Moreover, we discuss the recent advancements in Arabic Semantic Textual Similarity and identify the existing gaps and challenges that researchers face. In addition, we propose potential future research directions to further improve the Arabic Semantic Textual Similarity models. By addressing these areas, our work aims to foster a deeper understanding and more robust development of Semantic Textual Similarity for the Arabic language, ultimately expanding the scope and effectiveness of Semantic Textual Similarity applications.

Keywords. Semantic textual similarity, question similarity, Arabic NLP.

1 Introduction

In NLP, Semantic Textual Similarity (STS) is a task that aims to quantify the degree of similarity between two texts at the semantic level rather than just their surface form. This similarity taking into account the semantics and contextual meaning is essential for tasks requiring deeper understanding of textual data, such as Information Retrieval (e.g., [1]), Machine Translation (e.g., [2]), Text Summarization (e.g., [3]), Question Answering (e.g., [4]), and Plagiarism Detection (e.g., [5, 6]).

In essence, STS encompasses a hierarchy of tasks, ranging from word-level comparisons to sentence-level analysis and document-level assessments, each contributing to achieve a deeper comprehension of text similarity and enabling various applications across the field of NLP.

STS has a crucial role in the evaluation of language models, and word and sentence representation models. The performance of these models is judged through their performance in STS Benchmark.

While the methods for STS have seen important progress for English, there exists a convincing need to explore the state of the art for languages with distinct linguistic characteristics and cultural backgrounds, such as Arabic. A survey on Arabic STS remains highly relevant despite some existing survey works [7, 8, 9].

Indeed, while these previous works provided valuable insights into the state of STS research for the Arabic language, they miss the latest developments and breakthroughs that have emerged in NLP through deep learning techniques. The present survey attempts to highlight recent works, advancements, challenges, and opportunities, enabling a more comprehensive

doi: 10.13053/CyS-28-3-5035

understanding of the current landscape and inspiring further innovation in this important area of NLP.

It explores the various techniques and methodologies employed to tackle this task, ranging from traditional lexical and distributional approaches to modern deep learning-based methods. By doing so, we aim to shed light on the strengths and weaknesses of each approach in the specific context of the Arabic language. We also review the available Arabic STS datasets.

Furthermore, we address the unique challenges associated with Arabic STS, including the complexity of Arabic morphology, the presence of dialectal variations, and the scarcity of annotated data. We also explore recent advancements and innovations that have emerged to address these challenges, as well as propose potential avenues for future research exploration.

In this survey, the initial parts provide a comprehensive summary of the main approaches in the field leveraging the research conducted in English as the most targeted language in every NLP task. This segmentation aims to establish a foundational understanding of STS before the detailed exploration of the specific advancements made in Arabic STS.

Moreover, this sequential presentation allows for establishing a deep comparative analysis and gives valuable insights into the distinct characteristics and advancements within Arabic STS, facilitating an assessment of the progress and trends in Arabic STS in relation to the more extensively documented English-language research and then offering a nuanced perspective on the state of the field in the Arabic context.

The rest of this paper is structured as follows: In Section 2, we delve into the foundational aspects of STS. Section 3 is dedicated to the exploration of STS in the context of the Arabic language and discusses potential future directions in this field. Finally, we present our conclusions in Section 4.

2 Semantic Textual Similarity Background

In this section, we provide an in-depth examination of STS and its associated tasks, and we present the different approaches that researchers have developed to measure semantic similarity. These approaches encompass a wide spectrum of techniques, ranging from traditional methods to state-of-the-art deep learning models.

2.1 Semantic Textual Similarity and Related Tasks

In NLP, the concept of semantic similarity can be explored across multiple dimension. The basic case is when we study the similarity between words in terms of their meaning. Measuring the similarity between two sentences is more challenging, it needs to extend the analysis beyond individual words and take into account the overall meaning and coherence of the sentences.

More generally, semantic similarity can be also considered at the document level. One of the valuable applications of this case is plagiarism detection which helps users locate documents with related or similar content and maintain the integrity of academic and professional writing.

STS task made its debut at SemEval 2012. The task objective was to automatically assess the semantic similarity between a pair of texts using a predefined scale, such as a range from 0 (not similar) to 5 (completely equivalent) [10]. As Semantic similarity varies across a spectrum, ranging from stark dissimilarity to precise semantic equivalence, employing a graded similarity scale becomes essential to accurately reflect the level of similarity between text pairs.

Furthermore, the research in STS has led to a set of variants and special cases, Short Text Similarity Semantic (STSS) and Semantic Question Similarity are some of these sub-tasks (e.g., [11, 12]). Semantic Short Text Similarity (STSS) is a specialized case of STS that measures semantic similarity between short text fragments. This task is particularly relevant for applications like query reformulation, duplicate detection, and social media analysis, where users express their information needs or opinions in brief, concise formats (e.g., [13]).

Question Similarity can be considered as a case of SSTS where the task focuses on measuring the similarity between questions. This task is essential for various applications, including question-answering systems, community question-

Table 1. Example of semantic relatedness and STS

Text 1	Text 2
Darwin's theory of evolution	*Mendel's laws of inheritance*
A child reading a book	*A child engrossed in a book*

answering platforms, and chatbots, where the goal is to identify questions that share a similar meaning or intent.

Another case is Cross-Lingual STS [14]. In this case, the goal is to measure the semantic similarity between texts in different languages. While STS methods focus on nuances within a single language, Cross-Lingual STS tackles the added complexity of variations in language, structure, and cultural context across different languages.

STS finds applications in various specific domains, including clinical texts (e.g., [15, 16]), legal domain (e.g., [17, 18]), and scientific literature (e.g., [19]). While general-purpose STS measures may not perform effectively in certain specialized domains, domain-specific STS measures have been developed. In these specific domains, STS enhances efficiency and accuracy by assisting in content selection, summarization, and information retrieval, ultimately serving as a valuable tool for professionals in each field.

For instance, Clinical STS can assist in identifying relevant patient records, extracting essential information from unstructured medical documents, and generating concise, accurate summaries of patient histories which can help in clinical decision-making (e.g., [16]).

Another variant of STS is Interpretable Semantic Textual Similarity (iSTS) which aims to develop STS systems with the ability to understand and explain the similarity between texts in a clear and interpretable manner.

Recently, Deshpande et al. [20] introduced a new task named Conditional STS (C-STS) which assesses sentence similarity based on a feature described in natural language, referred to as a condition. C-STS aims to reduce the subjectivity and ambiguity of STS, allowing for nuanced language model evaluation across various natural language conditions.

Numerous other NLP tasks share similarities with STS, and although they tackle distinct challenges, it is notable that the methods employed in the literature largely remain consistent across these tasks. It's conceivable that the closest task to STS is Semantic Textual Relatedness. The key difference between these two concepts is that semantic textual relatedness assesses the level of association or connection between texts, while STS quantifies the degree of similarity in meaning or content between texts.

To provide further clarity regarding this distinction, we examine the two examples presented in Table 1. In the first example, the two texts present a degree of relatedness without sharing substantial content or meaning, whereas in the second case, the texts express a common idea with slight linguistic variances, thereby there is a high degree of similarity between these texts.

The tasks of Recognizing Textual Entailment, Machine Translation Evaluation, and Paraphrase Detection exhibit also some similarities with STS. While all of these NLP tasks deal with understanding and comparing text, their specific objectives and applications vary. STS measures text similarity, Machine Translation Evaluation assesses translation quality, Textual Inference evaluates logical relationships, and Paraphrase Detection identifies rephrased content, each contributing to the broader landscape of text analysis and language understanding.

2.2 State-of-the-art Approaches for Semantic Similarity

This subsection gives a summary of the main methods used to deal with the STS for English.

2.2.1 Knowledge-based Semantic-similarity Methods

Knowledge-based methods for STS use structured knowledge sources, such as ontologies, knowledge graphs, taxonomies, or dictionaries, to quantify the semantic similarity between texts. These approaches leverage the information contained in knowledge bases, including relationships between concepts as well as their definitions.

Among the most frequently employed knowledge bases in STS measurements are WordNet [21], Wikipedia, and BabelNet [22].

Knowledge-based semantic similarity methods can use the distance between nodes (concepts) within an ontology to measure the similarity, a technique commonly referred to as the path length or edge-counting method. In this approach, the shorter the distance between two concepts in the ontology, the greater their perceived similarity.

Another path length method assesses similarity based on the proximity of their Least Common Subsumer, which is defined as their closest common ancestor. For instance, the Li similarity measure proposed by Li et al. [23] combines both of these preceding techniques.

In addition to path length methods, feature-based and information content methods are also employed. feature-based methods (e.g., [24]) are used to assess semantic similarity by comparing the attributes or features of concepts. This approach involves examining shared characteristics between concepts to determine their similarity.

The information content is a measure of the amount of information or specificity associated with a word or concept in a given knowledge base. Highly specific words are associated with a low information content value, whereas more general and frequently occurring words have higher information content [25]. The semantic similarity between two concepts is determined by the amount of information shared by these concepts.

Whilst in feature-based measures, terms are represented as sets of features, an example of work applying this approach is [24]. This approach involves examining shared characteristics between concepts to determine their similarity. The degree of similarity is proportional to the number of common features they share and inversely related to the presence of distinctive features [26].

2.2.2 Feature Engineering and Handcrafted Representations

Supervised approaches with feature engineering in NLP tasks are based on transforming the input text to a vector of informative features that capture relevant linguistic patterns and semantic information. Through a set of experimentation and fine-tuning of feature engineering techniques, supervised NLP classification models can effectively learn to make good predictions based on the engineered features. These approaches are very well explored in many NLP tasks for several languages. They prove particularly advantageous when working with limited data resources. In the context of STS, these features may capture lexical, syntactic, or semantic information.

2.2.3 Deep Learning Paradigms in STS

Deep neural network-based methods have proven effective across various NLP tasks, especially in the case of STS. In this context, various deep learning models and architectures are explored to capture complex semantic features, facilitating precise quantification of textual similarity.

Recurrent neural networks and Siamese architectures: GRUs, LSTM, Bi-LSTM networks, and their variants have emerged as powerful tools in the realm of NLP tasks, particularly in the context of STS. These models with their ability to represent the sequential structure within textual data, yield enriched representations that effectively encapsulate the underlying meaning. These enriched representations are used to capture the similarity between texts.

Additionally, attention mechanisms have been integrated with RNNs to enable the networks to focus on relevant parts of the input, thereby improving their ability to discern subtle semantic nuances. Moreover, attention-based architectures used with RNN further enhance accuracy by focusing on relevant parts of the sentences. RNNs and their variants are also frequently combined with other neural network architectures like Convolutional Neural Networks (CNNs) within the same model to deal with STS (e.g., [27]).

Siamese neural networks are an architecture containing typically two identical sub-networks. Siamese networks aim to capture a meaningful representation of input pairs, ensuring that similar inputs are positioned closely to each other in the learned feature space. This style of architecture is suitable for tasks related to similarity measurement between two comparable things. This architecture is widely used in the literature for the task of STS (e.g., [28, 29, 30]).

Word and Sentence Embedding: In NLP, the words are commonly presented as continuous vectors in a multi-dimensional space, where words with similar meanings or contexts are located closer to each other. This representation captures semantic relationships between words. The early versions of word embedding were static like Word2Vec [31] or GloVe [32] each word or token in the vocabulary is associated with a pre-computed vector representation, and these vectors do not get updated or fine-tuned as the model learns from task-specific data or in the context.

Unlike traditional word embedding contextualized word embeddings, which are context-sensitive, provide different embeddings for a word depending on its context in a given sentence or document. This allows to capture the meaning of a word in a specific context within a sentence or text. Two prominent examples of models that generate contextualized word embeddings are ELMo [33] which uses a Bi-LSTM network to compute word embeddings based on the entire sentence's context.

ELMo embeddings capture syntactic and semantic nuances, making them valuable for various NLP tasks. Building on this foundation, BERT [34], a transformer-based model, further improved contextualized word embeddings.

It employs a masked language modeling objective to understand words in their surrounding context. BERT embeddings have gained widespread popularity and have been fine-tuned for numerous downstream NLP tasks, achieving state-of-the-art results.

Sentence embedding models have proven to perform well on various STS benchmarks, many models have been developed such as InferSent [35], which is trained on the SNLI dataset for natural language inference and is shown to perform well on STS. Another example of sentence embedding is the "Universal Sentence Encoder" [36], which has shown its effectiveness on a wide range of NLP tasks, including STS. Another recent sentence embedding mode is SimCSE [37]. The key idea behind it is to pre-train sentence embeddings in a contrastive learning framework, where semantically similar sentences are brought close together in the embedding space.

Transformer-based Models: The core concept of the Transformer architecture [38] is to substitute conventional recurrent and convolutional layers with a self-attention mechanism. This mechanism enables the model to prioritize different segments of the input sequence dynamically, enhancing its predictive capabilities. Self-attention is the key concept of the Transformer architecture. This mechanism calculates attention scores for each word/token in the input sequence based on its relationships with other words.

It allows the model to effectively capture long-range dependencies and contextual information. The computation in Transformers can be highly parallelizable, making them computationally efficient. This contrasts with RNNs, where computations are sequential, limiting their ability to utilize modern hardware effectively. Since the self-attention mechanism lacks inherent positional information, Transformers incorporate positional encodings to provide the model with information regarding the sequence of words in the input.

Transformer-based models like BERT [34] and RoBERTa [39], T5[40], and GPT [41] and their variants have achieved remarkable results in various NLP, particularly on STS. These models capture contextual word meanings and encode sentences into informative vectors. Sentence similarity can be measured by the cosine distance between these vector representations. Fine-tuning these models on STS datasets or using them as feature extractors can yield excellent results.

As a result, large language models like GPT-4 have become valuable tools offering enhanced performance in many NLP tasks including STS applications. The continuous advancements in transformer-based architectures contribute significantly to the improvement of STS models.

2.2.4 Cross-lingual Approaches

To tackle cross-lingual semantic similarity, two main methods are commonly employed: Machine Translation approaches and shared semantic space approaches.

Machine Translation (MT) is widely used to address the cross-lingual STS task. This involves converting the two texts under comparison into the same language, to apply a monolingual similarity approach. For instance, Tian et al. [42] proposed a method to deal with STS for Arabic-English, Spanish-English, and Turkish-English sentence pairs.

Their method relies on Machine Translation to convert the sentences into English. Subsequently, they apply a hybrid approach that combines both supervised learning and deep learning techniques to establish a semantic similarity measure. They used a set of features as input for the supervised machine learning module, including MT evaluation metrics, along with classical similarity features.

The limitation of MT-dependent approaches is their inapplicability to under-resourced languages lacking an efficient MT system. Therefore, some studies have introduced semantic similarity models across different languages that do not rely on Machine Translation.

To address this, various works have proposed methods based on a shared semantic space approach for different languages. Those methods articulate on representing words from different languages in a shared embedding space by training monolingual semantic representations independently of each other, then using a translation matrix, projecting word vector representations of one language into the representation space of the other language.

The translation matrix is computed using a small set of word pairs consisting of words of one language and their translation in the other language. subsequently, the similarity between words of each sentence is obtained by using traditional metrics such as the cosine similarity of their vectors within the shared embedding space. Monolingual approaches are then applied to compute the similarity between sentences (e.g., [43, 44]).

2.2.5 Hybrid Approaches

Hybrid methods for STS represent an integration of multiple techniques for similarity measurements between pieces of text. Typically, hybrid methods incorporate handcrafted features derived from linguistic and semantic analysis, to capture specific linguistic phenomena. Simultaneously, they leverage the power of machine learning algorithms, including deep neural networks, to autonomously learn complex patterns and representations from large datasets.

This combination enables hybrid models to capitalize on the interpretability of rule-based systems and the capacity of machine learning models to discern complex relationships. By merging these diverse methodologies, hybrid approaches aim to overcome the limitations of individual techniques and capitalize on the strengths of each.

As an example of studies using hybrid measures in Semantic Similarity, Panchenko and Morozova [45] combined a set of knowledge-based measures using Wordnet, corpus-based measures using the web, and different classical corpora, in addition to dictionary-based measures using glosses from Wikitionary, Wordnet, and Wikipedia, all combined with supervised learning to achieve better performance.

Moreover, Rychalska et al. [46] proposed a hybrid textual similarity model, incorporating recursive auto-encoders along with penalty or reward scores derived from WordNet. This model was combined with other similarity models in an ensemble to boost its performance.

3 Arabic Semantic Textual Similarity

In this section, we discuss the challenges related to the Arabic language, we explore the various approaches and datasets employed in the literature for measuring semantic similarity in Arabic text. Furthermore, we provide insights into potential future directions to advance research in this domain.

3.1 Arabic Language and its Challenges

The Arabic language consists of multiple variants. The Modern Standard Arabic (MSA), is regarded as the standardized form and serves as the official language variant and written standard across all Arab nations. Moreover, it predominantly serves as the primary mode of communication for public speaking, media, and education. MSA presents many challenges for NLP due to many levels of ambiguity.

These challenges are well-studied in the literature. In the next, we present a summary of some of these challenges. A more detailed presentation of the morphological and syntactic challenges of MSA can be found in [47] and [48]. One of the difficulties associated with MSA is the lack of diacritics. This difficulty is more complicated when it is associated with the inflectional nature of the Arabic language.

Table 2. Example of sentence ambiguity due to the absence of diacritics

Arabic sentence	English Translation
جاء يومَ العيد *jA' ywma AlEyd*	He came on the day of Eid
جاء يومُ العيد *jA' ywmu AlEyd*	The day of Eid came

Table 3. Example illustrating the pro-drop property of the Arabic language

Arabic sentence	Literal translation	English translation
ساعد غيرك، يساعدك sAEd gyrk, ysAEdk	Help others, help you	If you help another, he helps you

Table 4. Numerous clitic items

Arabic sentence	English Translation
والي	Ruler
و+ال+ي w+Al+y	And to me
و+ألي w+Oly	And I follow
و+آل+ي w+|l+y	And my clan
و+آلي w+|ly	And automatic

For instance, the following example illustrates a sentence's ambiguity arising from the absence of diacritics:

Another challenge faced by Arabic is the absence of capitalization which makes for instance the task of named entity recognition more difficult than in English.

Additionally, the task of sentence boundary detection is more challenging in Arabic since texts written in Arabic do not follow strict punctuation rules [47].

Another characteristic of Arabic that complicates automatic processing is the pro-drop property of the Arabic language. (i.e. a language that allows the omission of certain pronouns when they can be inferred from context) as demonstrated by the following example:

MSA exhibits high inflectional complexity due to its rich system of concatenation, which significantly complicates morphological analysis. Arabic words are structured around roots rather than stems [49].

Proclitics in linguistics are clitics that come before a word, resembling prefixes, such as the Arabic conjunction و (w) meaning 'and' or the definite article ال (Al) meaning 'the'. On the other hand, enclitics are clitics that follow the word, akin to suffixes, like the Arabic object pronoun هم (hm) meaning 'them'.

Arabic allows for multiple affixes and clitics within a single word. For instance, the word وسيكتبونها (wsyktbwnhA) contains two proclitics, one circumfix, and one enclitic [48]. For example (cf. Table 4), the word والي can be analyzed in five different ways [48]. Each of these cases has a different discretization.

Arabic NLP is still an underdeveloped field and then it suffers from a lack of open-source libraries, sufficient resources, and large corpora needed generally for many tasks on NLP and especially on STS.

Despite that the MSA is considered the official variant for all Arab countries. The language used in everyday communication in the Arab world is local dialects. These dialects pose numerous challenges due to their rich linguistic diversity and significant variation from MSA.

This variation encompasses differences in vocabulary, grammar, pronunciation, and even script, all compounded by the scarcity of annotated data. Additional complexities arise from code-switching and the absence of standardized norms. For example, in dialectal Arabic, there aren't consistent standard rules for vocabulary and spelling in written form.

3.2 Approaches Applied for Semantic Textual Similarity in Arabic

In this subsection, we explore the various approaches for measuring semantic similarity in Arabic text, including lexical, distributional, and deep learning-based methods. We discuss the strengths and limitations of each approach. Our objective is to provide a thorough and insightful overview of the present research landscape in Arabic STS, with a focus on the most relevant contributions and the latest advancements in the field.

3.2.1 Feature Engineering and Static Word Embedding with Similarity Measures

An early work that addressed sentence similarity was presented by Wali et al. [50]. They used data created using the word definition extracted from a collection of dictionaries designed for human users. The designed features cover lexical, semantic, and syntactic-semantic levels.

Furthermore, Hammad et al. [51] employed a supervised machine learning approach, incorporating feature engineering that encompasses morphological, semantic, and lexical-based features, to deal with the task of Semantic Question Similarity. Various machine learning algorithms, including SVM and AdaBoost, were experimented with. The Mowdoo3 dataset was employed as the experimental dataset.

Additionally, building on their prior work, the same authors in another study [52] opted for a classical machine learning approach by focusing on feature engineering of semantic, lexical, word embedding, morphological, word-level, and character-level features.

These features are designed to capture various aspects of textual information. The authors employed the XGBoost algorithm within a supervised machine learning framework, leveraging its robustness and effectiveness in handling complex feature sets. This combination of feature-rich representations and XGBoost's modeling prowess constitutes the core methodology employed in their study.

Correspondingly, to address the task of Semantic Question Similarity, Lichouri et al. [53] used a collection of n-gram features and lexical features and employed a variety of classifiers. Sharifi et al. [54] conducted their experiments using among others a set of shallow lexical similarity, word embedding, sentence embedding, word mover’s distance, and POS tag overlap.

The approach of static word embedding and similarity measures involves a set of measures applied to the vector representations of the input texts. The vectorization is made by replacing each token with its word embedding vector. Then the complete input text is represented as a sum of these vectors or a weighted sum of these vectors.

Several similarity measures such as cosine similarity, and Euclidean distance, were explored. An example using this approach is the work of Ferrero et al. [55]. They used a CBOW word representation for the Arabic model proposed by Zahran et al. [56].

The approach is based on summing the representation of the words of the sentence without weight or with weights depending on POS and the Inverse Document Frequency. The system outputs a float number ranging from ”0” (representing complete independence of sentence meanings) to ”1” (signifying meaning equivalence).

To conduct their experiments, they used a dataset comprising 750 pairs of sentences drawn from publicly Microsoft Research Video Description Corpus (MSR-Video) (MSRvideo, 2016), which were then manually translated into Arabic.

3.2.2 STS based on Siamese Neural Networks

Siamese neural networks find application in various NLP tasks, including paraphrase identification, and textual entailment, and have also been employed in several prior works for STS, especially for English [29], [57]. Furthermore, they are valuable for tasks involving comparisons between two inputs, extending their utility beyond NLP to domains like facial comparison, image retrieval, and visual object tracking [58].

These type of neural networks find their applications also in Arabic STS, as evidenced by the works of Hammad et al. [52], Othman et al. [59], Einea and Elnagar [60], and Lichouri et al. [53]. These works are elaborated upon below.

Hammad et al. [52] addressed the task of Semantic Question Similarity using a deep learning-based approach, specifically utilizing a Siamese-based recurrent architecture (bi-directional LSTM) trained with pre-defined features and a pre-trained deep bidirectional transformer based on the BERT model. The task is cast as a binary classification, distinguishing between similar and not similar pairs.

Table 5. Available Arabic STS datasets

Dataset	Year	Type	Language	Size	Scale
AWSS	2012	Word Semantic Similarity	MSA	70 pairs	[0,4]
SemEval 2016-task 3 (CQA-MD)	2016	Question Semantic Similarity	Arabic	1,531 questions and 45,164 related question/answer pairs	(Direct, Related, Irrelevant)
SemEval 2017-task 1	2017	Sentence Semantic Similarity / Cross-Lingual STS	MSA-English	2,435 pairs: 2185 for training and 250 for evaluation	[0,5]
SemEval 2017-task 1	2017	Sentence Semantic Similarity	MSA	1,354 pairs: 1104 for training and 250 for evaluation	[0,5]
Mawdoo3 Q2Q	2019	Question Semantic Similarity	MSA	15,712 pairs: 11997 for training and 3715 for evaluation	(0,1)
ASSD	2021	Sentence Semantic Similarity	MSA	887 pairs	[0,1]
Datasets presented by Al Sulaiman et al. [70]	2022	Sentence Semantic Similarity	MSA	1,379 pairs	[0,5]
			Egyptian Arabic	1,379 pairs for training and 250 for evaluation	[0,5]
			Saudi Arabic	1,379 pairs for training and 250 for evaluation	[0,5]
SemEval 2022-task 8	2022	Document Semantic Similarity	MSA	572 article pairs: 274 for training and 298 for evaluation	[1,4]

Table 6. Results comparison on identical dataset and evaluation settings

Study	F-score
Fadel et al. [63]	96.499 %
Al-Theiabat and Al-Sadi [65]	95.924%
Al-Bataineh et al. [62]	93.00%
Hammad et al. [52]	92.99%
Sharifi et al. [54]	82.58%
Lichouri et al. [53]	79.89%

The Mawdoo3 dataset serves as the dataset for their experiments. Moreover, Othman et al. [59] tackled the task of Semantic Question Similarity in a retrieval setting for both languages English and Arabic.

The approach relies on a deep learning architecture, specifically employing a Siamese-based framework with LSTM enhanced with Attention Mechanism. They also explored the utilization of CNNs incorporated within the Siamese architecture to retrieve pertinent questions. The questions texts are vectorized using Word2Vec CBOW. The evaluation is based on the dataset released by Othman et al. [61] for English. For Arabic, a translation of this same English collection was made using Google Translate with a post-manual verification.

The publicly accessible Quora Question Pairs dataset was employed for training the Siamese LSTM model. The input text is tokenized and vectorized using word2vec embeddings, each comprising 300 dimensions, which were trained on a corpus of 100 billion words. For Arabic, Word2Vec training was conducted using an

English dataset that was translated using Google Translate. Additionally, The work of Einea and Elnagar [60] is also based on a neural network architecture based on Siamese with different types of neural networks based on CNN and RNN. The input text is represented using a static word embedding based on Word2vec trained on two different datasets, namely the Mawdoo3 Q2Q dataset (described later in this paper) for assessing question pairs' similarity, and the Semeval-2016 Task 3 dataset, composed of query questions. Each query question is associated with a list of around 30 question/answer pairs, which vary in their degree of similarity to the query.

3.2.3 Contextualized Word Embedding and Transformers

Contextualized word embeddings are a type of word representation that captures the meaning of a word based on its context within a given sentence or document. In contrast to conventional static word embeddings, which assign a constant vector to each word without considering context, contextualized embeddings dynamically change based on the surrounding words and the overall context in which the word appears. Language models like BERT, ElMo, and GPT leverage the potency of contextualized word embeddings to achieve a heightened understanding of context, empowering them to excel in diverse linguistic tasks.

An instance of employing this method is demonstrated in the research by Al-Bataineh et al. [62] by training Embeddings from Language Models (ELMo) on a text corpus comprising both MSA and dialectic sentences, alongside a fine-grained pairwise similarity layer integrated to enhance the question-to-question similarity model, ensuring accurate predictions across different dialects, even though it has been exclusively trained on question-to-question MSA data.

The Mawdoo3 Q2Q Dataset served as the training data, while the test set was constructed using the MADAR dataset through the extraction of Q2Q pairs specifically focused on Arabic dialects. The training of word embedding is performed using a large dataset aggregated from three diverse sources: Tweets, Arabic Wikipedia, and Mawdoo3 articles.

Furthermore, Hammad et al. [52] utilized the BERT model to generate embeddings for question pairs. By encoding question pairs with BERT to produce a high-dimensional representation that retained semantic nuances.

Using the same dataset as the two previously mentioned works, Fadel et al. [63] performed some operations of data augmentation to enlarge the training dataset size. Subsequently, they used contextualized word embedding to represent the input. The sequence of vectors is fed to a special case of LSTM called ONLSTM [64] with self- attention.

Following the extraction of representations for each question, a function is employed to compute squared distances between vectors representing questions within each pair, facilitating their merging into a single vector. The result is then input into a deep fully connected neural network with a sigmoid output layer to produce the final binary decision.

Al-Theiabat and Al-Sadi [65] experimented using different deep learning models, a CNN-based, an RNN (bi-directional GRU), a multi-head attention network model, and a BERT model. The CNN model begins by encoding words, and subsequently, each question undergoes processing through three successive layers. In each of these layers, a convolutional operation is applied, followed by an activation function, and then max pooling.

Consequently, the output for each question is a feature representation, and the similarity label is determined by calculating the cosine similarity between the features of the two questions.

The input in this case consists of pairs of questions that are merged into one sequence. This sequence is then represented using the dictionary and fed into a bi-directional GRU neural network, which eventually generates the similarity label as the output.

In the case of the BERT model, the approach involved fine-tuning the multilingual model by employing a sentence pairs classification task specifically with Arabic questions.

Saidi et al. [66] investigated the integration of Arabic BERT models in Siamese neural networks to deal with sentence similarity. Their system assigns a discrete similarity score on a scale from 0 to 5, where 0 indicates complete semantic independence and 5 denotes semantic equivalence. Their system comprises a BERT-based Siamese Network that incorporates contextual embeddings from BERT, the attention mechanism, and the Siamese neural network.

The study explored various Arabic BERT models for embedding input sentences, including AraBERT [67], Arabic-BERT [68], CAMeL-BERT, and the multilingual mBERT, which is capable of handling Arabic texts [69]. The validation of their approach was conducted using Arabic STS datasets from the SEMEVAL 2017 Multilingual STS. The araBERT-based Siamese Network model achieved a Pearson correlation of 0.925 demonstrating the effectiveness of their approach.

In another study by Al Sulaiman et al. [70], transfer learning and knowledge distillation techniques were employed. The authors proposed three strategies for developing STS models for MSA, Egyptian Arabic, and Saudi Arabic. The first strategy involved using automatic machine translation to convert English data from SNLI [71] and MultiNLI [72] datasets into Arabic.

These translated datasets were then used in the fine-tuning stage to adapt Arabic BERT models into STS Arabic ones. The second approach focused on integrating Arabic BERT models with English data sources to enhance Arabic STS models. Lastly, the third approach aimed to improve the performance of knowledge distillation-based models by fine-tuning them with the use of a translated dataset, specifically tailored to enhance their performance in Arabic.

This study encompassed MSA and two Arabic dialects, Egyptian and Saudi Arabian, and proposed valuable datasets through the professional translation of 1.3K sentence pairs from English to MSA, Egyptian Arabic, and Saudi Arabic. The proposed MSA models were evaluated on the SemEval-2017 Arabic evaluation set [73], while the dialect models were tested on a translated version of this dataset crafted by native speakers of both dialects.

3.3 Arabic STS Datasets

Despite the presence of numerous publicly accessible STS datasets in English, there is still a considerable deficiency in both the number and size of such datasets in the Arabic language. In the following subsection, we present a set of datasets used in Arabic STS research. Our selection is focused on datasets that are openly accessible to the research community, notably those that have served as evaluation benchmarks for various Arabic STS systems. For STS at the word level, Almarsoomi et al., 2012 [74] have created a benchmark dataset, referred to as the AWSS dataset, which comprises 70 pairs of Arabic words. These pairs have been annotated by 60 native Arabic speakers.

For sentence similarity, in SemEval-2017, Cer et al. [73] provided two distinct datasets. One dataset was to evaluate the Cross-lingual Arabic-English Semantic Similarity, while the other one was for Arabic-Arabic Semantic Similarity. The pairs of sentences were retrieved from diverse English resources. These sentence pairs were subsequently annotated with STS labels and then translated into Arabic. Notably, the translation process was carried out independently from their corresponding pairs.

Translators were provided with an English sentence and its machine-generated Arabic translation, and their task involved correcting any errors before transferring the similarity scores. Another dataset for Arabic sentences was developed by Dahy et al. [75] named the ASSD dataset using collected sentences from Arabic Wikipedia, World Wide Web pages, and The Intermediate Lexicon.

The collected dataset covers different domains. The ASSD dataset underwent a manual evaluation process by seven annotators, who assigned values between 0 and 1 to each sentence pair. Furthermore, Al Sulaiman et al. [70] provided three datasets encompassing sentence pairs translated by experts from the SemEval 2017 English STS dataset into MSA, Egyptian dialect, and Saudi dialects. For Question Semantic Similarity, Seelawi et al. [11] produced the Mawdoo3 Q2Q dataset.

It contains pairs of questions labeled by annotators with 1 if they are semantically similar or 0 otherwise. Another Dataset was used for the study of Question Semantic Similarity, referred to as CQA-MD, was developed by Nakov et al. [76] as part of SemEval 2016, the dataset was released for the Arabic subtask aiming to rank pairs of question and answer, retrieved from Community Question Answering platforms, according to their relevance to a new question.

Table 7. Different Arabic STS models based on deep learning or pre-trained models

	Type of Arabic	Task	Notable Elements in the approach	Word Representation	Transformers	Attention Mechanism	Evaluation data	Results
Ferrero et al. [55]	MSA	Sentence Similarity	Calculate the cosine similarity between two sentences by summing their weighted or unweighted word vectors.	Pretrained CBOW Word2Vec model	NO	NO	Arabic STS datasets from the SEMEVAL 2017	Pearson correlation: 76.67%
Fadel et al. [63]	MSA	Q2Q	- Ordered Neurons LSTM[64] - Sequence weighted attention	Arabic ELMo	NO	YES	Mawdoo3 Q2Q dataset with Data Augmentation	F1-score: 96.499 %
Al-Bataineh et al. [62]	MSA + Arabic Dialect	Q2Q	Several models based on: - Word Embedding layer followed by LSTM or RandLSTM [79], Focus Layer/ Dot Product & Absolute Distance - Sent2Vec, Focus Layer/ Dot Product & Absolute Distance	- Word2Vec (AraVec) and ELMo, trained on Arabic Wikipedia, Mawdoo3, and Twitter - Sent2Vec	NO	NO	Mawdoo3 Q2Q Dataset + MADAR dataset	- Best model on Mawdoo3 Dataset (ELMo + TrainableLSTM + DPAD): F1-score 0.93) - Best model on MADAR dataset (ELMo + TrainableLSTM + FocusLayer): F1-score 0.82
Sharifi et al. [54]	MSA	Q2Q	- Similarity measures using embedding and Word Mover's distance - Doc2vec similarity - POS tag overlap - SVM classifier	FastText	NO	NO	Mawdoo3 Q2Q dataset	F1-score is 82.58%
Einea and Elnagar [60]	MSA	Q2Q	- Siamese neural networks with 1D-CNN, BiLSTM, BiGRU - Vector Similarity Layer with Manhattan Distance, Euclidean Distance, and Cosine Distance	A Word2Vec model trained on several datasets	NO	NO	Mawdoo3 Q2Q dataset and SemEval-2016 Task 3 dataset	- Results on NSURL Dataset: Best accuracy 76.9% obtained using 1D-CNN and Euclidean Distance - Results on SemEval Dataset: Accuracy 58.0%
Lichouri et al. [53]	MSA	Q2Q	Siamese neural networks	Not indicated	NO	NO	Mawdoo3 Q2Q dataset	F1-score: 79.89%

Al-Theiabat and Al-Sadi [65]	MSA	Q2Q	Different deep learning models: CNN-based, RNN (bi-directional GRU, Multi-head attention network model, BERT model)	Not indicated	YES	YES	Mawdoo3 Q2Q dataset	The top-performing model is generated by employing an ensemble of pre-trained multilingual BERT models with 95.924% F1-Score
Hammad et al. [52]	MSA	Q2Q	- Supervised machine learning with feature engineering using a set of morphological, semantic, and lexical-based features. - Siamese deep learning recurrent architecture - A Pre-trained deep bidirectional transformer based on the BERT model	BERT	YES	NO	Mawdoo3 Q2Q dataset	Best result (BERT-based model) with 92.99% F1-score
Al Sulaiman et al. [70]	MSA, Egyptian, and Saudi Arabic dialects	Sentence Similarity	- Tuning Arabic BERT - Combining English STS and Arabic BERT models to develop improved models for Arabic STS - Transfer learning and Knowledge distillation techniques	Several models including ArabicBERT [68] and ARBERT [80]	YES	-	SEMEVAL 2017 Multilingual Semantic Textual Similarity dataset and its manual translation to Egyptian and Saudi Arabic	- Evaluation Results for MSA STS: 81% Spearman rank correlation - Evaluation Results for Dialects: 77.5% Spearman rank correlation for the Egyptian dialect and 76% for the Saudi Arabia dialect
Saidi et al. [66]	MSA	Sentence Similarity	Combining BERT, Attention mechanism, and Siamese	BERT. including AraBER, Arabic-BERT, CAMeL-BERT, and the multilingual mBERT	YES	YES	Arabic STS datasets from the SEMEVAL 2017 Multilingual Semantic Textual Similarity	Best Results were obtained on the sub-dataset MSR-Paraphrase dataset with 92.50% Pearson's correlation.

The data was extracted from three different medical forums.

They used the questions extracted from one forum as original questions, and they associated every question with a set of related question/answer pairs collected from the other two forums using a search engine.

The available training dataset comprises 1,531 original questions associated with 45,164 question/answer pairs annotated with “Direct” if the pair contains a direct answer to the original question, “Related” if it provides a partial answer to the original question, and “Irrelevant” if it doesn’t cover any aspects of the original question.

At the document level, to the best of our knowledge, the only available Arabic dataset dedicated especially to document STS is the SemEval-2022 dataset [77] introduced for the news article similarity task.

The data comprises Arabic news article pairs, each annotated with labels ranging from 1 (Very similar) to 4 (Very Dissimilar) regarding seven score categories namely, "Geography," "Entities," "Time," "Narrative," "Overall," "Style," and "Tone". To determine the final label for each article pair, the average of the category scores is calculated. Furthermore, The Open Source Arabic Corpora (OSAC) [78] served as a source corpus for Arabic document similarity studies. The corpus contains 22,429 texts covering various categories (i.e. economics, history, education, health, etc.).

Table 5 describes the key Arabic STS datasets in terms of type, language, size, and the scale used for measuring semantic similarity.

3.4 Analysis of Arabic Semantic Textual Similarity and Future Directions

The objective of this subsection is to compare the results of various approaches when applied to Arabic. However, it is important to note that not all of these approaches have been assessed using identical evaluation methods or datasets. Table 7 resumes the different works previously presented. We highlighted the different approaches and the evaluation settings.

This study examines key works from 2017 onwards that pertain to the Arabic language. It specifically concentrates on research employing modern deep learning-based approaches, aiming to identify significant trends and noteworthy developments that have influenced the field during this period.

Table 6 shows a comparison of results from studies that were evaluated on the same dataset (Mawdoo3 Q2Q dataset) and using identical evaluation settings. We observe that the most effective systems utilize word embeddings in conjunction with advanced deep learning architectures, including ONLSTM, LSTM, and the BERT model.

While research in Arabic STS is expanding, certain limitations and challenges persist, impeding the advancement in this field from aligning with that of English. First, the available datasets for Arabic remain limited compared to the vast availability of English datasets. The inadequacy in both the size and number of Arabic datasets poses a significant challenge and negatively impacts the progress of research in this field in comparison to English STS. In NLP tasks, including STS, the availability of large and diverse datasets is crucial for training robust and effective models.

The Arabic language, with its unique characteristics including variations in dialects, introduces additional challenges. Unfortunately, the available STS datasets for Arabic are scarce when it comes to representing the diverse Arabic dialects as well as domain-specific datasets. Addressing these limitations and expanding the range and diversity of Arabic STS datasets is essential for the advancement of research in this field.

Additionally, despite the availability of datasets containing pairs sourced from native Arabic materials, it's noteworthy that some datasets are translations from English datasets. Translation can introduce additional complexities and potential discrepancies in linguistic nuances, cultural references, and idiomatic expressions between the source and target languages.

Consequently, models trained on translated datasets may encounter challenges in capturing the complexity of Arabic language usage and may exhibit biases or limitations in their performance. Thus, while leveraging translated datasets can expand the scope of available data for Arabic NLP tasks, careful consideration of the potential impacts of translation on dataset quality and model performance is essential for robust and reliable results.

On the other hand, Models trained on MSA or one dialect might not work well with others because of differences in words, grammar, and sentence structure. Ensuring the generalizability of STS models across Arabic dialects requires training data that encompasses a wide range of dialectal variations and fine-tuning or adapting models on dialect-specific data to improve performance. Syntactic differences including variations in word order, verb conjugation, and syntactic structures across Arabic dialects can affect the alignment of text pairs in STS tasks.

Arabic dialects differ in terms of idiomatic expressions, cultural references, and ways of expressing ideas, which may not be captured adequately by models trained in Standard Arabic. To enhance model performance, it's essential to incorporate linguistic features specific to Arabic dialects into model architectures, augment training data with dialectal variations and diverse linguistic phenomena, and develop evaluation benchmarks that take into account the dialectal differences and linguistic nuances of the Arabic language.

4 Conclusion

The paper provides a comprehensive examination of the STS task. It offers insights into various strategies employed to tackle the STS challenge in English while delving deeply into the specific progressions achieved in Arabic STS. Through comparative analysis, the paper unveils the attributes and advancements within Arabic STS, shedding light on valuable insights and future research directions. Despite strides made in Arabic STS, there remains significant work to be done, both in terms of the approaches used and the available datasets.

In comparison to the English language, there is little dataset-related work. Modern approaches based on deep learning, contextualized word embeddings, attention mechanisms, and vectorized text representations like the BERT language model have been minimally explored in the context of the Arabic language. The application of recent advancements, including Large Language Models (LLMs), is still underdeveloped, highlighting a gap compared to English and emphasizing the need for extensive research and development of comprehensive datasets.

References

1. **Severyn, A., Moschitti, A. (2015).** Learning to rank short text pairs with convolutional deep neural networks. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 373–382. DOI: 10.1145/2766 462.276773.
2. **Gonzales, A. R., Mascarell, L., Sennrich, R. (2017).** Improving word sense disambiguation in neural machine translation with sense embeddings. Proceedings of the Second Conference on Machine Translation, pp. 11– 19.
3. **Babar, S., Patil, P. D. (2015).** Improving performance of text summarization. Procedia Computer Science, 46, pp. 354-363. DOI: 10.1016/j.procs.2015.02.031.
4. **Risch, J., Möller, T., Gutsch, J., Pietsch, M. (2021).** Semantic answer similarity for evaluating question answering models. DOI: 10.48550/arXiv.2108.06130.
5. **Vrbanec, T., Meštrović, A. (2017).** The struggle with academic plagiarism: Approaches based on semantic similarity. 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, pp. 870-875. DOI: 10.23919/MIPRO. 2017.7973544.
6. **Arabi, H., Akbari, M. (2022).** Improving plagiarism detection in text documents using hybrid weighted similarity. Expert Systems with Applications, Vol. 207, p. 118034. DOI: 10.1016/j.eswa.2022.118034.
7. **Alian, M., Awajan, A. (2018).** Arabic semantic similarity approaches - review. 2018 International Arab Conference on Information Technology (ACIT), IEEE, pp. 1–6. DOI: 10.11 09/ACIT.2018.8672665.
8. **Alian, M., Awajan, A. (2020).** Semantic similarity for English and Arabic texts: a review. Journal of Information & Knowledge Management, Vol. 19, No. 4, p. 2050033. DOI: 10.1142/S0219649220500331.
9. **Abo-Elghit, A. H., Al-Zoghby, A. M., Hamza, T. T. (2020).** Textual similarity measurement approaches: A survey (1). Egyptian Journal of Language and Engineering, Vol. 7 No. 2, pp. 41–62. DOI: 10.21608/ejle.2020.42018.1012.
10. **Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A. (2012).** Semeval-2012 task 6: A pilot on semantic textual similarity. SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the

shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pp. 385–393.

11. **Seelawi, H., Mustafa, A., Al-Bataineh, H., Farhan, W., Al-Natsheh, H. T. (2019).** NSURL-2019 task 8: Semantic question similarity in Arabic. Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) colocated with ICNLSP 2019-Short Papers, pp. 1–8.
12. **Croft, D., Coupland, S., Shell, J., Brown, S. (2013).** A fast and efficient semantic short text similarity metric. 2013 13th UK Workshop on Computational Intelligence, IEEE, pp. 221–227. DOI: 10.1109/UKCI.2013.6651309.
13. **Alnajran, N., Crockett, K., McLean, D., Latham, A. (2018).** An empirical performance evaluation of semantic-based similarity measures in microblogging social media. 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), IEEE, pp. 126–135. DOI: 10.1109/BDCAT.2018.00023.
14. **Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Wiebe, J. (2016).** Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. SemEval-2016, 10th International Workshop on Semantic Evaluation, pp. 497–511.
15. **Hassanzadeh, H., Nguyen, A., Verspoor, K. (2019).** Quantifying semantic similarity of clinical evidence in the biomedical literature to facilitate related evidence synthesis. Journal of Biomedical Informatics, Vol. 100, p. 103321. DOI: 10.1016/j.jbi.2019.103321.
16. **Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Liu, H. (2018).** Overview of the BioCreative/OHNLP challenge 2018 task 2: clinical semantic textual similarity. Proceedings of the BioCreative/ OHNLP Challenge, Vol. 2018, pp. 1–5.
17. **Mandal, A., Chaki, R., Saha, S., Ghosh, K., Pal, A., Ghosh, S. (2017).** Measuring similarity among legal court case documents. Proceedings of the 10th Annual ACM India Compute Conference, pp. 1–9. DOI: 10.1145/ 3140107.314011.
18. **Sugathadasa, K., Ayesha, B., de-Silva, N., Perera, A. S., Jayawardana, V., Lakmal, D., Perera, M. (2017).** Synergistic union of word2vec and lexicon for domain specific semantic similarity. 2017 IEEE international conference on industrial and information systems, pp. 1–6. DOI: 10.1109/ICIINFS.2017. 8300343.
19. **Sellak, H., Ouhbi, B., Frikh, B. (2015).** Using rule-based classifiers in systematic reviews: a semantic class association rules approach. Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services, pp. 1–5. DOI: 10.1145/2837185.283727.
20. **Deshpande, A., Jimenez, C. E., Chen, H., Murahari, V., Graf, V., Rajpurohit, T., Narasimhan, K. (2023).** C-STS: Conditional semantic textual similarity. arXiv preprint arXiv:2305.15093. DOI: 10.48550/arXiv.23 05.15093.
21. **Miller, G. A. (1995).** WordNet: A lexical database for English. Communications of the ACM, Vol. 38, No. 11, pp. 39–41. DOI: 10.11 45/219717.219748.
22. **Navigli, R., Ponzetto, S. P. (2012).** BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial intelligence, Vol. 193, pp. 217–250. DOI: 10.1016/j.artint.2012. 07.001.
23. **Li, Y., Bandar, Z. A., McLean, D. (2003).** An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, pp. 871–882. DOI: 10.1109/TKDE.2003.1209005.
24. **Jiang, Y., Zhang, X., Tang, Y., Nie, R. (2015).** Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. Information Processing & Management, Vol. 51, No. 3, pp. 215–234. DOI: 10.1016/j.ipm.2015.01.001.
25. **Resnik, P. (1995).** Using information content to evaluate semantic similarity in a taxonomy. ArXiv Preprint cmp-lg/9511007. DOI: 10.485 50/arXiv.cmp-lg/9511007.

26. **Tversky, A. (1977).** Features of similarity. Psychological Review, Vol. 84, No. 4, p. 327. DOI: 10.1037/0033-295X.84.4.327.

27. **Tien, N. H., Le, N. M., Tomohiro, Y., Tatsuya, I. (2019).** Sentence modeling via multiple word embeddings and multi-level comparison for semantic textual similarity. Information Processing & Management, Vol. 56, No. 6, p. 102090. DOI: 10.1016/j.ipm.2019.102090.

28. **Ranasinghe, T., Orăsan, C., Mitkov, R. (2019).** Semantic textual similarity with siamese neural networks. Proceedings of the International Conference on Recent Advances in Natural Language Processing, pp. 1004–1011. DOI: 10.26615/978-954-452-056-4_116.

29. **Neculoiu, P., Versteegh, M., Rotaru, M. (2016).** Learning text similarity with siamese recurrent networks. Proceedings of the 1st Workshop on Representation Learning for NLP, pp. 148–157.

30. **Lv, C., Wang, F., Wang, J., Yao, L., Du, X. (2020).** Siamese multiplicative LSTM for semantic text similarity. Proceedings of the 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence, No. 28, pp. 1–5. DOI: 10.1145/3446132.3446160.

31. **Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013).** Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, pp. 3111–3119.

32. **Pennington, J., Socher, R., Manning, C. D. (2014).** Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543.

33. **Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018).** Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 pp. 2227–2237.

34. **Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019).** BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1. pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

35. **Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A. (2017).** Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364. DOI: 10.48550/arXiv.1705.02364.

36. **Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Kurzweil, R. (2018).** Universal sentence encoder for English. Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations, pp. 169–174. DOI: 10.18653/v1/D18-2029.

37. **Gao, T., Yao, X., Chen, D. (2021).** SimCSE: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821. DOI: 10.48550/arXiv.2104.08821.

38. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017).** Attention is all you need. 31st Conference on Neural Information Processing Systems, Vol. 30.

39. **Liu, Y. (2019).** Roberta: A robustly optimized bert pretraining approach. ArXiv Preprint ArXiv:1907.11692.

40. **Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Liu, P. J. (2020).** Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, Vol. 21, No. 140, pp. 1–67.

41. **Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018).** Improving language understanding by generative pre-training.

42. **Tian, J., Zhou, Z., Lan, M., Wu, Y. (2017).** 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. Proceedings of the 11th international

workshop on semantic evaluation, pp. 191–197. DOI: 10.18653/v1/S17-2028.

43. **Glavaš, G., Franco-Salvador, M., Ponzetto, S. P., Rosso, P. (2018).** A resource-light method for cross-lingual semantic textual similarity. Knowledge-Based Systems, Vol. 143, pp. 1–9. DOI: 10.1016/j.knosys.2017. 11.041

44. **Brychcín, T. (2020).** Linear transformations for cross-lingual semantic textual similarity. Knowledge-Based Systems, Vol. 187, p. 104819. DOI: 10.1016/j.knosys.2019.06.027.

45. **Panchenko, A., Morozova, O. (2012).** A study of hybrid similarity measures for semantic relation extraction. Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, pp. 10–18.

46. **Rychalska, B., Pakulska, K., Chodorowska, K., Walczak, W., Andruszkiewicz, P. (2016).** Samsung poland NLP team at SemEval-2016 Task 1: Necessity for diversity; combining recursive autoencoders, WordNet and ensemble methods to measure semantic similarity. Proceedings of the 10th International Workshop on Semantic Evaluation, pp. 602–608.

47. **Farghaly, A., Shaalan, K. (2009).** Arabic natural language processing: challenges and solutions. ACM Transactions on Asian Language Information Processing, Vol. 8, No. 4, DOI: 10.1145/1644879.1644881.

48. **Habash, N. Y. (2010).** Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies, Vol. 3, No. 1.

49. **Attia, M., (2008).** Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation. Manchester: University of Manchester, Vol. 279.

50. **Wali, W., Gargouri, B., Hamadou, A. B. (2015).** Supervised learning to measure the semantic similarity between Arabic sentences. Computational Collective Intelligence: 7th International Conference, Springer, pp. 158–167. DOI: 10.1007/978-3-319-24069-5_15.

51. **Hammad, M. M., Al-Smadi, M., Baker, Q. B., Al-Asa'd, M., Al-Khdour, N., Younes, M. B., & Khwaileh, E. (2020).** Question to question similarity analysis using morphological, syntactic, semantic, and lexical features. Journal of Universal Computer Science, Vol. 26, No. 6, pp. 671–697.

52. **Hammad, M., Al-Smadi, M., Baker, Q. B., Sa'ad, A. (2021).** Using deep learning models for learning semantic text similarity of Arabic questions. International Journal of Electrical and Computer Engineering, Vol. 11, No. 4, p. 3519. DOI: 10.11591/ijece.v11i4.pp3519-3528.

53. **Lichouri, M., Abbas, M., Benaziz, B., Freihat, A. A. (2019).** ST NSURL 2019 shared task: Semantic question similarity in Arabic. Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages, pp. 80–84.

54. **Sharifi, A., Hassanpoor, H., Maduyieh, N. Z. (2019).** AtyNegar at NSURL-2019 task 8: Semantic question similarity in Arabic. Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) colocated with ICNLSP 2019-Short Papers, pp. 31–36.

55. **Ferrero, J., Schwab, D. (2017).** LIM-LIG at SemEval-2017 task1: Enhancing the semantic similarity for Arabic sentences with vectors weighting. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 134–138. DOI: 10.18653/v1/S17-2017.

56. **Zahran, M. A., Magooda, A., Mahgoub, A. Y., Raafat, H., Rashwan, M., Atyia, A. (2015).** Word representations in vector space and their applications for Arabic. Computational Linguistics and Intelligent Text Processing: 16th International Conference, Springer, pp. 430–443. DOI: 10.1007/978-3-319-18111-0_32.

57. **Mueller, J., Thyagarajan, A. (2016).** Siamese recurrent architectures for learning sentence similarity. Proceedings of the AAAI conference on artificial intelligence Vol. 30, No. 1. DOI: 10. 1609/aaai.v30i1.10350.

58. **Ilina, O., Ziyadinov, V., Klenov, N., Tereshonok, M. (2022).** A survey on symmetrical neural network architectures and

applications. Symmetry, Vol. 14, No. 7. p. 1391. DOI: 10.3390/sym14071391.

59. **Othman, N., Faiz, R., Smaïli, K. (2022).** Learning English and Arabic question similarity with Siamese neural networks in community question answering services. Data & Knowledge Engineering, Vol. 138, p. 101962. DOI: 10.1016/j.datak.2021.101962.

60. **Einea, O., Elnagar, A. (2019).** Predicting semantic textual similarity of Arabic question pairs using deep learning. 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), IEEE, pp. 1–5. DOI: 10.1109/AICCSA47632.2019.9035362.

61. **Othman, N., Faiz, R., Smaïli, K. (2019).** Enhancing question retrieval in community question answering using word embeddings. Procedia Computer Science, Vol. 159, pp 485–494. DOI: 10.1016/j.procs.2019.09.203.

62. **Al-Bataineh, H., Farhan, W., Mustafa, A., Seelawi, H., Al-Natsheh, H. T. (2019).** Deep contextualized pairwise semantic similarity for Arabic language questions. 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, pp. 1586–1591. DOI: 10.1109/ICTAI.2019.00229.

63. **Fadel, A., Tuffaha, I., Al-Ayyoub, M. (2019).** Tha3aroon at NSURL-2019 task 8: Semantic question similarity in Arabic. arXiv preprint arXiv:1912.12514. DOI: 10.48550/arXiv.1912.12514.

64. **Shen, Y., Tan, S., Sordoni, A., Courville, A. (2018).** Ordered neurons: Integrating tree structures into recurrent neural networks. ArXiv Preprint, ArXiv181009536. DOI: 10.48550/arXiv.1810.09536.

65. **Al-Theiabat, H., Al-Sadi, A. (2020).** The inception team at NSURL-2019 task 8: Semantic question similarity in Arabic. ArXiv Preprint, ArXiv200411964. DOI: 10.48550/arXiv.2004.11964.

66. **Saidi, R., Jarray, F., Alsuhaibani, M. (2023).** SiameseBERT: A bert-based siamese network enhanced with a soft attention mechanism for Arabic semantic textual similarity. ICAART No. 3, pp. 146–151. DOI: 10.5220/0011624800003393.

67. **Antoun, W., Baly, F., Hajj, H. (2020).** AraBERT: Transformer-based model for Arabic language understanding. ArXiv Prepr, ArXiv200300104. DOI: 10.48550/arXiv.2003.00104.

68. **Safaya, A., Abdullatif, M., Yuret, D. (2020).** KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. ArXiv Prepr, ArXiv200713184. DOI: 10.48550/arXiv.2007.13184.

69. **Libovickỳ, J., Rosa, R., Fraser, A. (2019).** How language-neutral is multilingual BERT? ArXiv Prepr, ArXiv191103310. DOI: 10.48550/arXiv.1911.03310.

70. **Al-Sulaiman, M., Moussa, A. M., Abdou, S., Elgibreen, H., Faisal, M., Rashwan, M. (2022).** Semantic textual similarity for modern standard and dialectal Arabic using transfer learning. PloS one, Vol. 17, No. 8, p. e0272991. DOI: 10.1371/journal.pone.0272991.

71. **Bowman, S. R., Angeli, G., Potts, C., Manning, C. D. (2015).** A large annotated corpus for learning natural language inference. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 632–642. DOI: 10.18653/v1/D15-1075.

72. **Williams, A., Nangia, N., Bowman, S. (2018).** A broad-coverage challenge corpus for sentence understanding through inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 1112–1122. DOI: 10.18653/v1/N18-1101.

73. **Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L. (2017).** Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. ArXiv Preprint, ArXiv:1708.00055. DOI: 10.48550/arXiv.1708.00055.

74. **Almarsoomi, F. A., O'Shea, J. D., Bandar, Z. A., Crockett, K. A. (2012).** Arabic word semantic similarity. Proceedings of the World Academy of Science, Engineering and Technology. World Academy of Science, Engineering and Technology.

75. **Dahy, B., Farouk, M., Fathy, K. (2021).** ASSD: Arabic Semantic Similarity Dataset. Proceedings of the 2021 9th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC), IEEE, pp. 130–134. DOI: 10.1109/JAC-ECC54461.2021.9691424.

76. **Nakov, P., Márquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, A. A., Glass, J., Randeree, B. (2016).** SemEval-2016 Task 3: Community Question Answering. In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California: Association for Computational Linguistics, pp. 525–545. DOI: 10.18653/v1/S16-1083.

77. **Chen, X., Zeynali, A., Camargo, C. Q., Flöck, F., Gaffney, D., Grabowicz, P. A., Samory, M. (2022).** SemEval-2022 Task 8: Multilingual news article similarity. pp. 1094-1106. DOI: 10.18653/v1/2022.semeval-1.155.

78. **Saad, M., Ashour, W. (2010).** OSAC: Open source Arabic corpora. 6th International Conference on Electrical and Computer Systems. DOI: 10.13140/2.1.4664.9288.

79. **Wieting, J., Kiela, D. (2019).** No training required: Exploring random encoders for sentence classification. arXiv preprint arXiv:1901.10444. DOI: 10.48550/arXiv.1901.10444.

80. **Abdul-Mageed, M., Elmadany, A., Nagoudi, E. M. B. (2020).** ARBERT & MARBERT: Deep bidirectional transformers for Arabic. arXiv preprint arXiv:2101.01785. DOI: 10.48550/arXiv.2101.01785.

Article received on 16/06/2024; accepted on 08/07/2024.
**Corresponding author is Samira Boudaa.*