Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 1635708, 9 pages
https://doi.org/10.1155/2021/1635708

Research Article

Sentence Similarity Calculation Based on Probabilistic Tolerance Rough Sets

Ruiteng Yan,[1] Dong Qiu ,[1,2] and Haihuan Jiang[1]

[1]*School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Nanan, Chongqing 400065, China*
[2]*College of Science, Chongqing University of Posts and Telecommunications, Nanan, Chongqing 400065, China*

Correspondence should be addressed to Dong Qiu; dongqiumath@163.com

Received 15 August 2020; Revised 1 December 2020; Accepted 15 January 2021; Published 28 January 2021

Academic Editor: Jun Shen

Sentence similarity calculation is one of the important foundations of natural language processing. The existing sentence similarity calculation measurements are based on either shallow semantics with the limitation of inadequately capturing latent semantics information or deep learning algorithms with the limitation of supervision. In this paper, we improve the traditional tolerance rough set model, with the advantages of lower time complexity and becoming incremental compared to the traditional one. And then we propose a sentence similarity computation model from the perspective of uncertainty of text data based on the probabilistic tolerance rough set model. It has the ability of mining latent semantics information and is unsupervised. Experiments on SICK2014 task and STSbenchmark dataset to calculate sentence similarity identify a significant and efficient performance of our model.

1. Introduction

With the rapid development of information technique, innumerable text data are continuously growing. Unlike digital data, the processing of text data is more complex and difficult. Sentence similarity aims at calculating the degree of resemblance or distance between two sentences. It plays an important role in the application of natural language processing (NLP), like text summarization [1, 2], machine translation [3], question answering systems [4], and information retrieval [5]. These applications are based on sentence similarity to a certain extent, whose development makes the research of sentence similarity become urgent.

Text data are characterized by uncertainty, inaccuracy, and incompleteness. Existing sentence similarity computation methods are almost all based on the relation among words and words in the sentences or based on the deep learning algorithms. Methods based on the relation among the words and words such as word cooccurrence mainly consider sentence semantics from the shallow level and cannot capture the latent semantics information behind the sentences. Methods based on the deep learning algorithms such as convolutional neural network (CNN) can capture deep semantics information, but most of them are with high time complexity and supervision. In addition, both of the classes of methods cannot commendably process the uncertainty and imprecision of text sentences. In this paper, we start with the uncertainty and imprecision of text data. We improve the tolerance rough set model [6] by Ho et al.and present a sentence similarity computation model based on the probabilistic tolerance rough set model. Our model can not only process the uncertainty and imprecision of text data, but also overcome the shortcomings mentioned before.

This paper is organized as follows. Some related works on sentence similarity measures are reviewed in Section 2. Section 3 presents our proposed probabilistic tolerance rough sets-based model for sentence similarity computation in detail. Section 4 demonstrates the experimental results and discusses on sentence similarity tasks. In Section 5, some conclusions are made.

2. Related Work

The main work is to improve the traditional tolerance rough set model, and then establish a sentence similarity computation model based on the probabilistic tolerance rough set model. In this section, we discuss some related works about sentence similarity calculation methods and tolerance rough set models in NLP.

2.1. Sentence Similarity Calculation. Traditional works about sentence similarity are generally categorized into two classes, methods based on shallow semantics and methods based on deep learning algorithms. The idea of shallow semantics methods is to calculate the similarity between words. Methods based on words' cooccurrence and on corpus are two representatives. Methods based on words' cooccurrence are mentioned in [7–9]. Han et al. used the Bag-of-Words (BoW) technique [8], and Jones et al. [7] applied the term frequency inverse document frequency (TF-IDF) technique to represent sentences, and then the cosine distance or Euclidean distance was utilized to calculate the similarity between sentences. A keyword-based approach was proposed [9], which calculates the keywords' ranking score extracted in the sentences. Methods based on corpus such as WordNet, HowNet are mentioned in [10, 11]. In [12], Prasad et al. combined common words and semantic features for measuring sentence similarity. They extracted both syntactic features by searching for common words between sentences and semantic features by utilizing information content of sentences. Methods based on shallow semantics can only obtain the literal meaning of sentences, and fail to capture high-level semantics information behind sentences.

Nowadays, neural network and deep learning have been widely used in NLP and have made great achievements. By training sentences with deep learning algorithms, deep semantics information can be captured in the computation of sentence similarity. In [13], a CNN-based parallel semantic matching model was established; two parallel CNNs were built to train two sentences, respectively. Then, the two CNNs were cascaded into one multilayer neural network for matching the similarity of sentences. An elaborate convolutional network (ConvNet) variant was presented [14], which inferred sentence similarity by integrating differences of convolutions at different scales. For the problem of variable length sentences and complex sentences, Mueller et al. proposed a Siamese Network on the basis of the long short-term memory (LSTM) model [15]. Methods mentioned before mainly concentrate on the similar information of two sentences; on the bias, methods concentrated on the dissimilar information of two sentences were proposed. Wang et al. developed a sentence similarity learning model by decomposing and composing lexical semantics which considered both the similar information and dissimilar information between sentences [16]. In [17], a context-aligned recurrent neural network (CA-RNN) model was put forward. In this model, the contextual information of the aligned words was integrated in the neural network. Liu et al. incorporated the shallow semantics and deep information to evaluate the sentence similarity [18]. The shallow part is represented by the lexical similarity based on keywords and sentence lengths; and the deep part is modeled by a parallel CNN which extracts both the whole sentence and their context as the features. However, most of the sentence similarity learning algorithms based on neural network and deep learning are supervised, which need to train the data set first. Jacob et al. [19] proposed an unsupervised Bidirectional Encoder Representations from Transformers (BERT) model, which has reached excellent results for language representation.

It is undeniable that text data possess uncertainty, imprecision, and incompleteness. However, methods mentioned above do not measure the similarity between sentences from the perspective of uncertainty and imprecision. Fuzzy set theory and rough set theory are created to process such uncertainty and imprecision. A fuzzy set and rough sets-based approach was developed for measuring cross-lingual semantic similarity [20]. In [1], Chatterjee et al. proposed a fuzzy rough sets-based model. Sentence similarity was computed according to the upper approximation and lower approximation of two sentences.

We improve the traditional tolerance rough set model and propose a sentence similarity computation model based on the probabilistic tolerance rough set model. With the model from the point of the uncertainty of text data to process text data, it can not only solve the problem of inability to obtain high-level semantics information on methods based on shallow semantics, but also overcome the drawback of supervision on methods based on deep learning algorithms, with the advantages of capturing more latent semantics information and nonsupervision.

2.2. Tolerance Rough Sets in NLP. Rough set theory was proposed by the Polish scholar Pawlak for handling uncertainty, imprecision, and fuzziness in 1982 [21]. It has been effectively applied in the field of machine learning, data mining, and NLP [22–24]. Rough sets partition a set X by using an equivalence relation. Whether one certain object belongs to a set X or not is represented by a pair of concepts called lower approximation space and upper approximation space. A possible part is the upper approximation except the lower approximation, called the boundary region. Researchers generalized rough set theory to some expanded models according to different requirements, including probabilistic rough set model [25], decision rough set model [26], and tolerance rough set model [27]. An equivalence relation contains three properties of reflexivity, symmetry, and transitivity, in which the limitation of transitivity leads to the inapplicability in some cases. The tolerance relation was introduced to replace the equivalence relation by Skowron et al. since some applications cannot achieve the condition of transitivity, and the corresponding model was tolerance rough set model [27].

With the tolerance rough set model applied in NLP, a search result clustering method was put forward [28], in which the tolerance relation was defined as the number of word cooccurrences in documents. In [29], a tolerance

rough sets-based semantic clustering algorithm is introduced by Meng et al. for web search results, extending the original text semantics and processing the limitation on the sparsity of data. A nonhierarchical document clustering algorithm was established by Ho et al. [6] for information retrieval based on a tolerance rough set model, which can capture more potential semantics information. Patra and Nandi developed a single-link clustering algorithm on the basis of tolerance rough set model to obtain a better clustering result [30]. In this paper, we adopt the tolerance rough set model via expressing each sentence as a pair of upper approximation and lower approximation to separately compute the upper approximation similarity and lower approximation similarity.

3. Proposed Method

In this section, we firstly describe the traditional tolerance rough set model briefly. Then, we introduce the probabilistic tolerance rough sets-based sentence similarity calculation model detailedly.

3.1. Tolerance Rough Set Theory. A tolerance space was defined as a quadruple $\mathbf{R} = (U, I, \nu, P)$ [6], where $U = \{x_1, x_2, \ldots, x_n\}$ is the universe of all the objects, $I(x)$ is an uncertainty function, I: $U \longrightarrow 2^U$, a set of tolerance classes, ν: $2^U \times 2^U \longrightarrow [0,1]$ is a vague inclusion, and P: $I(U) \longrightarrow \{0,1\}$ is a structural function. The uncertainty function I: $U \longrightarrow 2^U$ is defined as a tolerance class. If an object shares similar information with x, it is an element of $I(x)$. Any function satisfying reflexivity and symmetry can be defined as an uncertainty function $I(x)$, that is, for arbitrary $x, y \in U, x \in I(x)$ iff $x \in I(y)$. The vague inclusion ν is monotonous, i.e., for any $X, Y, Z \subseteq U$ and $Y \subseteq Z$, $\nu(X,Y) \leq \nu(X,Z)$. It measures the degree of inclusion of sets, whether a set X contains the tolerance class $I(x)$ of an object $x \in U$. The structural function P is defined as two classes—structural subsets $(P(I(x)) = 1)$ and nonstructural subsets $(P(I(x)) = 0)$—which are on functions of $I(x)$ for each $x \in U$ [6]. The upper approximation $\mathbf{U}(\mathbf{R}, X)$ and lower approximation $\mathbf{L}(\mathbf{R}, X)$ of any $X \subseteq U$ are defined as

$$\begin{aligned} \mathbf{U}(\mathbf{R}, X) &= \{x \in U | P(I(x)) = 1 \,\&\, \nu(I(x), X) > 0\}, \\ \mathbf{L}(\mathbf{R}, X) &= \{x \in U | P(I(x)) = 1 \,\&\, \nu(I(x), X) = 1\}. \end{aligned} \tag{1}$$

If the upper approximation and lower approximation are with parameters α and β, which are denoted as

$$\begin{aligned} \mathbf{U}(\mathbf{R}, X) &= \{x \in U | P(I(x)) = 1 \,\&\, \nu(I(x), X) > \alpha\}, \\ \mathbf{L}(\mathbf{R}, X) &= \{x \in U | P(I(x)) = 1 \,\&\, \nu(I(x), X) \geq \beta\}, \end{aligned} \tag{2}$$

where $\alpha \in [0,1)$, $\beta \in (0,1]$, $\alpha \leq \beta$, then it is called as the probabilistic tolerance rough set model [25].

3.2. Probabilistic Tolerance Rough Sets-Based Sentence Similarity Model. Firstly, we introduce the definition of the quadruple of tolerance rough sets in our model. Suppose that $W = \{w_1, w_2, \ldots, w_N\}$ is the set of all the words in the corpus, where N is the vocabulary size. Then, we define the universe as $U = W$. The determinations of tolerance relation and tolerance classes are the essential steps for formulating a tolerance rough set model. In the tolerance rough set model proposed by Ho et al. [6], the cooccurrence of terms in all the documents in the corpus was applied to construct the tolerance relation. However, it suffers from two disadvantages: (1) whenever the whole corpus gets some changes, even increasing or decreasing by only one document, all the procedures need to be recalculated; (2) the time complexity is relatively high. Hence, we choose the word similarity between words as the tolerance relation. Generally, the semantics similarity between two words is defined as the cosine similarity between the word vectors of the two words [31]. When the corpus increases or decreases one document, the number of cooccurrences of all the words will change and recalculation is needed, but the word similarity between words does not need to change. It provides the model employing the new tolerance relation to be incremental. According to the algorithm flow of the tolerance rough set model, the time complexity decreases from $O(N^3)$ to $O(N^2)$.

For a positive threshold θ, $0 \leq \theta \leq 1$, the uncertainty function I_θ of w_i is defined as follows:

$$I_\theta(w_i) = \{w_i\} \cup \{w_j | \mathrm{sim}(w_i, w_j) \geq \theta\}, \tag{3}$$

where $\mathrm{sim}(w_i, w_j)$ denotes the cosine similarity degree between the word w_i and w_j.

$$\mathrm{sim}(w_i, w_j) = \cos(w_i, w_j) = \frac{\mathbf{w_i} \cdot \mathbf{w_j}}{\|\mathbf{w_i}\| \|\mathbf{w_i}\|}, \tag{4}$$

where $\mathbf{w_i}$ and $\mathbf{w_j}$ denote the word vectors of w_i and w_j, respectively. It is evident that the uncertainty function I_θ satisfies the condition of reflexive and symmetric.

Here, we give a counterexample to illustrate when I_θ does not satisfy the property of transitivity. Using the trained word2vec embeddings by Google [32], we can obtain similarity('beautiful', 'nice') = 0.5341, similarity('nice', 'pretty') = 0.5106, and similarity('beautiful', 'pretty') = 0.3299. Let the cosine similarity degree threshold $\theta = 0.5$; it is obvious that similarity('beautiful', 'nice') $> \theta$, similarity('nice', 'pretty') $> \theta$, and similarity('beautiful', 'pretty') $< \theta$. So, we conclude that the uncertainty function I_θ does not satisfy the condition of transitivity.

The vague inclusion function ν is defined the same as in [6]:

$$\nu(X, Y) = \frac{|X \cap Y|}{|X|}. \tag{5}$$

Let $S = \{S_1, S_2, \ldots, S_n\}$ be a collection of sentences, where S_i is represented by a group of words of the universe W, $1 \leq i \leq n$. Then, the fuzzy membership function μ for $w_j \in W, S_i \in S$ is expressed as

$$\mu(w_j, S_i) = \nu(I_\theta(w_j), S_i) = \frac{|I_\theta(w_j) \cap S_i|}{|I_\theta(w_j)|}. \tag{6}$$

Suppose that all the tolerance classes of words are structural subsets in the whole process, i.e., for any

$w_i \in W$, $P(I_\theta(w_i)) = 1$. Then, we define the upper approximation $\mathbf{U}(\mathbf{R}, S_i)$ and lower approximation $\mathbf{L}(\mathbf{R}, S_i)$ in R of any $S_i \in D$ as

$$\mathbf{U}(\mathbf{R}, S_i) = \left\{w_j \in W | \nu\left(\mathbf{I_\theta}\left(\mathbf{w_j}\right), \mathbf{S_i}\right) \geq \alpha\right\}, \tag{7}$$

$$\mathbf{L}(\mathbf{R}, S_i) = \left\{w_j \in W | \nu\left(\mathbf{I_\theta}\left(\mathbf{w_j}\right), \mathbf{S_i}\right) \geq \beta\right\}, \tag{8}$$

where $\alpha \in [0, 1)$, $\beta \in (0, 1]$, $\alpha < \beta$. $\mathbf{U}(\mathbf{R}, S_i)$ and $\mathbf{L}(\mathbf{R}, S_i)$ in R are also written as $\overline{S_i}$ and $\underline{S_i}$.

If S_i is regarded as one certain concept about the vague description of feature w_i; then $\mathbf{U}(\mathbf{R}, S_i)$ can be explained as a collection of concepts that share some semantics with S_i, and $\mathbf{L}(\mathbf{R}, S_i)$ can be explained as a collection of the core concepts of S_i. The probability values α and β can be used to adjust the accuracy of upper approximation and lower approximation.

Each sentence is denoted by two fuzzy sets on both upper approximation and lower approximation. Assume that one sentence S_1 is made up of a collection of words $\{w_1, w_2, \ldots, w_m\}$; then the upper approximation and lower approximation of S_1 are represented by

$$\overline{S_1} = \sum_{i=1}^{|\mathbf{U}(\mathbf{R}, S_1)|} \frac{\mu\left(w_j, S_1\right)}{w_j}, \tag{9}$$

$$\underline{S_1} = \sum_{j=1}^{|\mathbf{L}(\mathbf{R}, S_1)|} \frac{\mu\left(w_j, S_1\right)}{w_j}. \tag{10}$$

Considering the membership degree only, the upper approximation and lower approximation of S_1 can also be written as

$$\begin{aligned} \overline{S_1} &= \{u_{11}(w_1), u_{12}(w_2), \ldots, u_{1i}(w_k), \ldots, u_{1n}(w_n)\}, \\ \underline{S_1} &= \{l_{11}(w_1), l_{12}(w_2), \ldots, l_{1i}(w_i), \ldots, l_{1n}(w_n)\}, \end{aligned} \tag{11}$$

where $u_{11}(w_k)$ and $l_{11}(w_k)$ denote the membership degrees of w_k to $\overline{S_1}$ and $\underline{S_1}$, respectively. The upper approximation represents the expanded semantics of sentence S_1, capturing the latent semantics that S_1 contains. The similarity between two sentences can be measured by both the upper approximation similarity and lower approximation similarity of the two sentences. From the two different perspectives, both the expanded semantics similarity and the core semantics similarity can be captured sufficiently. For each sentence has been represented by two fuzzy sets, we employ two measurements to calculate the similarity between the two fuzzy sets, as defined as follows.

Measurement 1.

$$\mathrm{sim}_1\left(\overline{S_1}, \overline{S_2}\right) = \frac{\sum_{k=1}^{N} \min\left(u_{1k}, u_{2k}\right)}{\sum_{k=1}^{N} \max\left(u_{1k}, u_{2k}\right)}, \tag{12}$$

$$\mathrm{sim}_1\left(\underline{S_1}, \underline{S_2}\right) = \frac{\sum_{k=1}^{N} \min\left(l_{1k}, l_{2k}\right)}{\sum_{k=1}^{N} \max\left(l_{1k}, l_{2k}\right)}. \tag{13}$$

Measurement 2.

$$\mathrm{sim}_2\left(\overline{S_1}, \overline{S_2}\right) = \frac{\sum_{k=1}^{N} \min\left(u_{1k}, u_{2k}\right)}{(1/2)\sum_{k=1}^{N}\left(u_{1k} + u_{2k}\right)}, \tag{14}$$

$$\mathrm{sim}_2\left(\underline{S_1}, \underline{S_2}\right) = \frac{\sum_{k=1}^{N} \min\left(l_{1k}, l_{2k}\right)}{(1/2)\sum_{k=1}^{N}\left(l_{1k} + l_{2k}\right)}. \tag{15}$$

On the basis of the upper approximations and lower approximations of the two sentences, except for representing each sentence by a pair of fuzzy sets, we propose another method to measure the similarity. Assume that the elements of the upper and lower approximation of sentence S_1 and sentence S_2 are

(i) $\mathbf{U}(\mathbf{R}, S_1) = \{w_1, w_2, \ldots, w_a\}$,
(ii) $\mathbf{L}(\mathbf{R}, S_1) = \{w_1, w_2, \ldots, w_b\}$,
(iii) $\mathbf{U}(\mathbf{R}, S_2) = \{w_1, w_2, \ldots, w_c\}$,
(iv) $\mathbf{L}(\mathbf{R}, S_2) = \{w_1, w_2, \ldots, w_d\}$, then a new similarity degree measurement is defined as follows.

Measurement 3. Consider

$$\mathrm{sim}_3\left(\overline{S_1}, \overline{S_2}\right) = \cos\left(\sum_{i=1}^{a} \mathbf{w_i}, \sum_{j=1}^{c} \mathbf{w_j}\right), \tag{16}$$

$$\mathrm{sim}_3\left(\underline{S_1}, \underline{S_2}\right) = \cos\left(\sum_{l=1}^{b} \mathbf{w_l}, \sum_{k=1}^{d} \mathbf{w_k}\right). \tag{17}$$

The lower similarity determines the degree to which two sentences are similar assuredly. Correspondingly, the upper similarity determines the degree to which two sentences are similar possibly. To measure the final similarity degree of the two sentences, we utilize the linear combination of the upper and lower approximation similarity, which is given as

$$\mathrm{sim}_i\left(S_1, S_2\right) = \lambda \cdot \mathrm{sim}_i\left(\overline{S_1}, \overline{S_2}\right) + (1 - \lambda) \cdot \mathrm{sim}_i\left(\underline{S_1}, \underline{S_2}\right), \tag{18}$$

where $i = 1, 2, 3$, λ is the linear coefficient. λ indicates the proportion of the upper approximation similarity degree and $(1 - \lambda)$ indicates the proportion of the lower approximation similarity degree. On account that the lower approximation is composed of the core semantics, the proportion of the lower approximation similarity degree is assigned a higher value than the upper approximation similarity degree. Generally, $0 \leq \lambda \leq 0.5$. (Algorithm 1)

Example 1. Here, we give an example of our proposed methods to calculate the sentence similarity. Assume that the corpus contained four sentences as follows:

(i) Three boys are jumping in the leaves.
(ii) Three kids are jumping in the leaves.

Algorithm 1: Probabilistic tolerance rough sets-based sentence similarity model
Input: A collection of sentences $S = \{S_1, S_2, \ldots, S_n\}$.
Parameters: The cosine similarity degree threshold: θ; the probabilistic value: α, β; the linear combination parameter: λ.
Output: The similarity degree between S_i and S_l,.

(1) Preprocess the sentence corpus $S = \{S_1, S_2, \ldots, S_n\}$, and generate the universe including all the distinct words of the corpus.
(2) Compute the uncertainty function $I_\theta(w_i)$ of each word in the universe according to equation (3).
(3) Suppose that the similarity degree between sentence S_i and S_l is to be calculated. Apply equation (6) to calculate the fuzzy membership degree $\mu(w_j, S_i)$ of each word in sentence S_i, $1 \le j \le N$, $1 \le i \le n$.
(4) Obtain the upper approximation $\mathbf{U}(\mathbf{R}, S_i)$ and lower approximation $\mathbf{L}(\mathbf{R}, S_i)$ of each sentence $S_i \in S$ according to equation (7) equation and (8). Similarly, acquire $\mathbf{U}(\mathbf{R}, S_l)$ and $\mathbf{L}(\mathbf{R}, S_l)$.
(5) Represent the upper approximation and lower approximation of S_i and S_l as fuzzy sets according to equation (9) and equation (10), which are written as $\overline{S_i}$, $\underline{S_i}$, $\overline{S_l}$ and $\underline{S_l}$.
(6) Calculate the upper approximation similarity $\mathrm{sim}(\overline{S_i}, \overline{S_l})$ between $\overline{S_i}$ and $\overline{S_l}$ and the lower approximation similarity $\mathrm{sim}(\underline{S_i}, \underline{S_l})$ between $\underline{S_i}$ and $\underline{S_l}$ according to equations (12)-(17) of the three measurements, respectively.
(7) Obtain the final sentence similarity degree $\mathrm{sim}(\overline{S_i}, \overline{S_l})$ between S_i and S_l utilizing the linear combination in equation (18).

Algorithm 1: The procedure of our proposed model in detail.

(iii) Three kids are sitting in the leaves.

(iv) Children in red shirts are playing in the leaves.

After preprocessing every sentence, 9 words are included in the corpus. Then, let the universe be the set of words $U = \{\text{boys, jumping, leaves, kids, sitting, children, red, shirts, playing}\}$. Then, we illustrate the proposed probabilistic tolerance rough sets-based sentence similarity model for computing the similarity degree of the following sentences:

(i) S_1: Three boys are jumping in the leaves.

(ii) S_2: Three kids are jumping in the leaves.

Here, we set the similarity degree threshold $\theta = 0.6$ and the probabilistic values $\alpha = 0$ and $\beta = 0.7$. Then, the upper and lower approximations of these two sentences are shown in Table 1.

The upper approximation similarity degrees and lower approximation similarity degrees by the proposed three measurements are listed in Table 2.

Let the linear combination coefficient $\lambda = 0.4$; then the final similarity degrees between S_1 and S_2 by three measurements are as follows:

(i) $\mathrm{sim}_1(S_1, S_2) = 0.967$,
(ii) $\mathrm{sim}_2(S_1, S_2) = 0.940$,
(iii) $\mathrm{sim}_3(S_1, S_2) = 0.987$.

It is apparent that our proposed probabilistic tolerance rough sets-based sentence similarity algorithm can reflect the similarity relation between sentences commendably. Firstly, from the sentences S_1 and S_2, it is evident that both of them express the core semantics of "jumping" and "leaves," just like the lower approximation obtained by our algorithm. Secondly, the lower approximation similarity degree is computed to be 1, which means that S_1 and S_2 share the same core meaning. Thirdly, from the upper approximation of S_2, it can be seen that the word "children" did not originally belong to S_2, but the meaning of "children" is mined through our method. The new meaning "children" comes from the tolerance class of the word "kid," so, in a sense, "children" is the explanation of "kids." Therefore, our proposed methods can capture some latent semantics behind texts from upper approximation, which can better distinguish whether two sentences are similar from a more general perspective. Analogously, our proposed algorithms can refine the core semantics of texts by the lower approximation, which can preferably analyze sentence similarity from a more accurate perspective.

Table 1: Approximations of each sentence.

	Upper approximation	Lower approximation
S_1	Boys, jumping, leaves	Jumping, leaves
S_2	Jumping, leaves, kids, children	Jumping, leaves

Table 2: Upper and lower approximation similarity degrees on three measurements.

	Upper approximation similarity	Lower approximation similarity
M_1	0.918	1.000
M_2	0.850	1.000
M_3	0.969	1.000

Example 2. We use the traditional tolerance rough set model [6] on Example 1 for comparison. The word cooccurrence degree is set as 2. Then, the upper and lower approximations can be seen in Table 3.

Table 4 displays the corresponding upper and lower approximation similarity degrees.

Then, the sentence similarity degrees of the three measurements are as follows:

(i) $\mathrm{sim}_1(S_1, S_2) = 0.517$,
(ii) $\mathrm{sim}_2(S_1, S_2) = 0.476$,
(iii) $\mathrm{sim}_3(S_1, S_2) = 0.799$.

From the results, we can see that it provides a worse performance in contrast with our methods.

Then, we discuss the condition that one sentence "Three boys are sitting in the leaves" is added to the corpus in

Table 3: Approximations of each sentence by traditional tolerance rough set model.

	Upper approximation	Lower approximation
S_1	Boys, jumping, leaves, kids	Boys, jumping
S_2	Jumping, leaves, kids	Jumping, leaves, kids

Table 4: Upper and lower approximation similarity degrees on three measurements by traditional tolerance rough set model.

	Upper approximation similarity	Lower approximation similarity
M_1	0.542	0.500
M_2	0.703	0.324
M_3	0.936	0.708

Example 1. The whole computational process and results by the probabilistic tolerance rough sets do not alter. However, the procedures have to repeat from the calculation of uncertainty function by the traditional tolerance rough sets; then the new upper and lower approximation of S_1 and S_2 are illustrated in Table 5. Thus, the applicability of the model [6] has been greatly reduced.

4. Experimental Results and Discussion

In this section, we take from SICK2014 task and STSbenchmark dataset to evaluate the performance of our methods.

4.1. Dataset and Preprocessing. SICK2014 [33] is a dataset for the similarity evaluation of sentence pairs, which contains the training set, trial set, and testing set for a total of 15000 sentence pairs. Since our proposed model is unsupervised, which do not require additional training on the dataset, we select the 5000 sentence pairs of the training set for the experiments. And each sentence pair has been assigned a similarity score from 0 to 5 by experts. Table 6 shows two examples in the SICK2014 dataset.

STS is the abbreviation for Semantic Textual Similarity. The SemEval STS datasets from 2012 [34] to 2017 [35] were selected for this dataset. Each sentence pair has been assigned a similarity score from 0 to 5 by experts. STS-train, STS-dev, STS-test, and MSRvid are chosen for the experiments.

For better comparison with our experimental results, we have normalized the similarity score. We take the word embedding trained by Google [23] as the word vector in the experiment.

4.2. Evaluation Metrics. We exploit the Pearson correlation coefficient (Pcc) [36] and mean square error (MSE) [37] to evaluate the performance of sentence similarity measurements.

Pcc is a linear correlation coefficient that reflects the linear correlation of two variables. As for two variants X and Y, the mathematical expression of Pcc is defined as

Table 5: New approximations of each sentence by traditional tolerance rough set model.

	Upper approximation	Lower approximation
S_1	Boys, jumping, leaves, kids, sitting	Boys, jumping
S_2	Boys, jumping, leaves, kids, sitting	Jumping, kids

Table 6: Example in dataset.

Sentence A	Sentence B	Relatedness score
A man is jumping into an empty pool	A man is jumping into a full pool	3
Two young girls are sitting on the ground	Two girls are sitting on the ground	4.4

$$\rho_{xy} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X-EX)(Y-EY))}{\sqrt{D(X)}\sqrt{D(Y)}}, \quad (19)$$

where $\mathrm{Cov}(X,Y)$ is the covariance of X and Y, $D(X)$ and $D(Y)$ denote the standard deviation of X and Y individually, and EX refers to the mathematical expectation of X. The greater the absolute value of Pcc, the stronger the correlation is.

MSE is a measure reflecting the degree of difference between estimator value and real value. The definition of MSE is

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y-\widehat{y})^2, \quad (20)$$

where N is the sample size, y is the real value, and $\widehat{y}$ is the estimator value. A smaller value of MSE demonstrates a smaller deviation between the estimator value and the real value.

4.3. Experimental Results and Analysis. We proposed three sentence similarity measurements based on the probabilistic tolerance rough set model. The performances on the SICK2014 dataset are displayed in Table 7. In the table, BERT-687 and BERT-1024 are two different BERT models for sentence representation, and the sentence similarity is calculated by the cosine similarity. Fuzzy rough is the model proposed in [12]. As can be seen in Table 7, on the whole, the three measures have much better performance than the other three models. Obviously, the results of Measurement 3 achieve at the optimal performance of Pcc as 0.725 and MSE as 0.033. Particularly for the value of MSE, it is evident that there is very small error between the sentence similarity degree calculated by our methods and the real value.

Tables 8 and 9 show the Pcc and MSE results on the STSbenchmark dataset. From the tables, we can see that all of the three measures have much better performance than the results by BERT on the four datasets of STSbenchmark. The reason is that more latent semantics behind sentences can be captured by our models. Therefore, the experimental results confirm the efficiency and applicability of our methods.

4.4. Cosine Similarity Degree Threshold. In our improved probabilistic tolerance rough set model, the cosine similarity degree threshold θ controls the accuracy of the uncertainty function. The higher the value of θ, the more precise the

Table 7: Experimental results of various measurements of sentence similarity on SICK2014.

	Pcc	MSE
BERT-687	0.611	0.104
BERT-1024	0.656	0.097
Fuzzy rough	0.609	0.863
M1	0.625	0.086
M2	0.572	0.137
M3	**0.725**	**0.033**

Table 8: PCC of various measurements of sentence similarity on STSbenchmark.

Dataset	Pcc				
	BERT-768	BERT-1024	M1	M2	M3
MSRvid	0.060	0.581	0.668	0.694	**0.819**
STS-train	0.514	0.597	0.628	0.655	**0.690**
STS-dev	0.569	0.620	0.655	0.637	**0.693**
STS-test	0.454	0.579	0.633	0.655	**0.714**

Table 9: MSE of various measurements of sentence similarity on STSbenchmark.

Dataset	MSE				
	BERT-768	BERT-1024	M1	M2	M3
MSRvid	0.349	0.327	0.020	**0.014**	0.019
STS-train	0.251	0.238	0.021	**0.015**	0.018
STS-dev	0.311	0.256	0.021	**0.015**	0.022
STS-test	0.279	0.283	0.024	**0.017**	0.021

uncertainty function. However, too high value of θ will result in inadequate semantics mining. Too small value of θ will lead to more redundancy and noisy information. The interaction of θ on Example 1 can be identified in Table 10.

Figures 1 and 2 reveal the interactions of different cosine similarity threshold value θ on Pcc and MSE, respectively. In this experiment, α is set as 0, β is set as 0.6, and λ is set as 0.3. θ ranges from 0.5 to 1. As shown in Figure 1, the value of Pcc increases from $\theta = 0.5$, achieves the peak at $\theta = 0.9$, then decreases. Similarly, the value of MSE decreases from $\theta = 0.55$, achieves the minimum at $\theta = 0.95$, then increases. We can conclude that the interaction of θ satisfies the regular analyzed above.

4.5. Probability Value. The values of α and β are used for adjusting the precision of upper approximation and lower approximation. α determines the range of upper approximation. The smaller the value of $\alpha$is, the more elements the upper approximate set has. In the traditional tolerance rough sets, α is 0, by which the upper approximation contains the most information. It can reduce the generation of redundant information and does not lose too much potential semantics information via adjusting the value of α. The influence of α on Example 1 can be observed in Table 11. Similarly, β determines the range of lower approximation. A larger value of β leads to fewer elements of the lower approximate set. When $\beta = 1$, the fewest elements are included in the lower approximation, which may cause the loss of some core semantics information. Adjusting β properly may better and more adequately mine core semantics information. The effect of β on Example 1 can be observed in Table 12.

Table 10: Approximations of each sentence on different θ.

	θ	Upper approximation	Lower approximation
S_1	0.5	Boys, jumping, leaves, kids, children	Jumping, leaves
S_2	0.5	Boys, jumping, leaves, kids, children	Jumping, leaves
S_1	0.7	Boys, jumping, leaves	Jumping, leaves
S_2	0.7	Jumping, leaves, kids, children	Jumping, leaves

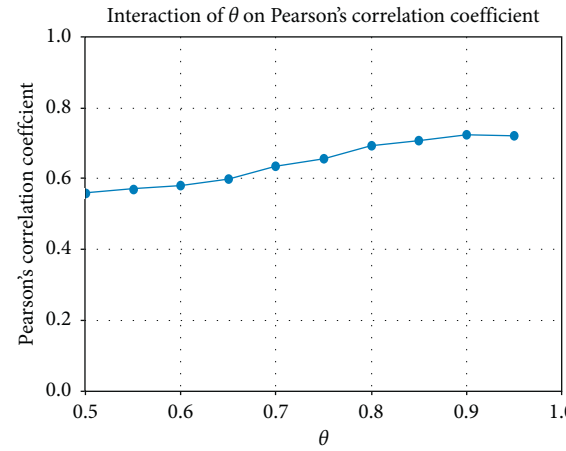

Figure 1: Interaction of cosine similarity threshold value θ on Pearson's correlation coefficient.

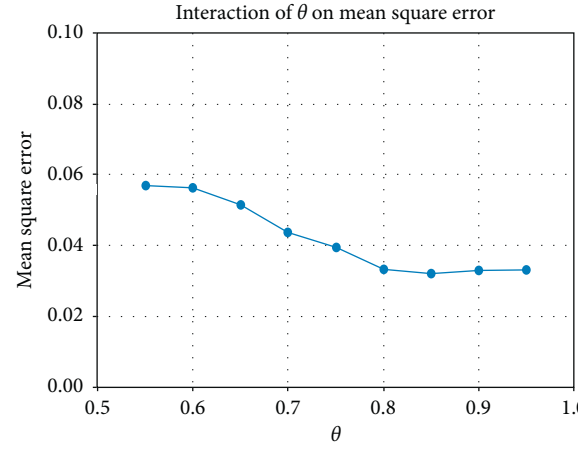

Figure 2: Interaction of cosine similarity threshold value θ on mean square error.

Table 11: Approximations of each sentence on different α.

	α	Upper approximation	Lower approximation
S_1	0	Boys, jumping, leaves, kids, children	Jumping, leaves
S_2	0	Boys, jumping, leaves, kids, children	Jumping, leaves
S_1	0.3	Boys, jumping, leaves	Jumping, leaves
S_2	0.3	Jumping, leaves, kids, children	Jumping, leaves
S_1	0.5	Jumping, leaves	Jumping, leaves
S_2	0.5	Jumping, leaves	Jumping, leaves

Table 12: Approximations of each sentence on different β.

	β	Upper approximation	Lower approximation
S_1	0.3	Boys, jumping, leaves, kids, children	Boys, jumping, leaves, kids, children
S_2	0.3	Boys, jumping, leaves, kids, children	Boys, jumping, leaves, kids, children
S_1	1	Boys, jumping, leaves, kids, children	Jumping, leaves
S_2	1	Boys, jumping, leaves, kids, children	Jumping, leaves

5. Conclusion

In this paper, owing to the property of uncertainty of text data, we incorporate the probabilistic tolerance rough sets to establish a novel sentence similarity computation model. For the reason that the traditional tolerance rough set model is not incremental and has high complexity, we make some improvement to it, making the model becoming incremental and reducing the time complexity. Through introducing the probability values α and β, the accuracy of the upper approximation and lower approximation can be adjusted. The upper approximation and lower approximation are served to represent every sentence. And on this basis, three sentence similarity calculation measurements are proposed. Upper approximation similarity and lower approximation similarity are individually calculated of each sentence pair. The linear combination of the upper approximation similarity and the lower approximation similarity is used to indicate the total sentence similarity. On the one hand, it can dig out more latent semantics information than the traditional methods based on shallow semantics. On the other hand, it is unsupervised, which relieves the defect of supervised deep learning-based methods. We carry out some experiments on the SICK2014 task to evaluate the performance of our proposed model. The results verify the efficiency and applicability of the proposed models.

The proposed model is established without considering the order of sentences, in which our future work will include it.

Data Availability

The SICK2014 task data used to support the findings of this study are available from clic.cimec.unitn.it/composes/sick.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant nos. 11671001 and 61876201).

References

[1] N. Chatterjee and N. Yadav, "Fuzzy rough set-based sentence similarity measure and its application to text summarization," *IETE Technical Review*, vol. 36, no. 5, pp. 517–525, 2019.

[2] S. Abujar, M. Hasan, and S. A. Hossain, "Sentence similarity estimation for text summarization using deep learning," in *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology*, pp. 155–164, Springer, Comilla, Bangladesh, October 2019.

[3] K. Wu, X. Wang, and A. Aw, "Bilingual word embedding with sentence similarity constraint for machine translation," in *Proceedings of the 2017 International Conference on Asian Language Processing (IALP)*, pp. 119–122, IEEE, Singapore, December 2017.

[4] W. An, Q. Chen, W. Tao et al., "ECNU at 2017 LiveQA track: learning question similarity with adapted long short-term memory networks," in *Proceedings of the Twenty-Sixth Text REtrieval Conference*, pp. 1–9, TREC, Gaithersburg, MD, USA, 2017.

[5] Y. Wang, Q. Hu, Y. Song, and L. He, "Potentiality of healthcare big data: improving search by automatic query reformulation," in *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data)*, pp. 807–816, Morgan Kaufmann, San Mateo, CA, USA, December 2017.

[6] T. B. Ho and N. B. Nguyen, "Nonhierarchical document clustering based on a tolerance rough set model," *International Journal of Intelligent Systems*, vol. 17, no. 2, pp. 199–212, 2002.

[7] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 60, no. 5, pp. 493–502, 2004.

[8] L. Han, T. Finin, P. Mcnamee, A. Joshi, and Y. Yesha, "Improving word similarity by augmenting PMI with estimates of word polysemy," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1307–1322, 2013.

[9] Y. Bi, K. Deng, and J. X. Cheng, *A Keyword-Based Method For Measuring Sentence Similarity*, pp. 379-380, ACM, New York, NY, USA, 2017.

[10] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and wordnet-based approaches," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 19–27, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009.

[11] P. D. Turney, "Mining the web for synonyms: PMI-IR versus LSA on TOEFL," in *Proceedings of the European Conference on Machine Learning*, pp. 491–502, Springer, Freiburg, Germany, September 2001.

[12] M. K. Prasad and P. Sharma, "Combining common words and semantic features for sentence similarity," in *Proceedings of the 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–4, IEEE, Bangalore, India, July 2018.

[13] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Advances in Neural Information Processing Systems*, pp. 2042–2050, 2014.

[14] H. He, K. Gimpel, and J. Lin, "Multi-perspective sentence similarity modeling with convolutional neural networks," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1576–1586, Lisbon, Portugal, September 2015.

[15] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 2786–2792, Phoenix, AZ, USA, February 2016.

[16] Z. Wang, H. Mi, and A. Ittycheriah, "Sentence similarity learning by lexical decomposition and composition," 2016, https://arxiv.org/abs/1602.07019.

[17] Q. Chen, Q. Hu, J. X. Huang, and L. He, "CA-RNN: Using context-aligned recurrent neural networks for modeling sentence similarity," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 265–273, New Orleans, LA, USA, February 2018.

[18] P. Liu, Z. Zheng, and Q. Su, "Sentence similarity computation by integrating shallow and deep information," in *Proceedings of the 2018 International Conference on Asian Language Processing (IALP)*, pp. 308–311, IEEE, Bandung, Indonesia, November 2018.

[19] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," 2019, https://arxiv.org/abs/1810.04805.

[20] H.-H. Huang and Y.-H. Kuo, "Cross-lingual document representation and semantic similarity measure: a fuzzy set and rough set based approach," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 6, pp. 1098–1111, 2010.

[21] Z. a. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.

[22] C. Luo, T. Li, H. Chen, H. Fujita, and Z. Yi, "Incremental rough set approach for hierarchical multicriteria classification," *Information Sciences*, vol. 429, pp. 72–87, 2018.

[23] R. K. Nowicki, M. Korytkowski, and R. Scherer, "Rough neural network ensemble for interval data classification," in *Proceedings of the 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–7, IEEE, Rio de Janeiro, Brazil, July 2018.

[24] Y. Ma, X. Luo, X. Li, Z. Bao, and Y. Zhang, "Selection of rich model steganalysis features based on decision rough set α-positive region reduction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 336–350, 2018.

[25] Y. Yao, "Probabilistic rough set approximations," *International Journal of Approximate Reasoning*, vol. 49, no. 2, pp. 255–271, 2008.

[26] Z. Pawlak and R. Sɵwinski, "Rough set approach to multi-attribute decision analysis," *European Journal of Operational Research*, vol. 72, no. 3, pp. 443–459, 1994.

[27] A. Skowron and J. Stepaniuk, "Tolerance approximation spaces," *Fundamenta Informaticae*, vol. 27, no. 2, 3, pp. 245–253, 1996.

[28] C. L. Ngo and H. S. Nguyen, "A method of web search result clustering based on rough sets," in *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pp. 673–679, IEEE, 2005.

[29] X.-J. Meng, Q.-C. Chen, and X.-L. Wang, "A tolerance rough set based semantic clustering method for web search results," *Information Technology Journal*, vol. 8, no. 4, pp. 453–464, 2009.

[30] B. K. Patra and S. Nandi, "Fast single-link clustering method based on tolerance rough set model," in *Proceedings of the International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, pp. 414–422, Springer, Delhi, India, December 2009.

[31] R. Zhao and K. Mao, "Fuzzy bag-of-words model for document representation," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 2, pp. 794–804, 2017.

[32] http://code.google.com/archive/p/word2vec.

[33] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "Semeval-2014 task 1: evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment," in *Proceedings of the 8th International Workshop on Semantic Evaluation*, pp. 1–8, August 2014.

[34] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "Semeval-2012 task 6: a pilot on semantic textual similarity," in *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval-2012)*, p. 385C393, Association for Computational Linguistics, Montréal, Canada, June 2012.

[35] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, September 2017, Article ID 670C680.

[36] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, pp. 1–4, Springer, Berlin, Germany, 2009.

[37] J. M. Lowerre, "On the mean square error of parameter estimates for some biased estimators," *Technometrics*, vol. 16, no. 3, pp. 461–464, 1974.