Received: 5 March 2020 | Revised: 1 July 2020 | Accepted: 11 July 2020

DOI: 10.1002/cpe.5971

WILEY

RESEARCH ARTICLE

A survey on the techniques, applications, and performance of short text semantic similarity

Mengting Han[1] | Xuan Zhang[1,2] | Xin Yuan[1] | Jiahao Jiang[1] | Wei Yun[1] | Chen Gao[1]

[1]School of Software, Yunnan University, Kunming, China

[2]Key Laboratory of Software Engineering of Yunnan Province, Kunming, China

Correspondence
Xuan Zhang, School of Software, Yunnan University, Kunming 650091, Yunnan, China.
Email: zhxuan@ynu.edu.cn

Funding information
the Data Driven Software Engineering Innovative Research Team Funding of Yunnan Province under Grant, Grant/Award Number: 2017HC012; the National Natural Science Foundation of China under Grant, Grant/Award Numbers: 61262025, 61502413, 61862063; the National Social Science Foundation of China under Grant, Grant/Award Number: 18BJL104; the Natural Science Foundation of Key Laboratory of Software Engineering of Yunnan Province under Grant, Grant/Award Number: 2015SE202; the Natural Science Foundation of Yunnan Province under Grant, Grant/Award Number: 2016FB106

Summary

Short text similarity plays an important role in natural language processing (NLP). It has been applied in many fields. Due to the lack of sufficient context in the short text, it is difficult to measure the similarity. The use of semantics similarity to calculate textual similarity has attracted the attention of academia and industry and achieved better results. In this survey, we have conducted a comprehensive and systematic analysis of semantic similarity. We first propose three categories of semantic similarity: corpus-based, knowledge-based, and deep learning (DL)-based. We analyze the pros and cons of representative and novel algorithms in each category. Our analysis also includes the applications of these similarity measurement methods in other areas of NLP. We then evaluate state-of-the-art DL methods on four common datasets, which proved that DL-based can better solve the challenges of the short text similarity, such as sparsity and complexity. Especially, bidirectional encoder representations from transformer model can fully employ scarce information of short texts and semantic information and obtain higher accuracy and F1 value. We finally put forward some future directions.

KEYWORDS

BERT, deep learning, semantic similarity, short text

1 | INTRODUCTION

Short text similarity plays a significant role in natural language processing (NLP) and has been widely used in many language processing fields such as question-answering,[1] text categorization[2,3], paraphrase identification,[4] and information retrieval.[5] With the increasing demand for these applications, text similarity has become a hot spot for NLP. However, the short text is different from common long text such as news and magazines. The content of the short text is too sparse so that the effect of traditional string-based measures is no longer applicable. Therefore, short text similarity measurement requires specific solutions, and research in this field has broad prospects and research value.

In the early stages of the study, string-based similarity metrics[6-8] such as Levenshtein Distance, Cosine, Jaccard, Euclidean distance, and Hash, have been proposed to deal with different kinds of short text similarity and other NLP problems. However, these string-based similarity measures are unable to solve the semantic problems such as polysemous and synonyms. Furthermore, since the biggest feature of short sentences is insufficient context, string-based similarity measures are difficult to calculate the sentence similarity accurately. Therefore, how to make the machine better recognize the meaning expressed by the short text is an important problem of the similarity computation. We have learned that relying on string measures alone is far from accurate. Semantic similarity makes up for the shortcomings of traditional methods and calculates the similarity more accurately by identifying the semantic information of the text. In fact, correct understanding of semantic information can lay a solid theoretical foundation and application conditions for similarity calculation. By identifying the context information, the similarity can be calculated more accurately than traditional string-based measures, because the meaning of the text will be more accurately understood. Therefore, semantic similarity has become a key technology in NLP. At present, there are many applications that use semantic similarity technology and have achieved good results. Such as text classification,[9,10] sentiment analysis,[11-13] information retrieval,[14] social network.[15-17]

With the increasing interest in neural networks, the extraction technology of semantic information has been improved, especially the emergence of deep learning (DL) models.[18-20] In the following, semantic similarity measures are divided into non-DL measure and DL measure. Non-DL measure is divided into two categories: corpus-based measure and knowledge-based measure. In addition, according to the mainstream DL methods, we summarize the DL similarity measures. These DL similarity measures are divided into three categories: general model,[21,22] attention model,[23,24] and hybrid model.[25,26] The main contributions of this article are summarized as follows.

1. The traditional semantic similarity measures and their advantages and disadvantages are summarized. The DL-based semantic similarity measures are analyzed and summarized.
2. Semantic similarity is a widely used technique. The application scenarios of different semantic similarity measures are also presented.
3. Typical and representative methods are used to performed sentence pair similarity experiments. The performance of these models on different datasets is discussed.

The remainder of this article is organized as follows. Section 2 introduces the theory and processing of text similarity. The techniques of non-DL measures and DL measures are reviewed in Section 3. Section 4 describes the applications of semantic similarity. The experiments and analysis are presented in Section 5. Sections 6 concludes our work and proposes future work.

2 | SHORT TEXTS SIMILARITY (STS) THEORY

The definition, data preprocessing, and feature engineering of short text similarity are described in this section. The definition helps to understand the general principle of short text similarity. Data preprocessing and feature engineering are both the essential premise of similarity calculation. Without them, the textual content will not be recognized by the computer or computational results will be affected.

2.1 | Definition

Short text similarity is widely used in various fields. The similarity definition is related to a specific application or a form of knowledge representation. Therefore, there is no uniform definition of text similarity. Lin[27] puts forward a similarity definition of information theory: the greater the difference between the two texts, the smaller the similarity; conversely, the greater the similarity. At the same time, the similarity theorem is deduced based on the hypothesis, as shown in the following formula[27]:

$$\mathrm{sim}(A,B)=\frac{\log\mathrm{P}(\mathrm{common}(A,B))}{\log\mathrm{P}(\mathrm{description}(A,B))} \tag{1}$$

where A and B represent sentence 1 and sentence 2 respectively. common (A, B) is the common information of A and B. The more commonality they share, the more similar they are. description (A, B) is all information describing A and B, which contains both information for A and B.

The similarity result values range from 0 to 1. If the two texts are identical, the sim value is 1. On the contrary, if the two texts are completely different, the sim value is 0.

2.2 | Data preprocessing

Short text similarity is not a direct calculation of two original sentences, it is necessary to convert the source sentences into data that the computer can understand and process. That is, the text needs to undergo a series of text preprocessing steps, including spell checking, tokenization, normalization, and so on. General steps for data preprocessing are presented in Figure 1. Please note that not all preprocessing steps need to be checked and fulfilled. It depends on the specific corpus form.

The tasks of each stage are described as follows:

1. The task of corpus cleaning is to keep valuable data in corpus and delete noise data.
2. The purpose of checking the spelling is to find spelling errors and improve spelling quality, which affects the accuracy of similarity.
3. Tokenization is to break the sentence into words.

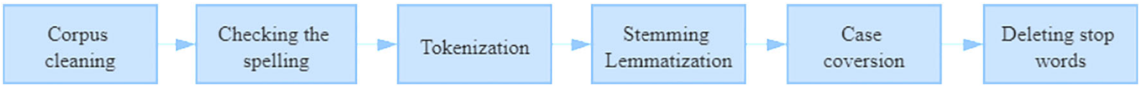
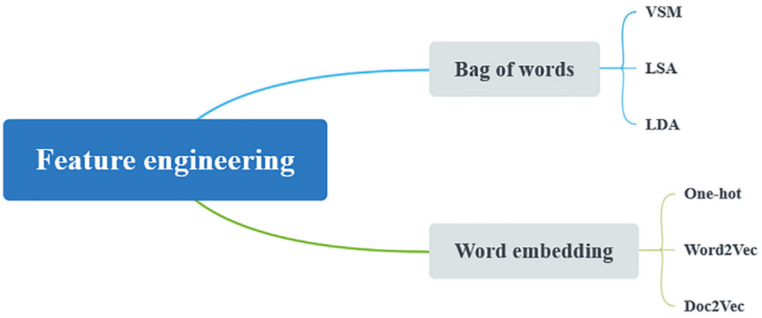


FIGURE 1 Basic steps for data preprocessing

FIGURE 2 Common representation models for feature engineering



4. Stemming and Lemmatization are used to convert words to their basic form.
5. Case conversion is to convert all uppercase letters to lowercase.
6. Deleting stop words is to delete words that do not affect sentence understanding and have no specific meaning.

2.3 | Feature engineering

The text needs to be expressed in a form that the computer can understand. Usually, a vector is a form that a computer can handle. Therefore, words are gathered in the word vector space, which facilitates subsequent similarity calculation. This transformation process is called feature engineering. The commonly used models are bag-of-words (BOW) and word embedding.

BOW counts word frequency on vocabulary in the corpus. The vocabulary in the corpus is used as the feature. The frequency of words is the feature value. The BOW model is usually used in conjunction with the TF-IDF[28] (Term Frequency-Inverse Document Frequency) model. TF indicates the frequency of words in the corpus. IDF is used to improve the ability to distinguish categories. The BOW model is simple to use and has achieved great success in practical applications. Common BOW models are vector space model (VSM),[29] Latent Semantic Analysis (LSA),[30] and Latent Dirichlet Allocation (LDA).[31] But BOW model has limitations, for example, the sparse problem of vectors and word order is not taken into account.

Word embedding converts a word into a fixed-length vector representation for mathematical processing. The simplest method of word embeddings is one-hot encoding. But if the corpus contains thousands of words, vectors of these words may be very long and complicated. Therefore, word embeddings tools, Word2Vec[32] and Doc2Vec,[33] were put forward to solve the problem and became popular. They both can construct word embeddings according to text semantic information. Word2Vec and Doc2Vec models will be introduced in detail in the following specific methods (see Section 3). Common representation models for feature engineering is summarized in Figure 2.

3 | SEMANTIC SIMILARITY MEASURES

In NLP, short text similarity has attracted wide attention, and understanding semantics correctly is a key challenge to understand lexical diversity and ambiguity. This is also the best way to solve the complexity of short texts. Specifically, there are the following challenges to short text similarity.

1. Sparsity: short texts lack sufficient context and rich semantic information. Short sentences contain fewer meaningful words, making it difficult to extract effective feature words. For example, “How are you?” contains too few keywords. Therefore, how to make the machine better recognize the correct meaning of short texts is the first challenge of semantic similarity.
2. Complexity: Irregular and Internet buzzwords are common in short texts, increasing textual noise. Polysemous words and synonyms often appear in text messages. The identical word may have different meanings. Different words may have the same meaning. These complicated characteristics make information identification more difficult.

Therefore, we will focus on evaluating the semantic similarity measures for these two challenges. As mentioned earlier, semantic similarity measures are divided into non-DL measures and DL measures. In Figure 3, we expand and subdivide the classification system based on these measures. We also summarize the detailed information about semantic similarity measures in Table 1.

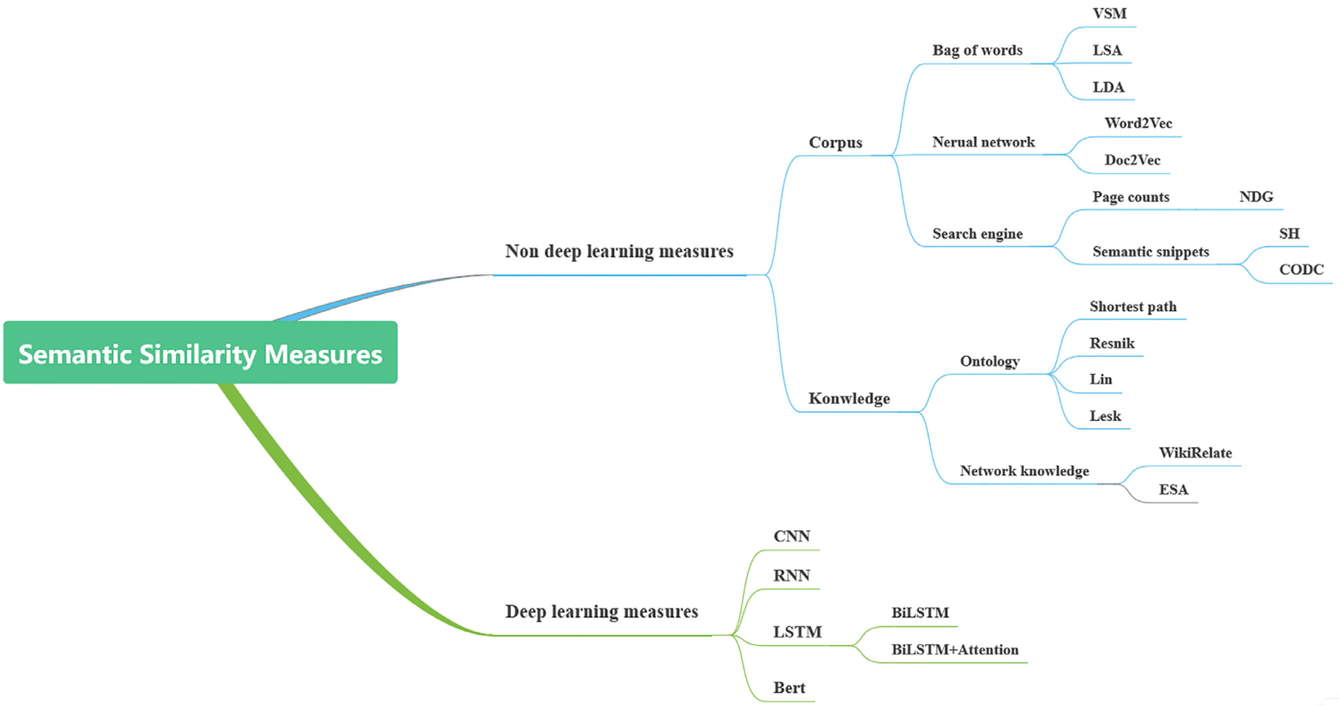


FIGURE 3 Semantic similarity measures

TABLE 1 The detailed information about non-deep learning measures

	Method	Year	Published	Citations
Corpus-based similarity	VSM[29]	1975	Communications of the ACM	8508
	LSA[30]	1990	Journal of the American Society for Information Science	12 000
	LDA[31]	2003	Journal of Machine Learning Research	7471
	Word2Vec[32]	2013	ICLR	1358
	Doc2Vec[33]	2014	International Conference on Machine Learning	1438
	NGD[34]	2007	IEEE Transactions on Knowledge and Data Engineering	2179
	SH[35]	2006	IWWW	900
	CODC[36]	2006	COLING	204
Knowledge-based similarity	Shortest Path[37]	1989	IEEE Transaction on Systems Man and Cybernetics	2275
	Resnik[38]	1995	International Joint Conference on Artificial Intelligence	3742
	Resk[27]	1998	International Conf. on Machine Learning	4593
	Li[39]	2013	IEEE Transaction on knowledge & data engineering	1350
	WikiRelate[40]	2006	Artificial Intelligence	897
	ESA[41]	2007	IJCAI	415
DL-based similarity	CNN[21]	2014	ACL	853
	RNN[42]	2010	International Speech Communication Association	734
	LSTM[43]	1997	Neural Computation	705
	BERT[44]	2019	ARXIV	/

3.1 | Non-DL measures

Non-DL measures include corpus-based measures and knowledge-based measures. Corpus-based measures calculate similarity of two or more texts obtained from the corpus. Knowledge base is constructed by domain experts based on experience. Knowledge-based measures use the information of the semantic network to calculate the similarity between two words.

3.1.1 | Corpus-based measure

Corpus-based similarity measures are to perform similarity calculations on two or more similar texts obtained from the corpus. Three methods are introduced below.

The semantic similarity between sentences can be calculated by the spatial distance between vectors. In 1975, Salton et al[29] proposed a VSM that expresses semantic similarity by measuring similarity in the space. The smaller the angle between two vectors, the more similar the two vectors are. The cosine similarity derivation formula is as follows:

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sum_{i=1}^{n}(x_i)^2 \times \sum_{i=1}^{n}(y_i)^2} \tag{2}$$

where x and y represent two vectors. Considering the number of same feature words reflects two STS, Li et al[45] proposed an improved VSM model. However, the VSM model cannot solve the problem of polysemy and synonyms. For example, human beings can easily distinguish "I like you" and "I don't like you", but the two sentences are highly similar by VSM calculation. In order to improve the shortcoming of VSM, Deerwester et al[30] and Blei et al[31] respectively proposed the LSA model and LDA model. Both models improved the similarity accuracy of the VSM model.

Neural network-based distribution is commonly called word embedding, which uses neural network technology to model the context. The most representative model is Word2Vec[32] proposed by Mikolov et al Word2Vec contains two models: Continues bag-of-words (CBOW) and Skip-Gram. CBOW predicts the target word embedding based on the context word embeddings, while the Skip-Gram model is the opposite. Although Word2Vec can perform semantic analysis on sentence pairs, it ignores the influence of the order of the words in sentences. In order to make up for this deficiency, Mikolov and Quoc proposed a new model Doc2vec,[33] which also contains two models distributed memory and distributed bag-of-words. Different from Word2Vec, Doc2vec uses a document feature vector and word order analysis. Moreover, Doc2vec also accepts sentences of different lengths as training samples.

The development of Internet technology has presented an explosive growth trend. In this era, search engine technology enables users to search for all the content they want. Search engine-based semantic similarity measures are generally divided into two categories: page counts-based measure and semantic snippets-based measure. Cilibrasi and Vitanyi[34] first proposed a normalized Google distance (NGD) semantic similarity calculation. The formula[34] is as follows:

$$\mathrm{NGD}(k_1, k_2) = \frac{\max\left(\log \mathrm{N}_1(k_1), \log \mathrm{N}_2(k_2) - \log \mathrm{N}_1(k_1, k_2)\right)}{\log \mathrm{N}_1 - \min\left(\log \mathrm{N}_1(k_1), \log \mathrm{N}_2(k_2)\right)} \tag{3}$$

where N_1 is the number of pages from the Google search engine. k_1 and k_2 are search terms. N (k_1) and N (k_2) represent the number of pages returned by the search engine for k_1 and k_2 respectively. Given a set of keywords, the similarity is measured by page counts returned by the Google search engine. Keywords with the same or similar meanings tend to be "close" in Google distance, while words with different meanings tend to be "far away."

However, one of the biggest shortcomings of NGD is that the similarity varies with search engines. So some scholars have proposed a semantic snippets-based measure, which calculates the similarity by analyzing the content of the returned web page. Semantic snippets give useful clues to describe the semantic relations that exist between query words, which are returned by search engines alongside the search results.[46] The semantic snippets-based measure has richer semantic information than using search quantity to compute similarity. Sahami and Heilman[35] introduced an algorithm to address the less satisfactory effect of similarity calculation between two short texts of shared term, which was then called SH. Chen et al[36] proposed a typical Co-occurrence Double Check (CODC) algorithm based on semantic snippets. This algorithm relies heavily on the ranking algorithm of the search engine, and it only has high similarity accuracy when calculating words with high relevance, otherwise, the similarity result is 0.

However, there are three major disadvantages of page counts-based measure: the synonyms in the web pages are ignored; the noise in the network data is ignored; the redundancy in the network data is ignored. Therefore, it is not enough to just use the page counts-based similarity measure, and it cannot improve the accuracy of the semantic similarity calculation.

In Table 2, the advantages and limitations of representative corpus-based similarity measures are summarized.

TABLE 2 The advantage and limitation of corpus-based similarity measures

Method	Advantage	Limitation
VSM	Simple and efficient.	It cannot distinguish polysemy and synonyms well.
LSA	It can distinguish polysemy and synonyms.	It ignores the order of words in a sentence.
LDA	It considers semantic association by adding a theme to the short texts.	It ignores the order of words in a sentence.
Word2Vec	It can maximize the use of semantic information.	It ignores the order of words in a sentence.
Doc2Vec	It adds words order analysis and can train sentences of different lengths.	It cannot distinguish polysemy and synonyms well.
NGD	It can recognize rich semantic information.	It ignores noise and redundancy in network data.
CODC	It can obtain better performance for relevant words.	The similarity of less relevant words is 0.

3.1.2 Knowledge-based measure

A knowledge base is a structured knowledge representation constructed by domain experts. Knowledge-based measures are also semantic similarity measures, which use the information of the semantic network to calculate the similarity between two words. Knowledge-based measures are usually divided into two categories: ontology-based measure and network knowledge-based measure.

The typical ontology includes a semantic dictionary and domain ontology. WordNet[47] is a complete and semantic dictionary, which contains vocabulary from multiple corpora and mixes the features and attributes of traditional dictionaries. WordNet not only gives the definition of vocabulary but points out the relationship between the concepts. WordNet is also a multilingual dictionary that can provide semantic knowledge in multiple languages. Shajalal and Aono[48] utilized bilingual word semantics to capture the semantic similarity between sentences, and proposed semantic similarity measures exploiting word embedding and WordNet. Similar to WordNet, BabelNet is a multilingual dictionary knowledge base. It aims to solve the problem of WordNet lacking non-English language data. Hassan et al[49] used BabelNet to solve the limitations of word aligner. When a single word alignment fails, its multi-word synonyms are retrieved from BabelNet. In addition to the semantic dictionary, other domain ontologies are also often used, such as medical gene ontology, aviation domain ontology, and social relationship ontology, and so on. A large number of professional vocabularies cannot be computed the similarity between them due to the lack of vocabulary in the dictionary, Yan et al proposed word2vec model to train word embedding and then use word embedding to depict the semantic similarity between words. The WordNet-based algorithm was used to verify that the method achieved good performance.[50]

At present, ontology-based semantic similarity can be computed by path length measures, information content measures, feature-based measures, and hybrid measures. Path length-based measures quantify the semantic distance between concept nodes by the path length of the concepts in the ontology tree. The larger the path length of the two concepts in the ontology tree, the smaller the similarity. Information content-based measures calculate similarity by measuring the amount of information contained in a concept. The attribute-based measures calculate the attribute similarity between two concepts. It is more suitable for solving the problem of semantic similarity across ontology. The hybrid measures calculate similarity by using factors such as path length and concept information. Table 3 lists the basic principles, representative algorithms, and characteristics of the methods.

Since ontology is not widely used, network knowledge is introduced to calculate semantic similarity. Compared with ontology, the description of network knowledge is more comprehensive. Network knowledge has richer semantic information and the update rate of information content is faster. Strube and Ponzetto[40] presented the first WikiRelate algorithm based on Wikipedia, which is able to compare the similarity of different

TABLE 3 Ontology-based semantic measures

Method	Fundamental	Algorithm	Feature
Path length	It quantifies the semantic distance between concept nodes.	Shortest path,[37] Wu,[51] Li[39]	It added influencing factors such as node depth, density, intensity, and width in the calculation method.
Information content	It computes the amount of information contained in a concept.	Resnik,[38] Lin[27]	The information shared by concepts is quantified as the semantic similarity between them.
Feature	It quantifies common attribute between two concepts.	Tversky[52]	The calculation effect depends on the integrity of the ontology attribute set.
Hybrid measures	Comprehensive calculation based on distance, content and attributes measures.	Li[39]	Setting of weight parameters depends on domain experts.

parts-of-speech of words. The algorithm computes semantic similarity by measuring the path between words related to Wikipedia pages. In order to improve the calculation accuracy and speed of the WikiRelate, Gabrilovich and Markovitch[41] introduced the Explicit Semantic Analysis (ESA) algorithm, which uses three steps to calculate the similarity: Convert words into concept vectors; use TF-IFD method to assign a weight to each element in the vector; calculate the cosine distance between two concept vectors.

Although non-DL measures take semantic information into account compared with string-based measures, there are still some shortcomings:

1. All methods ignore the impact of the order of the words in the sentence and lack sufficient context to determine the exact meanings of the words in a specific context.
2. The disadvantage of search engine-based measures is that noise and redundant information inevitably affect the similarity result because the information on the network is too much and cluttered.
3. The knowledge-based measures make full use of the prior knowledge of experts. These measures can avoid the sparseness of corpus data and the imbalance between different corpora. However, the construction of the knowledge base relies on domain experts, which requires a lot of manpower to maintain and update. In addition, knowledge reasoning and data completion of the knowledge base are difficult to realize.

3.2 | DL measures

In order to solve some challenges of non-DL measure, DL is applied to model sentence pairs. DL technology has achieved good results in the field of image processing and speech recognition and has also brought new progress to NLP. At present, more and more scientific institutions are using DL to handle more complex and abstract natural language understanding tasks.

In fact, several DL similarity measures have been proposed. The most typical and popular models among them are listed below:

1. Convolutional Natural Network (CNN)-based measures[53,54]input the extracted data features to the fully connected layer to obtain a vector representation of question pairs. The question pairs similarity is calculated by the traditional similarity measurement.
2. Recurrent Neural Networks (RNN) model can be viewed as multiple copies of the same neural network, and each neural network module delivers the message to the next one.[55] The model may cause gradient vanishing and gradient exploding. Therefore, Long Short-term Memory (LSTM)[43] was proposed. LSTM-based measures[19,24,56] avoid the gradient problem of RNN, has stronger "memory ability," and make good use of context feature information.
3. Bidirectional Encoder Representations from Transformer (BERT)[44] is a new model that stacks encoders of multiple Transformers[57] together. Transformer is a network structure that is proposed to replace RNN and CNN. It is essentially an attention structure which can directly obtain global information. Unlike RNN, which requires stepwise recursion to obtain global information, neither does CNN, which can only obtain local information. Two steps are used in BERT: pre-training and fine-tuning.

3.2.1 | CNN-based measure

Kalchbrenner et al[21] introduced a dynamic convolutional neural network that uses dynamic k-max pooling to extract pivotal semantic information in sentences. He et al[22] proposed a model based on CNN for model sentences, and the network facilitates subsequent similarity calculations by extracting features at multiple levels of granularity and using multiple types of pooling. To catch full semantic information, Wang et al[58] pay attention to the importance of dissimilar parts between two sentences and use a two-channel CNN to the decomposed similar and dissimilar components.

3.2.2 | RNN-based measure

To overcome the difficulties of traditional neural language models in capturing global semantic information, Mikolov et al[42] proposed a language model based on RNN, which uses hidden states to summarize all the previous contextual information. Considering a text contains many different latent topics, Song et al[59] proposed a novel fractional latent topic-based RNN (FraLT-RNN) model, which largely maintains the overall semantic information of the text. However, one of the biggest shortcomings of RNN is the problem of gradient vanishing and gradient exploding. It is difficult for RNN to train in long texts because of this defect. Thus, LSTM and various variants were proposed. The LSTM model not only overcame shortcomings but also successfully made success in tasks related to NLP.[43] Mueller and Thyagarajan[60] introduced a Siamese Recurrent Architectures to compare the similarity between two sentences with different length. The Siamese architecture uses two shared weighted LSTM to encode the embedding of the pre-processed sentences. Neculoiu et al[61] introduced the bidirectional LSTM (BiLSTM) model which consisting a forward calculation and a backward calculation. This makes it getting information from two directions of input text to better capture of bidirectional semantic information.

TABLE 4 Method, dataset and index value of DL measure

Model	Method	Year	Published	Dataset	ACC	Precision	Recall	F1
CNN	ABCNN[23]	2016	TACL	SICK	86.2			84.7
	Two-channel[58]	2017	arXiv	MSRP	78.4			82.3
	CNN[53]	2018	CCPE	MSRP	74.6			
RNN	Siamese LSTM[60]	2016	AAAI	SICK	84.2			
	AttSiaLSTM[24]	2018	IALP	MSRP		65.68		
	AttSiaBiLSTM[24]	2018	IALP	MSRP		63.19		
Hybrid CNN-RNN	CNN-LSTM[62]	2017	ICDS	PIT		74.8	60.4	72

3.2.3 | Hybrid measurement based on CNN and RNN

CNN and LSTM are the most commonly used semantic synthesis models for text similarity. The hybrid model can capture multiple layers of feature information for short text representation. Huang et al[62] proposed a multiple-granularity neural sentence model, which uses CNN to extract character-level and word-level features and uses LSTM to model sentence-level semantic representations to obtain fine-grained features, semantic representation, and important contextual and grammatical features. Zheng et al[26] introduced a hybrid bidirectional recurrent convolutional neural network, which captures contexts and long text information by BiLSTM. In addition, the model employed the maximum pool layer of CNN, it is determined by context information which word plays a key role in the text. All experiment results prove that the hybrid model not only outperforms traditional machine learning models but also works better than CNN and RNN.

3.2.4 | Measurement based on attention mechanism

In recent years, the attention mechanism has been widely applied to various tasks of NLP based on DL.[59-64] With the in-depth study of the attention mechanism, various attentions have been proposed by researchers. The attention mechanism is usually employed to weight keywords. Yin and Schütze[23] computed attention weights directly on the input representation, the output of convolution, and both directions to evaluate experiment effects. It is found that the effect of attention mechanism which is input into the convolutional layer is the best. Bao et al[24] proposed an Attention Siamese LSTM (AttSiaLSTM) to capture high-level semantic information. The model is proved the effectiveness of the method in three corpora and three language tasks. In 2017, Google heavily used the self-attention mechanism[57] to learn text representation. Self-attention mechanism adds attention and finds the connection inside the sequence. It has been proved to be effective in many areas such as machine reading, text summaries, and image description generation. Cheng et al[65] used LSTM model and self-attention mechanism in machine reading and exceed the effect of existing models.

3.2.5 | BERT-based computation

BERT uses two steps to perform NLP tasks: pre-training and fine-tuning. Pre-training is similar to word embedding. It uses an existing unlabeled corpus to train a language model. Fine-tuning uses pre-trained language models to complete sentence similarity tasks. Zhang et al[66] proposed a new structured language model. The model not only makes use of a simple context but also uses merging structured semantic information, which can provide rich semantics for language representation. Sakata et al[67] use BERT model to calculate the similarity between the user's query and answer. Their method has robust and high-performance retrieval. In addition, many other BERT-based NLP tasks[68,69] have been proposed and have achieved good results.

In Table 4, the DL similarity measure works of literature are enumerated and summarized.

4 | APPLICATIONS

Semantic similarity can be applied in various fields, these applications can be broadly classified as text classification and text clustering (see Section 4.1), sentiment analysis (see Section 4.2), information retrieval (see Section 4.3), social networks (see Section 4.4), academic plagiarism detection (see Section 4.5) and specific domain (see Section 4.6). We summarize each application area in Table 5.

TABLE 5 References and information for each application area

Domain	Year	Published	Method
Text classification	2014	EMNLP[70]	CNN
	2016	ACL[71]	BiLSTM
	2017	King University-Computer and Information Science[9]	LSI
	2019	IEEE Access[26]	BRCAN
Text clustering	2014	Information Sciences[72]	GA
	2019	IEEE Access[73]	WVDD
	2019	Knowledge and Information Systems[74]	FGTM
Sentiment analysis	2016	IJCNN[12]	LDA
	2019	Knowledge-Based Systems[13]	word embeddings
Information retrieval	2009	Expert Systems with Applications[14]	Ontology
	2012	World Congress on Intelligent Control and Automation[5]	Ontology
	2013	Expert Systems with Applications[75]	WordNet
	2015	ICLR[76]	LSTM
	2017	J Intell Inf Syst[77]	LSTM
Academic plagiarism detection	2016	MIPRO[77]	WordNet
	2018	COLING[78]	CNN
Specific Domain	2012	BMC Bioinformatics[79]	Ontology
	2019	BioMed Research International[80]	Ontology
	2019	International Joint Conference on Artificial Intelligence[38]	Resnik

4.1 | Text classification and text clustering

Manning and Schutze[9] defined text classification as a task of identifying the text content and match according to the defined categories. Fawaz[10] used a singular value decomposition method to extract latent semantic indexing-based text features and showed that cosine similarity is a better option to be considered for the Arabic language text classification. At present, many DL models such as CNN and RNN have been applied to text classification, and have achieved excellent results.[70,71,81] However, in order to solve the challenge of multi-class text classification and fine-grained sentiment analysis, Zheng et al[26] proposed a hybrid bidirectional RNN attention-based model to achieve fine-grained text classification task.

Clustering is a technique that compares the similarity of a group of document or text information and classifies the similar document or text information into the same group. Song et al[72] proposed a fuzzy control genetic algorithm, which combined with mixed semantic similarity measures for document clustering. The mixed measures took advantage of thesaurus-based and corpus-based semantic methods to obtain better performance. Afterward, more and more semantic similarity technologies have been applied to a document or text clustering, and have achieved good results.[73,74]

4.2 | Sentiment analysis

Sentiment analysis refers to the process of analyzing, processing, and extracting subjective text with emotional color using NLP and text mining technology.[11] Poria et al[12] proposed a novel framework, Sentic LDA, to turn aspect-based sentiment analysis from syntax to semantics. Their algorithm supervises the clustering process by exploiting the semantic similarity between two words, highly improve clustering. Araque et al[13] proposed a new method utilizing sentiment lexicons, which extract text features by calculating semantic similarity between input words and lexicon words. The method is tested on multiple datasets, and the experimental results show that the performance of sentiment analysis is improved.

4.3 | Information retrieval

Faced with increasingly complicated network information, it is difficult to retrieve and obtain information. How to accurately and quickly obtain the desired resources is a problem of information explosion. In particular, there are polysemy and synonyms in words of natural language, information retrieval has always been a major challenge.[14] Therefore, a lot of research has been proposed to utilize contextual semantic information to improve the accuracy of information retrieval.[4,75-77]

4.4 | Social networks

As a basic method of text related research and application, semantic similarity measures are often used in social network analysis.[15] Micro-blogging services such as Twitter constitute one of the most successful kinds of applications in the current social networks.

Yang[16] used LDA to discover similar topics and find interesting tweets for users. An algorithm called TS-LDA is proposed, which extracts topics from the content by modeling the time trend on twitter. Vicient and Moreno[17] presented a novel topic discovery methodology based on the mapping of semantic hashtags to WordNet terms and their posterior clustering. Besides, Automatic question-answering is an emerging application on the Internet. It is an active research area in NLP, which aims to design a system that can answer questions automatically and improve human social efficiency greatly. Minaee and Liu[82] proposed a DL model for automatic question-answering. The model contains two parts: the vector representations of questions and answers are obtained by using doc2vec; a neural network is trained to find the most similar pair of questions and answers.

4.5 | Academic plagiarism detection

Academic plagiarism is one of the most serious academic phenomena nowadays, which has seriously affected the goals we are pursuing. It is urgent to develop a qualified plagiarism detection tool. Now, there are many plagiarism detections tools and software that have been successfully used, but the plagiarism phenomenon remains unresolved because the types of plagiarism are diverse. Vrbanec and Mestrovi[83] proposed how to apply text semantic similarity to the plagiarism detection task. They classified existing semantic similarity measures and analyzed their possible usage for paraphrasing detection. In addition, the use of semantic similarity in multilingual detection has also made good progress.[84]

4.6 | Specific domain

Biological data are not only increasing in size but also in diversity.[85] Especially with the rapid development of gene sequencing technology, high-dimensional data such as nomenclature, chromosomal localization, and gene products present a wide variety of features. High-dimensional and complex internal structures make the calculation of semantic similarity in biomedicine become difficult. Garla and Brandt[79] analyzed various semantic similarity measures in the biomedical domain, showing knowledge-based measures perform better than distributed-based measures. Ameera et al[80] utilized parallel and distributed processing by splitting data into multiple partitions and applied semantic similarity measures to each partition. The solution consists of three steps: isolating gene ontology, data clustering, and improved performance of similarity calculation. In addition, many ontology-based semantic similarity measures[86-88] and neural network-based measures[89] also have been successfully applied in the biomedicine domain.

5 | EXPERIMENTS

In order to understand the performance of DL models in sentence pair similarity measurement more clearly and intuitively, sentence pairs similarity experiments on four datasets are performed.

5.1 | Datasets

Four datasets are two recent SemEval competition datasets and one question-question dataset. Each sentence pair has a relatedness score on four datasets and the score is 0 or 1. The label is 1 if the sentence pair is relevant, otherwise 0.

Microsoft Research Paraphrase Corpus (MSRP)91 is extracted from thousands of web-based news sources and contains 5801 English sentence pairs. Each sentence in the dataset comes from a different news article, that is, no more than one sentence is extracted from a news article.

Sentences Involving Compositional Knowledge (SICK) is taken from the competition of SemEval 2014—Task 1[90]and consists of 9840 annotated sentence pairs.

Semantic STS[91] Benchmark comprises a selection of the English datasets used in the STS tasks organized in the context of SemEval between 2012 and 2017. It comprises 8628 sentence pairs and includes textual entailment,[92,93] semantic relatedness,[93] and paraphrase detection.[89,94] Table 6 is the breakdown according to genres and train-dev-test splits:

TABLE 6 STS benchmark annotated examples by genres (rows) and by train, dev. test splits (columns)

Genre	Train	Dev	Test	Total
news	3299	500	500	4299
captain	2000	625	525	3250
forum	450	375	254	1079
total	5749	1500	1379	8628

The Quora dataset[95] is composed of more than 400 000 pairs of questions and is the first dataset open by Quora. Quora is a question-answering website where users ask questions and other users respond. Opening the Quora dataset to the world is to better reduce inefficient duplicate problem pages.

5.2 | Evaluation metrics

To evaluate the performance of DL-based semantic similarity measures on four datasets, we use MSE (mean squared error), Accuracy (ACC), and F1-score as our metrics. Metrics are defined as follows:

MSE is an indicator to evaluate the error between the true value and the predicted value. The smaller the MSE, the more accurate the measurement result. The formula is shown below:

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2 \tag{4}$$

ACC estimates the error between the average value of multiple measurements and the true value, high accuracy means that the test results are closer to the true value. It is defined as follows:

$$\text{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} \tag{5}$$

F1-score is a comprehensive evaluation index that balances precision (P) and recall (R), as follows:

$$\mathrm{F1} = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P}+\mathrm{R}} \tag{6}$$

where

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{7}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{8}$$

Various evaluation metrics can be more easily understood through the confusion matrix, confusion matrix is shown in Table 7.

5.3 | Experiments and analysis

The semantic similarity measures are evaluated and analyzed in the following. We perform sentence pair similarity experiments on semantic measures, including non-DL measures and DL measures. MSE, ACC, F1 of each method were evaluated in terms of MSRP, SICK, STS Benchmark, and

TABLE 7 Confusion matrix

	Predicted value	
True value	Positive	Negative
Positive	True Positive (TP)	True Negative (TN)
Negative	False Positive (FP)	False Negative (FN)

Quora datasets. We first compare one-hot, VSM, and Word2Vec, three vector representation methods. Then the cosine similarity is used to calculate the similarity of sentence pairs so as to better compare the effect of different representation methods. Next, one of the typical knowledge-based measures is compared: WordNet. Finally, we perform four different DL experiments.

To analyze the experimental results more accurately, the non-DL methods we use are the simplest and do not incorporate other algorithms. The framework based on the DL models uses the Siamese network. The architecture of Siamese network has been successfully used in many NLP applications, such as vision application[96,97] and acoustic modeling.[98,99] More recently, Siamese architecture has been applied to measure text similarity.[100] Mueller presented SiaLSTM architecture to learn sentence semantic similarity, which has been proved to be superior to the other methods.[60] Siamese Network has two sub-networks with the same structure and sharing weights. Two sub-networks receive two inputs respectively, convert them into a vector, and then calculate the distance between the two vectors by some distance metric. The analysis of the BERT model is also added in the paper, BERT is a model that has received widespread attention recently and has been successfully used in answer-question application.[67,101] The paper compares the LSTM models, BERT model with the traditional non-DL measures.

The experimental results are shown in Figures 4, 5, 6, and 7. It should be noted that non-DL metrics are calculated using the threshold that similarity is 0.5. If the similarity is greater than 0.5 and the label is 1, or the similarity is less than 0.5 and the label is 0, the matching is recorded as correct. 0.5 is the threshold with the highest accuracy.

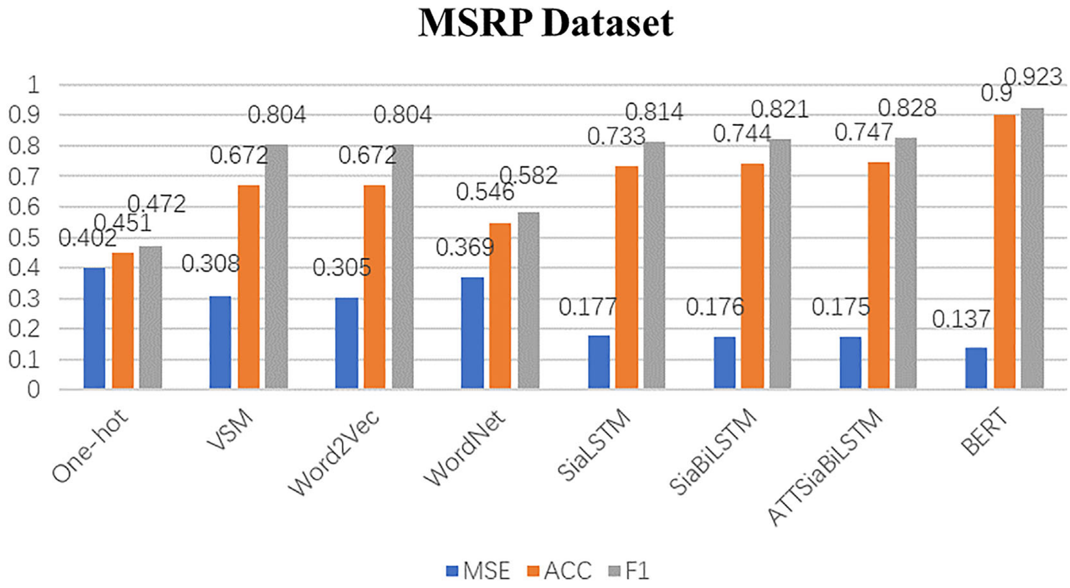


FIGURE 4 MSE, ACC, and F1 of sentence pair similarity on the MSRP dataset

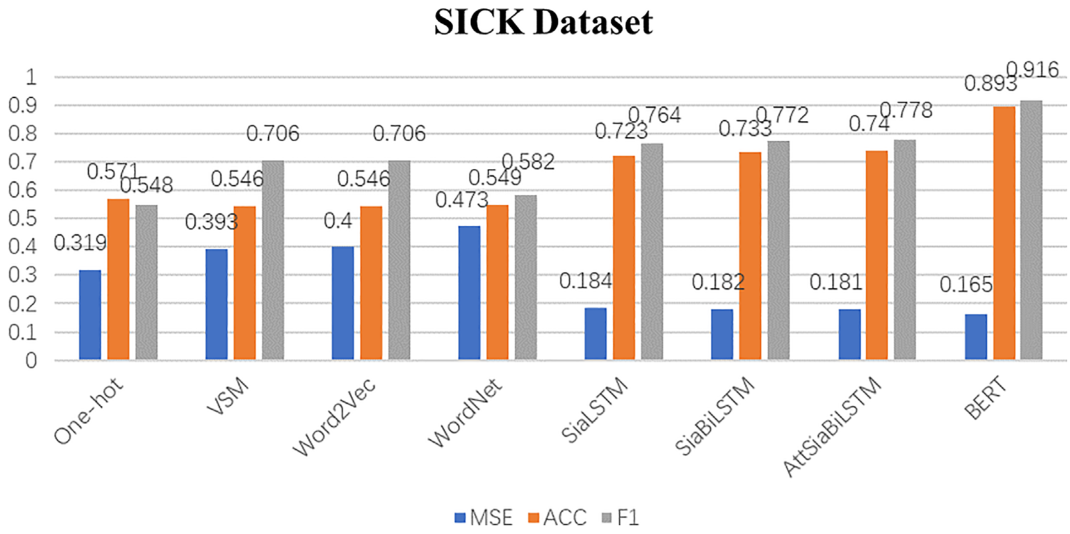


FIGURE 5 MSE, ACC, and F1 of sentence pair similarity on the SICK dataset

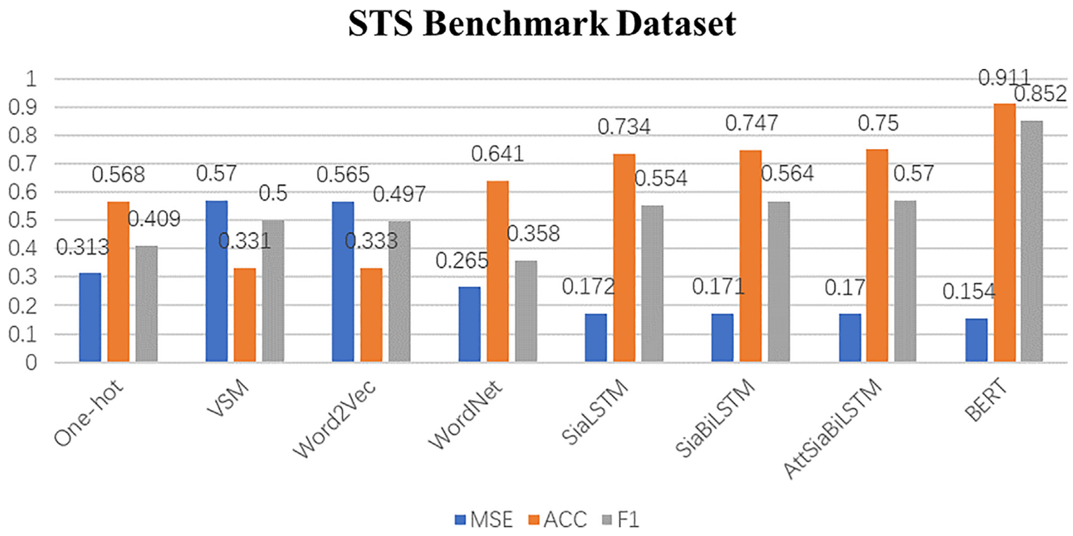


FIGURE 6 MSE, ACC, and F1 of sentence pair similarity on the STS dataset

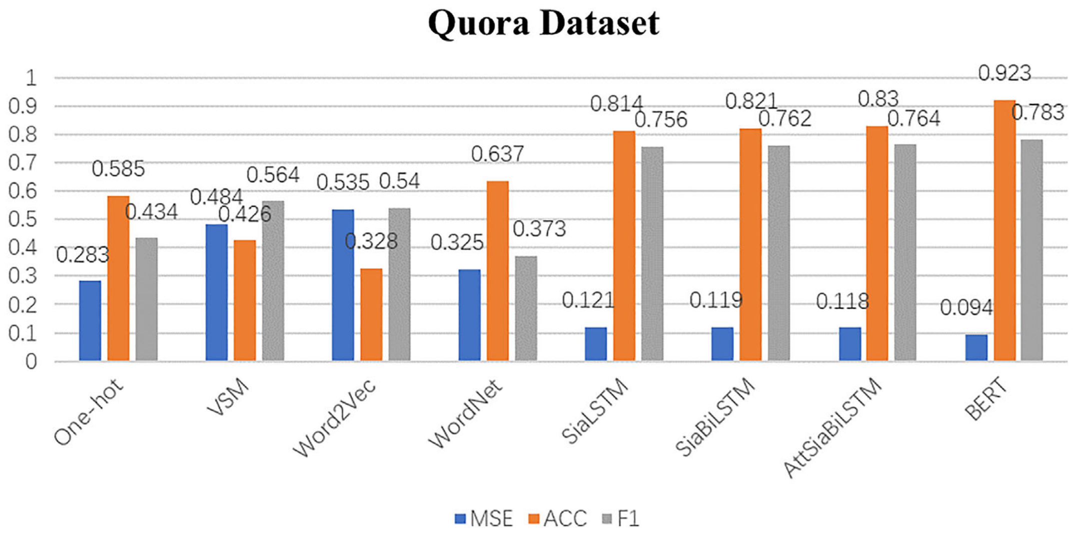


FIGURE 7 MSE, ACC, and F1 of sentence pair similarity on the Quora dataset

From the above four figures, we can see that different models have different results of evaluation metrics on the same dataset. Through the comparison of these different results, we can draw the following conclusions. The evaluation metric results of non-DL measures are obviously much lower than the DL measures. No matter which evaluation metric is used, DL measure is a better method to compute semantic similarity. Some evaluation metric results of VSM and Word2Vec on MSRP and SICK datasets are very close or even the same. This is caused by the threshold, VSM and Word2Vec both have achieved high similarity. Therefore, their evaluation metric results on ACC and F1 are close. It can be seen from the figures that the F1 obtained on the STS dataset is lower than the other datasets. The STS Benchmark includes English datasets in STS missions between 2012 and 2017, and the data come from multiple domains. The possible cause of this result is that the semantic relationship between the two sentences in the STS dataset is usually very subtle. The MSE of BERT model is the lowest, indicating that the measured value trained by the model is closest to the true value. Furthermore, the ACC and F1 of the BERT model are close to 0.9, which is a very good result in semantic similarity.

6 | CONCLUSION AND FUTURE WORK

The paper presents the techniques, applications, and performance of short text semantic similarity measures. First, the basic theory of short text similarity measures is described. Then, many methods are proposed to measure the similarity of short texts in the field. These measures are divided into three categories: corpus-based, knowledge-based, and DL-based measures. The principles of these measures are introduced. Also, the performances are analyzed on four datasets. From the experimental results, DL-based semantic similarity measures have obtained better results than the traditional methods from three evaluation metrics of MSE, ACC, and F1. Especially, BERT model gets the best performance in the short text similarity measures and the performance far exceeds other models.

We believe that there are two promising research directions in the field of semantic similarity: cross-linguistic information, and application in professional fields.

1. Cross-linguistic information: From the current literature on semantic similarity, monolingualism accounts for the majority. However, with the deepening of the degree of economic globalization, exchanges and cooperation between various countries have become more and more frequent. Cross-language semantic similarity may be valuable.
2. Application in professional fields: Most of the current research or competitions of semantic similarity focuses on the daily life of human beings. Most of the datasets are news extracted from Google. But in fact, there are many other areas that apply to text similarity, such as astronomy, geography, medicine, and other specific fields. In the future, we also hope to use existing resources and technologies to study more professional fields.

ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China under Grant No. 61862063, 61502413, 61262025; the National Social Science Foundation of China under Grant No. 18BJL104; the Natural Science Foundation of Yunnan Province under Grant No. 2016FB106; the Natural Science Foundation of Key Laboratory of Software Engineering of Yunnan Province under Grant No. 2015SE202; the Data Driven Software Engineering Innovative Research Team Funding of Yunnan Province under Grant No. 2017HC012.

ORCID

Mengting Han https://orcid.org/0000-0002-3302-5904
Xuan Zhang https://orcid.org/0000-0003-2929-2126

REFERENCES

1. Jimmy L. An exploration of the principles underlying redundancy-based factoid question answering. *ACM Trans Inform Syst*. 2007;25(2):1-55.
2. Wang Y, Wang Z. Text categorization rule extraction based on fuzzy decision tree. Paper presented at: 2005 Conference on Machine Learning and Cybernetics; 2005; Canton, China.
3. Wang Y, Zhu L. Research on improved text classification method based on combined weighted model. *Concurr Comput Pract Exp*. 2020;32(6):1-15.
4. Xu W. *Data-Drive Approaches for Paraphrasing Across Language Variations*. New York, NY: New York University; 2014.
5. Gao Q. Similarity matching algorithm for ontology-based semantic information retrieval model. Paper presented at: 2012 10th World Congress on Intelligent Control and Automation; 2012; Beijing, China.
6. Li M, Chen X, Li X, Ma B, Vitanyi PM. The similarity metric. *IEEE Trans Inform Theory*. 2004;50(12):3250-3264.
7. Monge A, Elkan C. An efficient domain-independent algorithm for detection approximately duplicate database records. Paper presented at: 1997 SIFMOD Workshop on Data Mining and Knowledge Discovery;1997; Tuscan, Italy.
8. Yu Y, Hu Z, Zhang Y. Research on Large Scale Documents Deduplication Technique based on Simhash Algorithm. Paper presented at: 2015 1th International Conference on Information Sciences; 2015; Machinery, Singapore.
9. Manning CD, Schutze H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press; 1999.
10. Al-Anzi SF, Abuzeina D. Toward an enhanced arabic text classification using cosine similarity and latent semantic indexing. *J King Saud Univ Comput Inform Sci*. 2017;29:189-195.
11. Pang B, Lee L. Opinion mining and sentiment analysis. *Found Trends Inf Retrieval*. 2008;2(1–2):130-135.
12. Poria S, Chaturvedi I, Cambria E, Bisio F. Sentic LDA: improving on LDA with semantic similarity for aspect-based sentiment analysis. Paper presented at: 2016 29th International Joint Conference on Neural Networks (IJCNN); 2016; Vancouver, Canada.
13. Araque O, Zhu GG, Iglesias C. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowl Based Syst*. 2019;165:346-359.
14. Song W, Li CH, Park SC. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. *Expert Syst Appl*. 2009;36:9095-9104.
15. Zhang S, Zheng X, Hu CJ. A survey of semantic similarity and its application to social network analysis. Paper presented at: 2015 IEEE International Conference on Big Data (Big Data); 2015; Santa Clara, USA.
16. Yang M, Rim H. Identifying interesting twitter contents using topical analysis. *Expert Syst Appl*. 2014;41:4330-4336.
17. Vicient C, Moreno A. Unsupervised topic discovery in micro-blogging networks. *Expert Syst Appl*. 2015;42:6472-6485.
18. He H, Jimmy L. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. Paper presented at: 2016 NAACL-HLT; 2016; San Diego, CA.
19. Palangi H, Deng L, Shen Y, Gao J, He X, Chen J, Song X. Semantic modelling with long-short-term memory for information retrieval. arXiv preprint arXiv:1412.6629. 2014.
20. Hinton G, Salakhutdinov R. Reducing the dimensionality of data with natural networks. *Science*. 2006;313(5786):504-507.
21. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. Paper presented at: 2014 52th Annual Meeting of the Association for Computational Linguistics; 2014; Baltimore, MD.

22. He H, Gimpel K, Lin J. Multi-perspective sentence similarity modeling with convolutional neural networks. Paper presented at: 2015 Conference on Empirical Methods in Natural Language Processing; 2015; Lisbon, Portugal.
23. Yin WP, Schütze H. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *Trans Assoc Comput Linguist*. 2016;4:259-272.
24. Bao W, Du JH, Yang YY, Zhao XB. Attentive Siamese LSTM network for semantic textual similarity measure. Paper presented at: 2018 14th Conference on Asian Language Processing (IALP); 2018; Bandung, Indonesia.
25. Huang JP, Yao SX, Lyu C, Ji DH. Multi-granularity neural sentence model for measuring short text similarity. *Database Syst Adv Appl*. 2017;10177:439-455.
26. Zheng J, Zheng LM. A hybrid bidirectional recurrent convolutional neural network attention-based model for text classification. *IEEE Access*. 2019;7:106673-106685.
27. Lin D. An information-theoretic definition of similarity. Paper presented at: 1998 International Conference on Machine Learning; 1998; New York, NY.
28. Salton G. Precision weighting: an effective automatic indexing method. *J ACM*. 1976;23(1):76-88.
29. Salton G, Wong A, Yang C. A vector space model for automatic indexing. *Commun ACM*. 1975;18(11):613-620.
30. Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic analysis. *J Am Soc Inf Sci*. 1990;41(6):391-407.
31. Blei D, Ng A, Jordan M. Latent Dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022.
32. Mikolov T, Chen K, Corrado G. Efficient estimation of word representations in vector space. Paper presented at: 2013 Proceedings of Conference on Learning Representations Workshop Track; 2013; Scottsdale, AZ.
33. Mikolov T, Quoc L. Distributed representations of sentences and documents. Paper presented at: 2013 Proceedings of International Conference on Machine Learning; 2014; Beijing, China.
34. Cilibrasi R, Vitanyi P. The Google similarity distance. *IEEE Trans Knowl Data Eng*. 2007;19(3):370-383.
35. Sahami M, Heilman T. A web-based Kernel function for measuring the similarity of short text snippets. Paper presented at: 2006 15th World Wide Web Conference.
36. Chen HH, Lin MS, Wei YC. Novel association measures using web search with double checking. Paper presented at: 2006 44th International Conference on Computational Linguistics; 2006; Sydney, Australia.
37. Rada R, Mi H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. *IEEE Trans Syst Man Cybernet*. 1983;19(1):17-30.
38. Resnik P. Using information content to evaluate semantic similarity. Paper presented at: 1995 14th International Joint Conference on Artificial Intelligence.
39. Li Y, Bandar ZA, Mclean D. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans Knowl Data Eng*. 2003;15(4):871-882.
40. Strube S, Ponzetto SP. Wikirelate! Computing semantic relatedness using Wikipedia. Paper presented at: 2006 21th Conference on Artificial Intelligence; 2006; Menlo, CA.
41. Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. Paper presented at: IJCAI; 2007.
42. Mikolov T, Karafiát M, Burget L, Cernock H, Khudanpur S. Recurrent neural network based language model. Paper presented at: 2010 11th Annual Conference of the International Speech Communication Association; 2010; Makuhari, Japan.
43. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780.
44. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding; 2019. arXiv:1810.04805v2 [cs.CL].
45. Li L, Zhu A, Su T. An improved text similarity calculation algorithm based on VSM. *Adv Res Autom Commun Architecton Mater*. 2011;225-226(1–2):1105-1108.
46. Kavitha A. An integrated approach for measuring semantic similarity between words and sentences using web search engine. *Int Arab J Inform Technol*. 2015;12(6):589-596.
47. Miller GA. Wordnet:a lexical database for English. *Commun ACM*. 1995;38(11):39-41.
48. Shajalal M, Aono M. Semantic textual similarity between sentences using bilingual word semantics. *Prog Artif Intell*. 2019;8:263-272.
49. Hassan B, Abdelrahman S, Bahgat R, Farag I. UESTS: an unsupervised ensemble semantic textual similarity method. *IEEE Access*. 2019;7:85462-85482.
50. Yan F, Fan Q, Lu M. Improving semantic similarity retrieval with word embeddings. *Comput Pract Exp*. 2018;30(23):1-6.
51. Wu Z, Palmer M. Verb semantic and lexical selection. Paper presented at: 1994 32th Annual Meeting of the Associations for Computational Linguistics.
52. Tversky A. Features of similarity. *Psychol Rev*. 1977;84(4):327-352.
53. Severyn A, Moschitti A. Learning to rank short text pairs with convolutional deep neural networks. Paper presented at: 2015 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR); 2015; Santiago, Chile.
54. Yao H, Liu H, Zhang P. A novel sentence similarity model with word embedding based on convolutional neural network. *Concurr Comput Pract Exp*. 2018;30(23):1-12.
55. Elman JL. Finding structure in time. *Cogn Sci*. 1990;14:179-211.
56. Kamineni A, Yenala H, Shrivastava MM, Chinnakotla M. Siamese LSTM with convolutional similarity for similar question retrieval. Paper presented at: 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP); 2018; Pattaya, Thailand.
57. Vaswani A, Shazzer N, Parmar N, Uszkoreit J, Jones LN, Gomez A, Kaiser L. Attention is all you need; 2017. arXiv:1706.03762v5[cs.CL].
58. Wang Z, Mi H, Ittycheriah A. Sentence similarity learning by lexical decomposition and composition; 2017. arXiv:1602.07019v2.
59. Song Y, Hu XW, He L. Using fractional latent topic to enhance recurrent neural network in text similarity modeling. Paper presented at: 2019 Proceedings of the 18th International Conference on Database Systems for Advanced Applications; 2019; Chiang Mai, Thailand.
60. Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity. Paper presented at: 2016 3th AAAI Conference on Artificial Intelligence (AAAI-16); 2016; Phoenix, AZ.
61. Neculoiuv P, Versteegh M, Rotaru M. Learning text similarity with Siamese recurrent networks. Paper presented at: 2016 1th Workshop on Representation Learning for NLP; 2016; Berlin, Germany.
62. Huang JP, Yao SX, Lyu C, Ji DH. Multi-granularity neural sentence model for measuring short text similarity. Paper presented at: proceedings of the 22th International Conference on Database Systems for Advanced Applications; 2017; Suzhou, China.
63. Chorowski J, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech recognition. Paper presented at: 2015 28th Proceedings of International Conference on Neural Information Processing Systems; 2015; Cambridge, England.

64. Xu K, Ba J, Kiros R, Cho K, Courville AC, Salakhutdinov R, Zemel RS, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention; 2015. CoRR vol. abs/1502.03044.
65. Cheng JP, Dong L, Lapata M. *Long Short-Term Memory-Networks for Machine Reading*. Austin, TX: EMNLP; 2016.
66. Zhang ZS, Wu YW, Zhao H, Li ZC. Semantics aware BERT for Language Understanding; 2019. arXiv:1909.02209v2 [cs.CL].
67. Sakata W, Shibata T, Tanaka R, Kurohashi S. FAQ retrieval using query-question similarity and BERT-based query-answer relevance. Paper presented at: 2019 42th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2019; Paris, France.
68. Jin D, Jin ZJ, Zhou JT, Szolovits P. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment; 2019. arXiv:1907.11932v3 [cs.CL].
69. Wang QC, Liu PY, Zhu ZF, Yin HX, Zhang QY, Zhang LD. A text abstraction summary model based on BERT word embedding and reinforcement learning. *Appl Sci*. 2019;9:1-19.
70. Kim Y. Convolutional neural networks for sentence classification. Paper presented at: 2014 Conference on Empirical Methods in Natural Language Processing EMNLP; 2014; Doha, Katar.
71. Hassan A, Mahmood A. Efficient deep learning model for text classification based on recurrent and convolutional layers. Paper presented at: 2014 16th IEEE Int Conf Mach Learn Appl (ICMLA); 2014; Cancun, Mexico.
72. Song W, Liang JZ, Park SC. Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering. *Inform Sci*. 2014;273:156-170.
73. Zhou S, Xu XX, Liu YL, Chang RF, Xiao YY. Text similarity measurement of semantic cognition based on word vector distance decentralization with clustering analysis. *IEEE Access*. 2019;7:107247-107258.
74. Soares V, Ricardo J, Nourashrafeddin S, Milios E, Naldi MC. Combining semantic and term frequency similarities for text clustering. *Knowl Inform Syst*. 2019;61:1485-1516.
75. Mohammed Nazim U, Trong Hai D, Ngoc Thanh N, Xin Min Q, Jo GS. Semantic similarity measures for enhancing information retrieval in folksonomies. *Expert Syst Appl*. 2013;40:1645-1653.
76. Palangi H, Deng L, Shen Y, Gao J, He X, Chen J, Song X, Ward R. Semantic modeling with long-short-term memory for information retrieval. Paper presented at: under review as a workshop contribution at ICLR; 2015; San Diego, CA.
77. Samir A, Adrian T, Djamel AZ. Sentence similarity based on semantic kernels for intelligent text retrieval. *J Intell Inf Syst*. 2017;48:675-689.
78. Mahmoud A, Zrigui A, Zrigui M. A text semantic similarity approach for arabic paraphrase detection. Paper presented at: 2018 18th International Conference on Computational Linguistics; 2018; Santa Fe, NM.
79. Garla V, Brandt C. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC Bioinform*. 2012;13:1-13.
80. Ameera MA, Hend SA, Abdulmalik SA. Handling big data scalability in biological domain using parallel and distributed processing: a case of three biological semantic similarity measures. *Biomed Res Int*. 2019;2019:1-20.
81. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B. Attention-based bidirectional long short-term memory networks for relation classification. Paper presented at: 2016 54th Meeting of the Association for Computational Linguistics; 2016; Berlin, Germany.
82. Minaee S, Liu Z. Automatic question-answering using a deep similarity neural network. Paper presented at 2017 5th IEEE Global Conference on Signal and Information Processing (GlobalSIP); 2017; Montreal, Canada.
83. Vrbanec T, Mestrovi A. The struggle with academic plagiarism: approaches based on semantic similarity. Paper presented at: 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO); 2017; Opatija, Croatia.
84. Ezzikouri H, Oukessou M, Erritali M. Semantic similarity/relatedness for cross language plagiarism detection. Paper presented at: 2016 13th International Conference Computer Graphics, Imaging and Visualization; 2016; Porto, Portugal.
85. Dolinski K, Troyanskaya OG. Implications of big data for cell biology. *Mol Biol Cell*. 2015;26(14):2575-2578.
86. Zhang SM, Chen JW, Wang BY. The research of semantic similarity algorithm consideration of multi-factor ontology-based in access control. Paper presented at: international conference on computer application and system modeling; 2010; Taiyuan, China
87. Harispe S, Sánchez D, Ranwez S, Janaqi J, Montmai M. A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain. *J Biomed Inform*. 2014;48:38-53.
88. Ferreira JD, Couto FM. Multi-domain semantic similarity in biomedical research. *BMC Bioinform*. 2019;20(10):23-31.
89. Blagec K, Xu H, Agibetov A, Samwald M. Neural sentence embedding models for semantic similarity estimation in the biomedical domain. *BMC Bioinform*. 2019;20:1-10.
90. Marelli M, Bentivogli L, Baroni M, Bernardi R, Menini S, Zamparelli R. SemEval-2014 task 1: evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. Paper presented at: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval), Dublin, Ireland; 2014.
91. Daniel C, Mona D, Agirre A. SemEval-2017 Task 1: semantic textual similarity multilingual and crosslingual focused evaluation. Paper presented at: 2017 Meeting of the Association for Computational Linguistics; 2017; Vancouver, Canada.
92. Bowman SR, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. Paper presented at: International Conference on Empirical Methods on Natural Language Processing (EMNLP); 2015; Lisbon, Portugal.
93. Bentivogli L, Bernardi R, Marelli M, Menini S, Baroni M, Zamparelli R. SICK through the SemEval glasses. Lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Lang Resour Eval*. 2016;50(1):95-124.
94. Xu W, Callison-Burch C, Dolan B. SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). Paper presented at: 2015 9th International Workshop Semantic Evaluation (SemEval); 2015; Denver, CO.
95. Lakshay S, Laura G, Nikita N, Utku E. Natural language understanding with the quora question pairs dataset; 2019. CoRR abs/1907.01041.
96. Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. Paper presented at: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); 2005.
97. Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. Paper presented at: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06); 2006; MA.
98. Thiollière R, Dunbar E, Synnaeve G, Versteegh M, Dupoux E. A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. Paper presented at: INTERSPEECH; 2015.
99. Gabriel S, Emmanuel D. A temporal coherence loss function for learning unsupervised acoustic embeddings. *Procedia Comput Sci*. 2016;81:95-100.

100. Neculoiu P, Versteegh M, Rotaru M. Similarity with siamese recurrent networks. Paper presented at: 2016 1th Workshop on Representation Learning for NLP; 2016; Berlin, Germany.

101. Chen Q, Liu Y, Qiu MH, Bruce CW, Zhang YF, Lyyer M. BERT with history answer embedding for conversational question answering. Paper presented at: 2019 International ACM SIGIR Conference on Research and Development in Information Retrieval; 2019; Paris, France.

How to cite this article: Han M, Zhang X, Yuan X, Jiang J, Yun W, Gao C. A survey on the techniques, applications, and performance of short text semantic similarity. *Concurrency Computat Pract Exper*. 2021;33:e5971. https://doi.org/10.1002/cpe.5971