

Article

A Quantum Language-Inspired Tree Structural Text Representation for Semantic Analysis

Yan Yu ¹, Dong Qiu ^{2,*}  and Ruiteng Yan ¹

¹ School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; d180201018@stu.cqupt.edu.cn (Y.Y.); s180231039@stu.cqupt.edu.cn (R.Y.)

² College of Science, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

* Correspondence: qiudong@cqupt.edu.cn

Abstract: Text representation is an important topic in the field of natural language processing, which can effectively transfer knowledge to downstream tasks. To extract effective semantic information from text with unsupervised methods, this paper proposes a quantum language-inspired tree structural text representation model to study the correlations between words with variable distance for semantic analysis. Combining the different semantic contributions of associated words in different syntax trees, a syntax tree-based attention mechanism is established to highlight the semantic contributions of non-adjacent associated words and weaken the semantic weight of adjacent non-associated words. Moreover, the tree-based attention mechanism includes not only the overall information of entangled words in the dictionary but also the local grammatical structure of word combinations in different sentences. Experimental results on semantic textual similarity tasks show that the proposed method obtains significant performances over the state-of-the-art sentence embeddings.

Keywords: natural language processing; syntax tree; attention mechanism; semantic analysis; quantum language-inspired



Citation: Yu, Y.; Qiu, D.; Yan, R. A Quantum Language-Inspired Tree Structural Text Representation for Semantic Analysis. *Mathematics* **2022**, *10*, 914. <https://doi.org/10.3390/math10060914>

Academic Editors: Weihua Xu, Jinhai Li and Xibei Yang

Received: 8 February 2022

Accepted: 10 March 2022

Published: 13 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The parallelism of quantum computing has attracted more and more attention in different fields. Some scholars have combined quantum computing with natural language processing (NLP). In the quantum computing-based text representation model, the word vector was multiplied by its own transposed vector tensor to obtain a density matrix, and the weighted density matrix was summed to obtain a sentence tensor representation [1,2].

$$\rho = \sum_{i=1}^n p_i |w_i\rangle \langle w_i|, \quad (1)$$

$$\text{Sim}(S_1, S_2) = \text{tr}(\rho_1 \rho_2) = \sum_{i,j} \lambda_i \lambda_j \langle w_i | w_j \rangle^2, \quad (2)$$

where tr denoted the trace operation. The probability that event $|u\rangle \langle u|$ belongs to a system was defined as the semantic measurement [3],

$$\mu_\rho(|u\rangle \langle u|) = \text{tr}(\rho |u\rangle \langle u|) = \langle u | \rho | u \rangle \quad (3)$$

where $\mu_\rho \in [0, 1]$.

The association between adjacent words with all of the entanglement coefficients set to 1 was discussed, ignoring the long-range modified relationship between non-adjacent words [4]. When the sentence structure is complex, the two words that have a direct modified relationship are not necessarily in close proximity. Especially for long sentences, the complex syntactic structure makes adjacent words not necessarily grammatically related, and the grammatically related words separated by several words. Take the following sentence for example with the constituency parser and the dependency parser (<https://>

[//nlp.stanford.edu/software/lex-parser.shtml](http://nlp.stanford.edu/software/lex-parser.shtml) (accessed on 1 October 2020)) of S_1 shown in Figure 1.

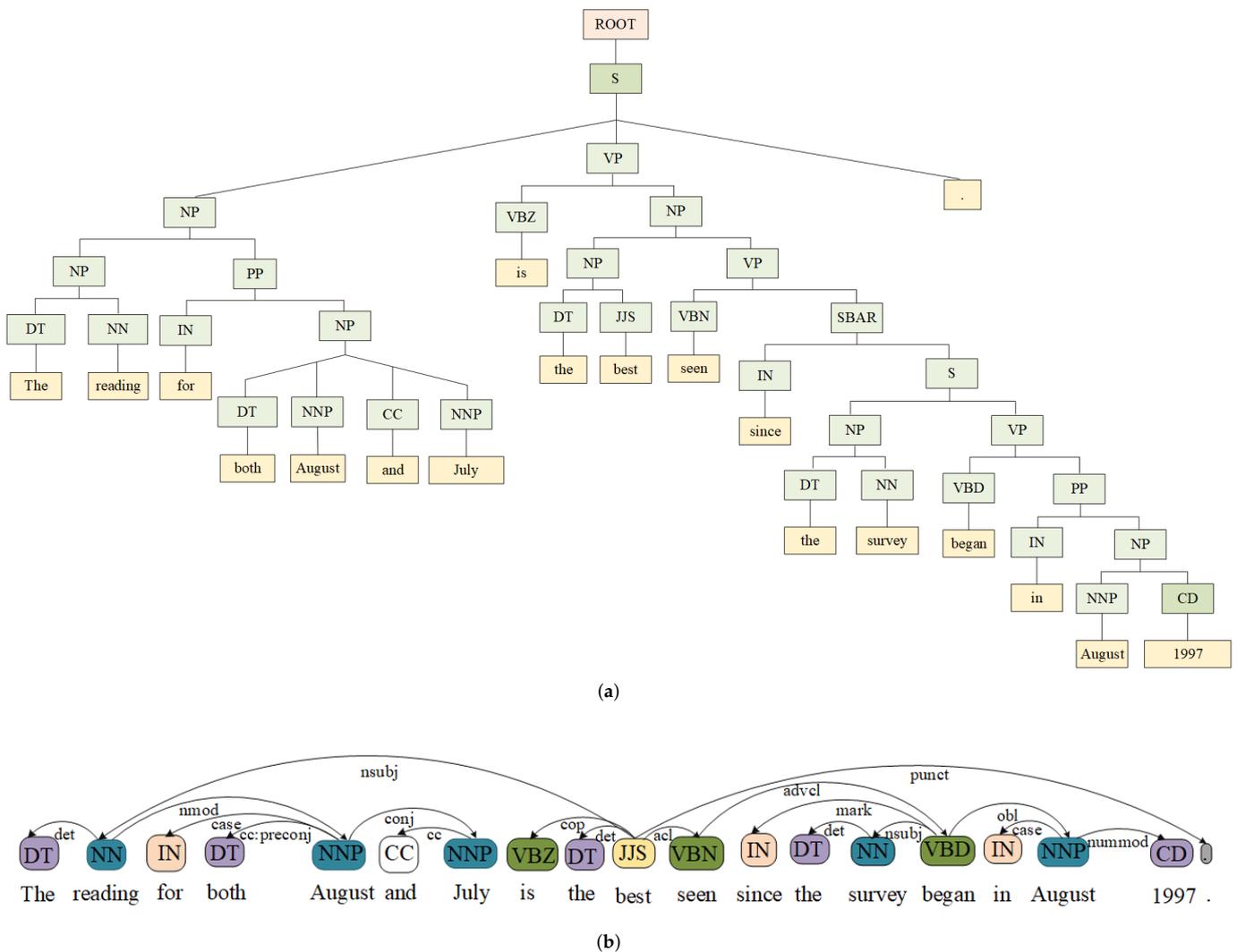


Figure 1. S_1 : (a) constituency parser; (b) dependency parser. S_1 : The reading for both August and July is the best seen since the survey began in August 1997.

According to the adjacent words to form a related phrase, the closest to the word *best* is the nominal word *July*, but the two syntax trees in Figure 1 both show that there is no direct modification relationship between the two words. The dependent relation between *reading* and *best* is *nsubj* (nominal subject), but the distance between them is 8 words, meaning that at least 9-gram can contain both of the words. If the 9-gram is considered, it will contain at least seven irrelevant words except *reading* and *best*, thus introducing semantic errors. Furthermore, the distance difference between two words of different relation entities is different, which requires the size of the kernel function to change according to the change of the distance difference between the two ends of the relation entity. Obviously, this is difficult to achieve in reality. As different weights of part-of-speech (PoS) combinations of entangled word have different influences on sentence semantics [5], PoS combination weight can be integrated with the attention mechanism to express the different modified relationship. Inspired by the density matrix and attention mechanism, a quantum language-inspired tree structural text representation model is established to reflect the association between variable distance words.

As different syntactic tree structures reflect different associations between words, different semantic association models between words according to the dependency parser and constituency parser of sentences are constructed. According to the association between relation entities in the dependency parser of the sentence, the text representation based on the dependency parser combines the word vector tensors of two words with relation entities to establish the semantics between long-distance dependent words with relation entities entanglement, so that distant words with a direct modified relationship can also be semantically related. According to the different degrees of modified relationship between words in the constituency parser, the text representation based on the constituency parser combines the semantic correlation coefficient with the distribution characteristics of adjacent words to establish the semantic association. The proposed model consists of two parts, as shown in Figure 2. The first part is composed of all of the tensor products of the two adjacent words in a sentence, combining the characteristics to establish the contribution of the short-range dependence between words to the semantics. The second part consists of the two words with long-range dependency of the direct modified relationship. Finally, the entanglement between adjacent words is integrated with the word entanglement of direct long-range modified relationship to form the sentence representation.

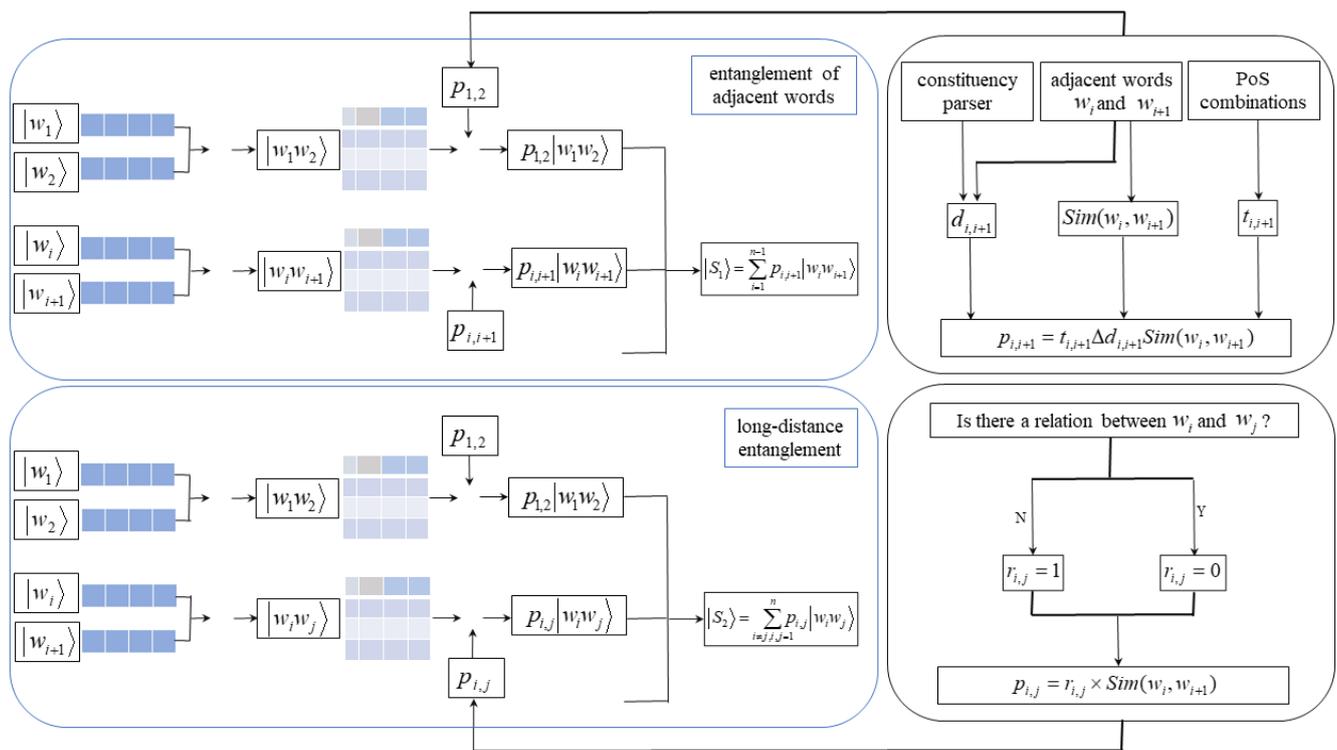


Figure 2. Flowchart of the quantum language-inspired tree structural text representation model.

In brief, the contributions of this work are as follows.

- (1) A quantum language-inspired text representation model based on relation entity and constituency parser is established, including long-range and short-range semantic associations between words.
- (2) The combination of attention mechanism and entanglement coefficient reduces the semantic impact of indirect modified relationships between adjacent words and enhances the semantic contribution of direct modified relationships with long-range associations.
- (3) The attention mechanism contains not only the overall information of the related words in the dictionary, but also the local grammatical structure of different sentences.

- (4) The semantic association between words with variable distances is established by combining the dependency parser.

The rest of the paper is organized as follows. Section 2 summarizes some related literature on attention-based semantic analysis, the dependency tree and quantum based NLP. Section 3 explains the approach in detail. The experimental settings are presented in Section 4. Section 5 describes the experimental results and lists the detailed effects of different parameters. In Section 6, some conclusions are drawn.

2. Related Work

2.1. Attention-Based Semantic Analysis

The neural network-based methods use the attention mechanism to assign different semantic weights to words with good experimental results in many downstream tasks, such as LSTM [6], BiLSTM [7] and BERT [8]. Semantic analysis based on attention mechanism has been involved in many works [9–11] and can reflect the different weights of words in different texts. The attention mechanism is introduced to obtain different weight of words in order to extract enough key information. Semantic analysis can be applied to many problems such as image-text matching [12], question answering [13,14], knowledge extraction [15–19], and entailment reasoning [20]. To discover visual-textual interactions across different dimensions of concatenations, memory attention networks were adopted while marginalizing the effect of other dimensions [12]. A deep multimodal network with a top- k ranking loss mitigated the data ambiguity problem for image-sentence matching [21]. LSTM with a bilinear attention function was adopted to infer the image regions [22]. A mutual attention mechanism between the local semantic features and global long-term dependencies was introduced for mutual learning [23]. A scheme of the efficient semantic label extraction was developed to achieve an accurate image-text similarity measure [24].

Compared to the previous works mentioned above, the method proposed here is mainly based on the traditional attention mechanism, which mainly reflects the relation dependency between input and hidden without considering the relations of input words. The transformer mainly relies on attention mechanisms [25]. An improved self-attention module was proposed by introducing low-rank and locality linguistic constraints [26]. With the introduction of the self-attention mechanism, new models based on transformer obtained much success in semantic analysis on large datasets [27–29]. The new model BERT [30] and its variants [31,32] based on transformer divided the pretraining methods into feature-based methods and fine-tuning methods [33].

2.2. Dependency Tree

Relation extraction plays a very important role in extracting structured information from unstructured text resources. The dependency tree not merely expresses the semantics of sentences but also reflects the modified relationship relationship between words. Each node in dependency trees represents a word, and every word has at least one grammatically related word. The dependency tree is constituted by the head word, PoS of the head word, dependent word, PoS of the dependent word and the label of dependency. The works on the dependency tree are mainly divided into two categories: statistics-based models and deep-learning-based models [34]. Wang et al. structured a regional CNN-LSTM model based on a subtree to analyze sentiment predictions [35]. A reranking approach for the dependency tree was provided utilizing complex subtree representations [36]. A bidirectional dependency tree representation was provided to extract dependency features from the input sentences [37]. Zhang et al. [38] tried to upgrade the synchronous tree substitution grammar-based syntax translation model by utilizing the string-to-tree translation model. A graph-based dependency parsing model was presented by Chen et al. [39]. A bidirectional tree-structured LSTM was provided to extract structural features based on the dependency tree [40]. Fei et al. utilized a dependency-tree-based RNN to extract syntactic features and used the CRF layer to decode the sentence labels [41]. Global reasoning on a depen-

dependency tree parsed from the question was performed [42]. A phrase-based text embedding considering the integrity of semantic with a structured representation is reported [43].

2.3. Quantum Based NLP

In recent years, the application of quantum language models (QLM) in NLP has attracted more and more attention [44]. Aerts and Sozzo theoretically proved that under some conditions, the joint probability density of the two entities selected to define the uncertainty of selection to establish an entanglement between the concepts was reasonable [45]. Quantum theory is applied to neural networks to form quantum neural networks. To achieve comparable or better results, quantum network needed far fewer epochs and a much smaller network [46]. Quantum probability was first practically applied in information retrieval (IR) with significant improvements over a robust bag-of-words baseline [47–52]. At the same time, an unseparable semantic entity was used in IR, considering the pure high-order dependence among words or phrases [53]. On this basis, quantum entanglement (QE) was applied to terms dependency co-occurrences on quantum language models with theoretical proof of the connection between QE and statistically unconditional pure dependence [54].

Commonly, semantic representation generalized quantum or quantum-like use Hilbert spaces to model concepts, and the similarity is measured by scalar product and projection operators [55]. Density matrix representation can be used to many fields, such as document interaction [2], different modality correlations [56] and sentiment analysis [1,3,57]. The mathematical formalism of quantum theory could resist traditional model with significant effectiveness in cognitive phenomena [58]. A semiotic interpretation of the role played based on quantum entanglement was provided to find a finite number of the smallest semantic units to form every possible complex meaning [59].

3. Approaches

3.1. Read Text and Generate Syntax Tree

Each sentence is read, and the dependency tree and relation entity are generated. The word2vec embedding, PoS tagging and tree depth of each word in a sentence are stored, forming a quadruple $\mathcal{T}(\mathcal{A}, \mathcal{P}, \mathcal{H}, \mathcal{R})$. The \mathcal{A} array stores all of the words with the original sequences in the sentence.

$$\mathcal{A} = \{w_1, w_2, \dots, w_i, \dots, w_n\}, \quad (4)$$

where w_i is the i th word in the sentence and n is the total number of the words.

The \mathcal{P} array stores the PoS of each word in sequence.

$$\mathcal{P} = \{P_1, P_2, \dots, P_i, \dots, P_n\}, \quad (5)$$

where P_i is the PoS of the i th word w_i .

The \mathcal{H} array denotes the tree depth of the i th word w_i in the dependency tree.

$$\mathcal{H} = \{h_1, h_2, \dots, h_i, \dots, h_n\}, \quad (6)$$

where h_i represents the tree depth of the i th word w_i .

The \mathcal{R} array denotes the relation entity of the i th word w_i in the dependency tree. We only consider whether an entity relationship exists between entangled words. If a relationship exists, the words, the associated vocabularies and the relationship between them are saved.

3.2. Entanglement between Words with Short-Range Modified Relationship

The model mainly describes the entanglement between two adjacent words. The attention mechanism is introduced to highlight word entanglement with a direct modi-

fied relationship while weakening the impact of the entanglement of indirect modified relationship on sentence semantics.

3.2.1. Normalize the Word Vector

Normalize the word2vec of an input word.

$$|w_i\rangle = \frac{\vec{s}_i}{|\vec{s}_i|}, \tag{7}$$

where \vec{s}_i is the d dimensional column vector of the i th word and $|\vec{s}_i|$ is the module of \vec{s}_i . Therefore, $|w_i\rangle$ is set as $|w_i\rangle = [a_1 \ a_2 \ \cdots \ a_d]^T$, with d dimensional column vector.

3.2.2. Embedding of Entangled Word

Two adjacent words are entangled together in order, forming the arrays

$$\mathcal{B} = \{(w_1w_2), (w_2w_3), \dots, (w_iw_{i+1}), \dots, (w_{n-1}w_n)\}, \tag{8}$$

and

$$\mathcal{C} = \{(P_1P_2), (P_2P_3), \dots, (P_iP_{i+1}), \dots, (P_{n-1}P_n)\}, \tag{9}$$

where w_iw_{i+1} is the combination of the adjoint words w_i and w_{i+1} , and P_iP_{i+1} is the PoS combination of w_iw_{i+1} . The representation of (w_iw_{i+1}) is defined by the tensor product between $|w_i\rangle$ and $|w_{i+1}\rangle$.

$$\begin{aligned} |w_iw_{i+1}\rangle &= |w_i\rangle|w_{i+1}\rangle \\ &= [a_1 \ a_2 \ \cdots \ a_d]^T [b_1 \ b_2 \ \cdots \ b_d]^T \end{aligned} \tag{10}$$

For simplicity and clarity, $|w_iw_{i+1}\rangle$ is rewritten into a square matrix form.

$$|w_iw_{i+1}\rangle = \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_d \\ a_2b_1 & a_2b_2 & \cdots & a_2b_d \\ \vdots & \vdots & \ddots & \vdots \\ a_db_1 & a_db_2 & \cdots & a_db_d \end{bmatrix}. \tag{11}$$

Similarly, $|w_{i+1}w_i\rangle$ is obtained,

$$|w_{i+1}w_i\rangle = \begin{bmatrix} b_1a_1 & b_1a_2 & \cdots & b_1a_d \\ b_2a_1 & b_2a_2 & \cdots & b_2a_d \\ \vdots & \vdots & \ddots & \vdots \\ b_da_1 & b_da_2 & \cdots & b_da_d \end{bmatrix}. \tag{12}$$

Obviously, the inequality is obtained, $|w_iw_{i+1}\rangle \neq |w_{i+1}w_i\rangle$, which reflects the influence of the order of the entangled words.

3.2.3. Attention Mechanism

The attention mechanism consists of three components: the cosine similarity between the entangled words, the influence of PoS combination and the dependency tree depth difference of the two words.

The similarity between the adjacent words is defined as follows:

$$Sim(w_i, w_{i+1}) = \frac{\langle w_i | w_{i+1} \rangle}{||w_i\rangle| \cdot ||w_{i+1}\rangle|}, \tag{13}$$

where

$$\langle w_i | w_{i+1} \rangle = [a_1 \ a_2 \ \dots \ a_d] [b_1 \ b_2 \ \dots \ b_d]^T = \sum_{i=1}^d a_i b_i, \tag{14}$$

$$||w_i\rangle| = \sqrt{\sum_{i=1}^d a_i^2}, \tag{15}$$

$$||w_{i+1}\rangle| = \sqrt{\sum_{i=1}^d b_i^2}. \tag{16}$$

The weight of PoS combination of the entangled words can reflect some common modified relationships between words [5]. The different combinations of the two adjacent words are set to different values to express the different contributions for grammatical structures. We use $t_{i,i+1}$ to represent the influence of the PoS combination of the i th word and the $(i + 1)$ th word, where $t_{i,i+1} \in (0, 1)$, reflecting the global information of the corpus.

The last part is a parameter $d_{i,i+1}$, which is determined by the dependent relationship in the syntax tree and the tree depth difference. The tree depth difference is the absolute value of the difference between the tree depth of the i th word and that of the $(i + 1)$ th word.

$$\Delta h_{i,i+1} = |h_i - h_{i+1}|. \tag{17}$$

The weight of $\Delta h_{i,i+1}$ is set as follows:

$$d_{i,i+1} = \begin{cases} d_1, & \text{for } \Delta h_{i,i+1} = 0 \\ d_2, & \text{for } \Delta h_{i,i+1} \leq a, \\ d_3, & \text{for } \Delta h_{i,i+1} > a \end{cases} \tag{18}$$

where d_1, d_2 and d_3 satisfy the condition of $d_1 > d_2 > d_3, a = 2$ and the value of a can be altered.

Hence, the attention mechanism is described as follows:

$$p_{i,i+1} = Sim(w_i, w_{i+1}) \times t_{i,i+1} \times d_{i,i+1}. \tag{19}$$

3.2.4. Adjacent Words Entanglement-Based Sentence Representation

The adjacent words entanglement-based sentence representation is defined as follows.

$$|S_1\rangle = \sum_{i=1}^{n-1} p_{i,i+1} |w_i\rangle |w_{i+1}\rangle = \sum_{i=1}^{n-1} p_{i,i+1} |w_i w_{i+1}\rangle, \tag{20}$$

where $p_{i,i+1} = \frac{p_{i,i+1}}{\sum_{i=1}^{n-1} p_{i,i+1}}$.

3.2.5. Sentence Similarity

Since both the representations of $|S_1\rangle$ and $|w_i w_{i+1}\rangle$ are the second-order tensor with d^2 dimensions, the normalized dot product between the two sentence representations $|S_1\rangle$ and $|S'_1\rangle$ is defined as the sentence similarity of the sentence pair,

$$Sim(|S_1\rangle, |S'_1\rangle) = \frac{\langle S_1 | S'_1 \rangle}{\|S_1\| \cdot \|S'_1\|}, \tag{21}$$

where $\langle S_1 | S'_1 \rangle$ denotes the inner product of $\langle S_1 |$ and $|S'_1\rangle$, $\|S_1\|$ and $\|S'_1\|$ are the norms of $|S_1\rangle$ and $|S'_1\rangle$, respectively, and $\langle S_1 |$ is the conjugate transpose of $|S_1\rangle$.

3.3. Optimize the Sentence Embedding

The ultimate goal of text representation is to enable computers to understand human language. For the calculation of the text semantic similarity in this paper, the calculated value x_i is infinitely close to the artificial score y_i . Therefore, the Pearson correlation coefficient (Pcc) reaches the maximum value and the mean square error (MSE) reaches the minimum values. To maximize the Pcc and minimize the MSE, we optimize the sentence embedding using two approaches.

3.3.1. Entanglement between Words with Long-Range Modified Relationship

For long sentences or sentences with complex structures, two modified words are not necessarily adjacent. Aiming at the modified relationship of the long-distance association, a long-range dependent relationship $R_{i,j}$ is defined as the R array between words w_i and w_j .

$$R = \{R_{i,j}\}, (i, j = 1, \dots, n) \tag{22}$$

$$R_{i,j} = (r_{i,j}, |w_i\rangle, |w_j\rangle), \tag{23}$$

where $r_{i,j}$ is a binary element, defined by whether there is a correlation entity between words w_i and w_j . If there is a relation entity between words w_i and w_j , $r_{i,j} = 1$; if not, $r_{i,j} = 0$. The entanglement between w_i and w_j is

$$p_{i,j} = r_{i,j} \times Sim(w_i, w_j). \tag{24}$$

Equation (24) indicates that only the word pair with relation entity is considered to expand the sentence semantics.

Therefore, the sentence embedding based on relation entity is altered as follows.

$$\begin{aligned} |S_2\rangle &= \sum_{i < j, i, j = 1}^n p_{i,j} |w_i\rangle |w_j\rangle \\ &= \sum_{i < j, i, j = 1}^n p_{i,j} |w_i w_j\rangle, \end{aligned} \tag{25}$$

where $\sum_{i < j, i, j = 1}^n$ defines that all of the long-range dependencies in the sentence are taken into account, and $p_{i,j} = \frac{p_{i,j}}{\sum_{i, j = 1}^N p_{i,j}}$.

3.3.2. Sentence Embedding Based on Constituency Parser and Relation Entity

The entanglement between the short-range modified words and the entanglement between the long-range modified relationship form a sentence representation. Hence, the optimized sentence representation is altered as

$$\begin{aligned}
 |T\rangle &= |S_1\rangle + |S_2\rangle \\
 &= \sum_{i=1}^{n-1} p_{i,i+1} |w_i w_{i+1}\rangle + \sum_{i < j, i, j=1}^n p_{i,j} |w_i w_j\rangle.
 \end{aligned}
 \tag{26}$$

$|S_1\rangle$ uses the entanglement coefficient $p_{i,i+1}$ to highlight the semantic contribution of adjacent words with modified relationships, and to weaken the semantic contribution of the adjacent words without modified relationship. $|S_2\rangle$ only considers the semantic contribution of word pairs with a relation entity. Therefore, $|T\rangle$ includes not only the modified relationship between adjacent words, but also the long-range modified relationship between related words. In addition, the entanglement coefficient analyzes the relevant degree between words from multiple perspectives by the relation entity of the text, combining the local information of the words in the same text with the global information in a dictionary.

3.3.3. Reduce Sentence Embedding Dimensions

The semantics of some phrases cannot be expressed by any of their constituent words alone, nor can the semantics of these two words be simply added together, such as *lung cancer*. The semantics of *lung cancer* is less than the semantics adding of *lung* and *cancer*. To reduce the redundant information after word entanglement, we use the two methods of dimensionality reduction on the level of sentence embedding and of entangled word representation. For dimensionality reduction at the sentence level, we delete some smaller absolute values in the sentence embedding. At the entangled word level, we delete some smaller absolute values of the entangled word embedding to reduce the dimension of the sentence representation.

4. Experimental Settings

4.1. Parameters Definition

Some parameters and variables are defined in Table 1.

Table 1. Definitions of the parameters and variables.

Parameter/Variable	Definition
$ w_i\rangle$	word embedding of the i th word
$ w_i w_j\rangle$	entangled word embedding of the i th word and the j th word
$ T\rangle$	sentence embedding
$Sim(w_i, w_j)$	direction cosine between the word embedding of the i th word and the $(i + 1)$ th word
$t_{i,i+1}$	part of speech combination weight of the i th word and the $(i + 1)$ th word
Δh	depth difference between two words in the parser tree
$d_{i,j}$	weight of the depth difference between the i th word and the j th word
$p_{i,j}$	entanglement coefficient between the i th word and the j th word
y	annotated sentence similarity by humans
x	calculated sentence similarity by the proposed model
σ, γ	threshold value of annotated sentence similarity
λ	threshold value of relative error
ΔE	relative error between the experimental result and annotated score of sentence similarity
D	dimensionality of the sentence representation
D_1	sentence dimension reduced at the level of sentence embedding
D_2	sentence dimension reduced at the level of entangled word embedding

4.2. Datasets

Datasets include the SemEval Semantic Textual Similarity (STS) Tasks (years of 2012 (STS'12), 2014 (STS'14), 2015 (STS'15)) and STS-benchmark (STSB). E. Agirre et al. selected and piloted annotations of 4300 sentence pairs to compose the STS tasks in SemEval, including machine translation, surprise datasets and lexical resources [60]. E. Agirre et al. added new genres to the previous corpora, including 5 aspects in STS'14 [61]. In 2015, sentence pairs on answer pairs and belief annotations were added [62]. STS-benchmark and STS-companion include English text from image captions, news headlines and user forums [63]. In the study, the public lib of word2vec with 300 dimensions is assigned (<http://code.google.com/archive/p/word2vec> (accessed on 1 October 2020)). The input vectors of fasttext are 300 dimensions (<http://fasttext.cc/docs/en/english-vectors.html> (accessed on 1 October 2020)). The total number of sentence pairs for each corpus is summarized in Table 2 (<http://groups.google.com/group/STS-semeval> (accessed on 1 October 2020)). The corpora consist of English sentence pairs and the annotated similarities ranging from 0.0 to 1.0 (divided by 5). All the grammatical structures of sentences in the provided model are generated by Stanford Parser models package (<https://nlp.stanford.edu/software/lex-parser.shtml> (accessed on 1 October 2020)).

Table 2. The number of sentence pairs in each corpus.

STS'12	STS'14	STS'15	
MSRvid (750)	deft-forum (450)	answers-forums (375)	STSB (4225)
SMTeuroparl (459)	deft-news (300)	answers-students (750)	STS'12 (3108)
OnWN (750)	headlines (750)	belief (375)	STS'14 (3750)
MSRpar (750)	images (750)	images (750)	STS'15 (3000)
SMTnews (399)	tweet-news (750)	headlines (750)	
	OnWN (750)		

4.3. Experimental Settings

To make the calculated sentence similarity x_i approximately equal to the annotated score y_i , we perform a simple classification of the calculated results based on the relative error ΔE :

$$\Delta E = \frac{|x_i - y_i|}{y_i}. \quad (27)$$

When $\Delta E > \lambda$, we introduce the optimized models to recompute the sentence similarities. This is performed in order to reduce the difference between the calculated value and the labeled value:

$$\min(\sum_{\Delta E > \lambda} |x_i - y_i| + \sum_{\Delta E \leq \lambda} |x_i - y_i|). \quad (28)$$

Then, we divide the calculated results into two parts. The first is computed by the short-range entanglement, and the other is modeled by the optimized models. When $\Delta E \leq \lambda$, the calculated result is considered feasible and is stored as x_{0i} . When $\Delta E > \lambda$, the calculated result is too large and must be optimized. To exploit the advantages of the two dimensionality reduction models, we divide the sentence pairs with $\Delta E > \lambda$ into two parts according to the annotated scores y_i . When $\sigma < y_i \leq 1$, the sentence similarity is recomputed by the sentence-level dimensionality reduction model with the result of x_{1i} . When $\gamma \leq y_i < \sigma$, we utilize the entangled-word-level dimensionality reduction model to recompute the sentence similarity, with the calculated result of x_{2i} . The algorithm is listed

in Algorithm 1.

Algorithm 1: Framework of sentence embedding based on constituency parser and relation entity for semantic similarity computation.

Input: sentence pairs (T_1, T_2) , annotated score y
Output: semantic similarity of (T_1, T_2)

```

1 for each sentence do
2   generate constituency parser and relation entity
3   for each adjacent words  $w_i$  and  $w_{i+1}$  do
4     get  $|w_i\rangle$  and  $|w_{i+1}\rangle$ , PoS combination weight  $t_{i,i+1}$  of  $w_i$  and  $w_{i+1}$ 
5     compute  $|w_i w_{i+1}\rangle$  by tensor product
6     compute  $d_{i,i+1}$  according to depth difference between  $w_i$  and  $w_{i+1}$ 
7     compute  $p_{i,i+1}$ 
8   end
9   sentence embedding based on constituency parser is obtained:
10   $|S_1\rangle \leftarrow \sum_{i=1}^{n-1} (p_{i,i+1} |w_i w_{i+1}\rangle)$ 
11  for each word pair  $(w_i, w_j)$  with relation entity do
12    get  $|w_i\rangle$  and  $|w_j\rangle$ 
13    compute  $|w_i w_j\rangle$  by tensor product
14    determine  $r_{i,j}$  whether there is a relation entity between  $w_i$  and  $w_j$ 
15    compute  $|w_i w_j\rangle$  by tensor product
16    compute  $p_{i,j}$ 
17    sentence embedding based on relation entity is obtained:
18     $|S_2\rangle \leftarrow \sum_{i < j, i=1}^{n-1} (p_{i,j} |w_i w_j\rangle)$ 
19  end
20  sentence embedding based on constituency parser and relation entity:
21   $|T_1\rangle \leftarrow |S_1\rangle + |S_2\rangle$ 
22 end
23 get  $|T_2\rangle$  in the same way
24 compute  $x \leftarrow Sim(|T_1\rangle, |T_2\rangle)$ 
25 determine whether the sentence embedding needs to be optimized
26 return  $x$ 

```

5. Experimental Results

5.1. Comparing with Some Unsupervised Methods

The Spearman's rank correlation and Pearson correlation are used to compare the experimental results, as shown in Tables 3 and 4. Table 3 shows that the Srcs of STS'14, STS'15 and STSb are significantly higher than the comparison models, but the average Src of our model is higher than all the comparison models. The results in Table 4 show that, except for 4 corpora, the Pcc of all corpora has increased. The average Pccs of STS'12, STS'14 and STS'15 have all been significantly improved, especially the maximum improvement rate of the provided model compared to ACVT on STS'12 has reached 29%. The growth rates of the top three are 20.3%, 9.8% and 7.2% for STS'12.MSRpar, STS'14.tweet-news and STS'12.SMTeuoparl, respectively. Moreover, STS'12.SMTeuoparl achieves a growth rate of 42.3%. Comparing all the Pccs of the 16 corpora, The Pccs of 3 corpora are above 0.9, of 3 corpora are below 0.8, and only of 1 corpus is below 0.7. Though the Pcc of STS'14.deft-forum is less than 0.7, it is also increased slightly, with an increase of 0.02.

Table 3. Comparison of the Spearman’s rank correlation (Src) in each dataset.

Model	STS’12	STS’14	STS’15	STSb	Avg.
SimCSE-base [64]	0.702	0.732	0.814	0.802	0.763
IS-Bert-NLI [65]	0.568	0.630	0.752	0.692	0.661
Bert-flow [66]	0.652	0.694	0.749	0.723	0.705
DINO [67]	0.703	0.713	0.805	0.778	0.750
SBERT-base [68]	0.710	0.732	0.791	0.770	0.751
Provided model	0.641	0.774	0.844	0.830	0.772

Table 4. Comparison of the Pearson correlation coefficient (Pcc) in each dataset.

Dataset	ACVT [69]	SCBOW-att [70]	PP-att [70]	No Tree-Based [4]	Provided Model
12’MSRpar	0.58	0.58	0.50	0.59	0.71
12’MSRvid	0.83	0.83	0.85	0.90	0.92
12’SMTeuroparl	0.43	0.52	0.52	0.69	0.74
12’OnWN	0.70	0.73	0.73	0.84	0.82
12’SMTnews	0.54	0.66	0.67	0.78	0.81
STS’12	0.62	0.66	0.65	0.76	0.80
14’deft-forum	0.48	0.54	0.56	0.66	0.68
14’deft-news	0.74	0.74	0.76	0.78	0.75
14’headlines	0.72	0.72	0.72	0.81	0.82
14’images	0.81	0.81	0.83	0.87	0.87
14’OnWN	0.87	0.87	0.85	0.92	0.93
14’tweet-news	0.75	0.82	0.79	0.82	0.90
STS’14	0.73	0.75	0.75	0.81	0.82
15’answers-forums	0.69	0.69	0.69	0.86	0.88
15’answers-students	0.79	0.79	0.79	0.86	0.89
15’belief	0.70	0.78	0.78	0.87	0.88
15’images	0.82	0.84	0.85	0.85	0.86
15’headlines	0.79	0.79	0.77	0.89	0.86
STS’15	0.76	0.78	0.78	0.86	0.87

To compare the semantic influence of the input word embedding, the entanglement coefficients of word2vec and fasttext for each corpus in Table 5 are set to the same value, that is, the influence of the syntax tree is mainly considered. For easy comparison, the entanglement coefficients $p_{i,j}$ of ‘word2vec’ and ‘fasttext’ are set to the same values for each corpus in the table; namely, the influence of the long-range dependency relation is mainly considered. Comparison of the experimental results in Table 5 shows that all of the MSEs of ‘word2vec’ are smaller than those of ‘fasttext’. There are 5 datasets with a difference of less than 0.01 and 5 corpora with an MSE with a multiplied value greater than 2. The main reason for this result is that all of the MSEs are adjusted under the parameters of ‘word2vec’, which are not the optimal parameters of ‘fasttext’. Additionally, the maximum value of ‘word2vec’ is less than 0.055, and that of ‘fasttext’ is less than 0.09.

Table 5. The semantic influence of the input word embedding.

		STS’12					STS’14					
		MSRpar	MSRvid	SMTeu	OnWN	SMTnews	deft-f	deft-n	headlines	images	OnWN	tweet-n
word2vec	Pcc	0.71	0.91	0.74	0.82	0.81	0.68	0.75	0.82	0.87	0.92	0.89
word2vec	MSE	0.022	0.017	0.047	0.020	0.022	0.053	0.033	0.029	0.022	0.022	0.015
fasttext	Pcc	0.51	0.84	0.42	0.74	0.60	0.51	0.63	0.76	0.76	0.79	0.86
fasttext	MSE	0.028	0.035	0.086	0.049	0.073	0.055	0.039	0.035	0.040	0.062	0.028
		STS’15					STSb					
		answ-for	answ-stu	belief	images	headlines						
word2vec	Pcc	0.88	0.89	0.88	0.86	0.86	0.86					
word2vec	MSE	0.016	0.017	0.023	0.033	0.030	0.028					
fasttext	Pcc	0.78	0.76	0.78	0.81	0.81	0.78					
fasttext	MSE	0.029	0.041	0.037	0.037	0.044	0.037					

5.2. Influence of PoS Combination Weight $T_{I,J}$

We mainly discuss the influence of notional words on sentence semantics including noun, verb, adjective and adverb. The four PoS types of notional words are combined in pairs to form 16 different PoS combinations. The weights of other PoS combinations are set to 0.5.

In Table 6, $t_{i,j}$ is set to different values to illustrate the influence of the parameters of the combination of PoS for different word entanglement. The word ‘first’ denotes the number combination of ‘0.5, 0.2, 0.7, 0.9, 0.3, 0.1, 0.5, 0.9, 0.6, 0.1, 0.6, 0.9, 0.5, 0.9, 0.9, 0.1’, for which the numbers are scattered around 0.5. The word ‘second’ denotes the number combination of ‘0.3, 0.5, 0.3, 0.6, 0.5, 0.5, 0.4, 0.3, 0.5, 0.6, 0.4, 0.3, 0.5, 0.4, 0.5, 0.4’, for which the numbers are concentrated around 0.5. $t_{i,j} = 0.5$ means that all sixteen weights of the PoS combinations are set to 0.5. Comparing all of the Pccs, there are 10 out of 16 corpora achieving the maximum for $t_{i,j} = second$, with the greatest improvement by 0.037 from STS’15.headlines. Additionally, the top three advances are derived from $t_{i,j} = second$, with an improvement of 0.037 from STS’15.headlines, an increase of 0.023 from STS’12.MSRpar and an increase of 0.012 from STS’14.headlines. Comparing the three different MSEs of the same corpus in Table 6, the ratio of the minimums among the three PoS parameters combinations is 6:5:6, approximately evenly distributed. However, there are 8 out of 16 corpora reaching the maximum from $t_{i,j} = first$, and the other corpora average from $t_{i,j} = 0.5$ and $t_{i,j} = second$. In brief, the Pcc and MSE are clearly affected by the distribution of the PoS combination weights. A more discrete parameter distribution corresponds to greater influence on the Pcc and MSE.

The detailed influence of $t_{i,j}$ on STS’12.OnWN is shown in Table 7. When D_2 changes, the trends of the Pcc and MSE with $t_{i,j}$ are still consistent. A higher concentration of $t_{i,j}$ corresponds to a larger Pcc and smaller MSE, as shown in Figure 3. Comparison of the Pccs in Figure 3a and the MSEs in Figure 3b shows that for $t_{i,j} = 0.5$, the Pcc reaches the maximum and the MSE reaches the minimum. This result can be explained based on the entanglement coefficient $p_{i,j} = Sim(|w_i\rangle, |w_j\rangle) \times t_{i,j} \times d_{i,j}$. When the entangled words $|w_i\rangle$ and $|w_j\rangle$ are determined, the cosine similarity of the two words is a fixed value. For the same input sentence, the dependency tree of the sentence is a certain structure; that is, the tree depth difference Δh between the two words is a constant value so that the difference weight $d_{i,j}$ is a certain value. Hence, when the input sentence is known, $p_{i,j}$ is only changed by $t_{i,j}$. The value of $t_{i,j}$ has a strong effect. For example, when the two weights are 0.1 and 0.9, the weight of the PoS combination corresponding to 0.9 is nine times that of the PoS combination corresponding to 0.1. To compare the influence of other parameters, $t_{i,j}$ is set to 0.5 for the experiments discussed below.

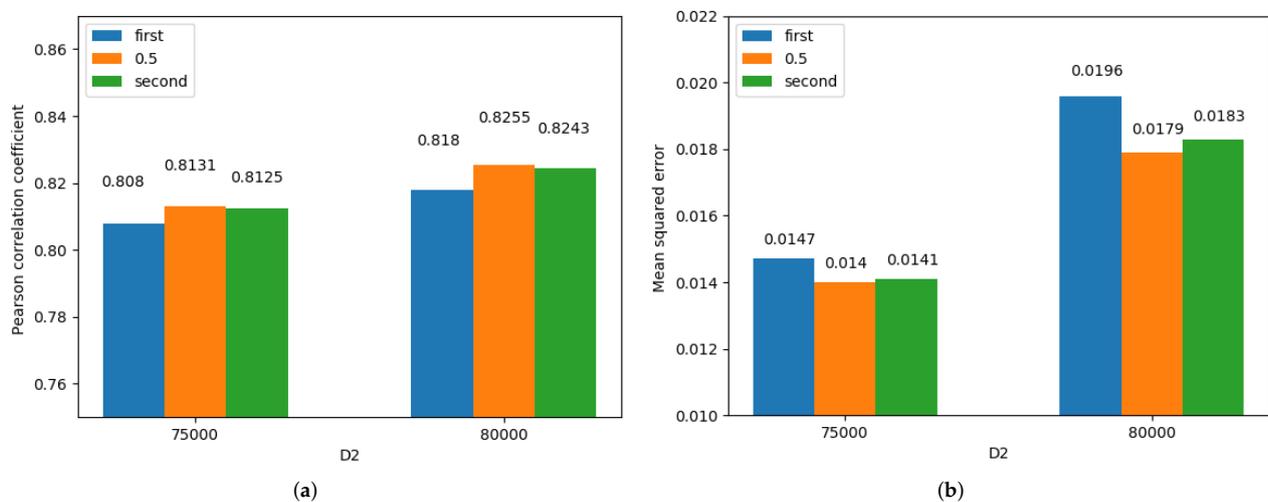


Figure 3. Detailed influence of $t_{i,j}$ on STS’12.OnWN. (a) Pearson correlation coefficient, (b) Mean squared error.

Table 6. Influence of parameters $t_{i,j}$ on sentence semantics of different corpora. The figures in bold-type refer to the maximum Pearson correlation coefficient of each corpus. The figures in red-type refer to the minimum mean squared error of each corpus.

Year	Dataset	Pcc			MSE		
		$t_{i,j} = 0.5$	$t_{i,j} = second$	$t_{i,j} = first$	$t_{i,j} = 0.5$	$t_{i,j} = second$	$t_{i,j} = first$
2012	MSRpar	0.688	0.711	0.703	0.0248	0.0224	0.0235
	MSRvid	0.882	0.881	0.872	0.0248	0.0250	0.0270
	SMTeuoparl	0.728	0.732	0.732	0.0987	0.1013	0.1023
	OnWN	0.825	0.824	0.818	0.0179	0.0183	0.0196
	SMTnews	0.773	0.779	0.770	0.0195	0.0200	0.0198
2014	deft-forum	0.654	0.658	0.665	0.0515	0.0519	0.0432
	deft-news	0.750	0.747	0.752	0.0328	0.0337	0.0326
	headlines	0.805	0.806	0.794	0.0294	0.0294	0.0320
	images	0.872	0.876	0.873	0.0225	0.0215	0.0222
	OnWN	0.918	0.919	0.916	0.0257	0.0255	0.0262
	tweet-news	0.906	0.909	0.909	0.0161	0.0154	0.0151
2015	answers-forums	0.891	0.891	0.892	0.0214	0.0211	0.0209
	answers-students	0.842	0.843	0.843	0.0274	0.0275	0.0272
	belief	0.888	0.886	0.885	0.0233	0.0234	0.0232
	images	0.858	0.858	0.857	0.0326	0.0327	0.0328
	headlines	0.909	0.910	0.863	0.0200	0.0199	0.0301

Table 7. Influence of parameters $t_{i,j}$ on sentence semantics of STS'12.surprise.OnWN.txt with the variables of $y_1 = 0.7, 0.3 < y_2 < 0.7, \nabla E_1 = \nabla E_2 = 0.2$ and $D_1 = 10,000$.

$t_{i,j}$	D_2	Pcc	MSE
<i>first</i>	$D_2 = 75,000$	0.80803	0.01465
<i>first</i>	$D_2 = 80,000$	0.81801	0.01962
0.5	$D_2 = 75,000$	0.81312	0.01404
0.5	$D_2 = 80,000$	0.82549	0.01792
<i>second</i>	$D_2 = 75,000$	0.81251	0.01413
<i>second</i>	$D_2 = 80,000$	0.82427	0.01833

5.3. Influence of the Tree Depth Difference Δh

5.3.1. On STS-Benchmark

The influence of the tree depth difference Δh on STS-benchmark is listed in Table 8. The list named ‘ Δh changes’ represents the condition, and the list labeled ‘ $d_{i,j}$ stays the same’ describes the values under this condition. For example, the condition combination of ‘(1.5, 1.2, 0.8)’ and ‘($\Delta h = 0, 0 < \Delta h \leq 2, \Delta h > 2$)’ denotes that when $\Delta h = 0, d_{i,j} = 1.5$; when $0 < \Delta h \leq 2, d_{i,j} = 1.2$; and when $\Delta h > 2, d_{i,j} = 0.8$. The condition $\Delta h = 0$ that denotes the tree depth difference between the two words is zero; that is, the two words have a direct modified relationship. When Δh changes, the Pcc and MSE show very small changes. The main reason for this result is that the corpus consists mostly of short sentences, with sentence lengths of less than 10. A shorter sentence tends to have a simpler parser tree structure.

Table 8. Influence of parameters p_h and Δh on sentence semantics of STS-benchmark with the variables of $y_1 = 0.7, 0.25 < y_2 < 0.7, \nabla E_1 = \nabla E_2 = 0.25, D_1 = 10,000$ and $D_2 = 75,000$.

$d_{i,j}$ Stays the Same	Δh Changes	Pcc	MSE
(1.5, 1.2, 0.8)	($\Delta h = 0, 0 < \Delta h \leq 2, \Delta h > 2$)	0.85837	0.02869
	($\Delta h = 0, 0 < \Delta h \leq 3, \Delta h > 3$)	0.85829	0.02868
	($\Delta h = 1, 1 < \Delta h \leq 3, \Delta h > 3$)	0.85812	0.02868

5.3.2. On STS'14.deft-News

The influences of the tree depth difference Δh and the tree depth difference weight $d_{i,j}$ on STS'14.deft-news are illustrated in Table 9. In the STS'14.deft-news corpus, there are

some long-sentences with more than 20 words. The complexity of the sentence structure tends to increase with the sentence length.

Table 9. Influence of parameters p_h and Δh on sentence semantics of STS’14.deft-news.txt with the variables of $y_1 = 0.7, 0.25 < y_2 < 0.7, \nabla E_1 = \nabla E_2 = 0.2, D_1 = 10,000$ and $D_2 = 85,000$.

$d_{i,j}$ Stays the Same	Δh Changes	Pcc	MSE
(1.5, 1.2, 0.8)	($\Delta h = 0, 0 < \Delta h \leq 2, \Delta h > 2$)	0.75023	0.03280
	($\Delta h = 0, 0 < \Delta h \leq 4, \Delta h > 4$)	0.74806	0.03296
	($\Delta h = 0, 0 < \Delta h \leq 3, \Delta h > 3$)	0.74431	0.03340
	($\Delta h \leq 1, 1 < \Delta h \leq 3, \Delta h > 3$)	0.74687	0.03313
	($\Delta h \leq 1, 1 < \Delta h \leq 4, \Delta h > 4$)	0.74554	0.03312
	($\Delta h \leq 1, 1 < \Delta h \leq 5, \Delta h > 5$)	0.74560	0.03310
	($\Delta h \leq 1, 1 < \Delta h \leq 6, \Delta h > 6$)	0.74532	0.03313
	($\Delta h \leq 1, 1 < \Delta h \leq 7, \Delta h > 7$)	0.74533	0.03313
	($\Delta h \leq 1, 1 < \Delta h \leq 8, \Delta h > 8$)	0.74540	0.03312
	($\Delta h \leq 1, 1 < \Delta h \leq 9, \Delta h > 9$)	0.74682	0.03302
Δh Stays the Same	$d_{i,j}$ Changes	Pcc	MSE
($\Delta h = 0, 0 < \Delta h \leq 2, \Delta h > 2$)	(1.5, 1.2, 0.6)	0.75083	0.03276
	(2.0, 1.2, 0.6)	0.75083	0.03276
($\Delta h = 0, 0 < \Delta h \leq 2, \Delta h > 2$)	(2.0, 1.5, 1.0)	0.74690	0.03292
	(2.0, 1.5, 0.8)	0.75023	0.03280
($\Delta h = 0, 0 < \Delta h \leq 2, \Delta h > 2$)	(1.5, 1.2, 0.6)	0.75083	0.03276
	(1.5, 1.5, 0.6)	0.75008	0.03285
($\Delta h = 0, 0 < \Delta h \leq 2, \Delta h > 2$)	(2.0, 1.5, 1.0)	0.74690	0.03292
	(1.5, 1.2, 0.6)	0.75083	0.03276

We use the variable-controlling approach to discuss the impact of Δh and $d_{i,j}$ on semantics from two aspects, as shown in the upper and lower parts of Table 9. The upper half shows the semantic changes when Δh changes. The detailed influence of the tree depth difference Δh between the entangled words with indirect modified relationship on sentence semantics is illustrated in Figure 4. As shown in Figure 4a, when $2 \leq \Delta h < 4$, the Pcc decreases markedly, whereas for $\Delta h \geq 4$, the change curve of the Pcc is almost horizontal. Examination of Figure 4b shows that the MSE changes only slightly. The lower part of Table 9 shows the influence of the weight $d_{i,j}$ of the direct modified relationship and that of the indirect modified relationship on sentence semantics. The values in blue are the values of the variable changes, and the values in red indicate the maximum Pcc and the minimum MSE. When the weight of direct modified relationship is altered, the Pcc and MSE remain unchanged. However, comparing the second group and the third group, it is found that when $d_{i,j}$ increases, the Pcc decreases and MSE increases, implying that entangled words modified indirectly will produce semantic errors. Additionally, when all three weights increase, the Pcc decreases and the MSE increases. Hence, when two adjacent words are entangled together by the tensor, if their modified relationship is indirect, semantic errors will be introduced.

Comparison of the results presented in Tables 8 and 9 shows that the influence on the sentence semantics of the long-range entanglement of words with modified relationship is more apparent than that of the short-range entanglement.

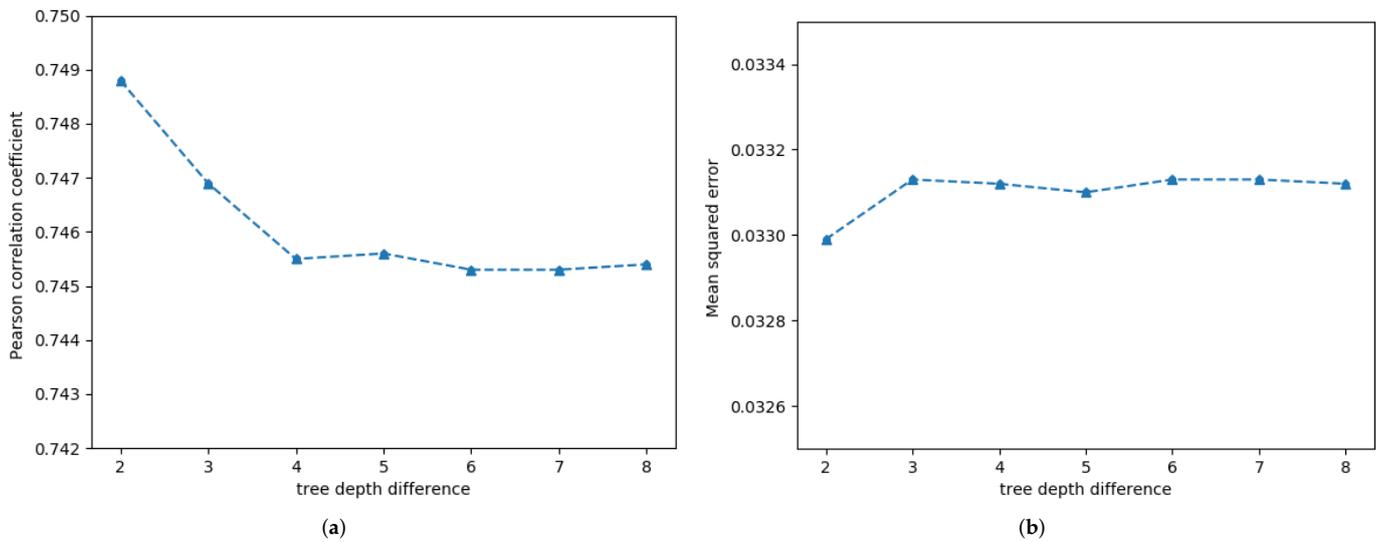


Figure 4. Detailed influence of the distance between the entangled words with indirect modified relationship on sentence semantics in STS'14.left-news. (a) Pearson correlation coefficient, (b) Mean squared error.

5.4. Influence of the Dimension Reduction

5.4.1. Influence on STS'15.Images

The influences of dimension D_1 of the sentence dimensionality reduction and dimension D_2 of the entangled word dimensionality reduction on STS'15.images are illustrated in Figure 5. As D_2 increases, the Pcc first decreases slowly and then fluctuates with smaller variables, as shown by the orange curve in Figure 5a. Excluding $D_1 = 15,000$, the overall change in the MSE is small; as D_1 increases, the MSE gradually increases, as explained by the orange curve in Figure 5b. As D_2 increases, the Pcc first increases and then decreases, and the MSE first decreases and then increases, as shown in Figure 5c and Figure 5d, respectively. When the Pcc reaches the maximum, the MSE is not the minimum, such that $D_2 = 75,000$. When $D_2 = 80,000$, the MSE reaches the minimum, but the Pcc is not at the maximum.

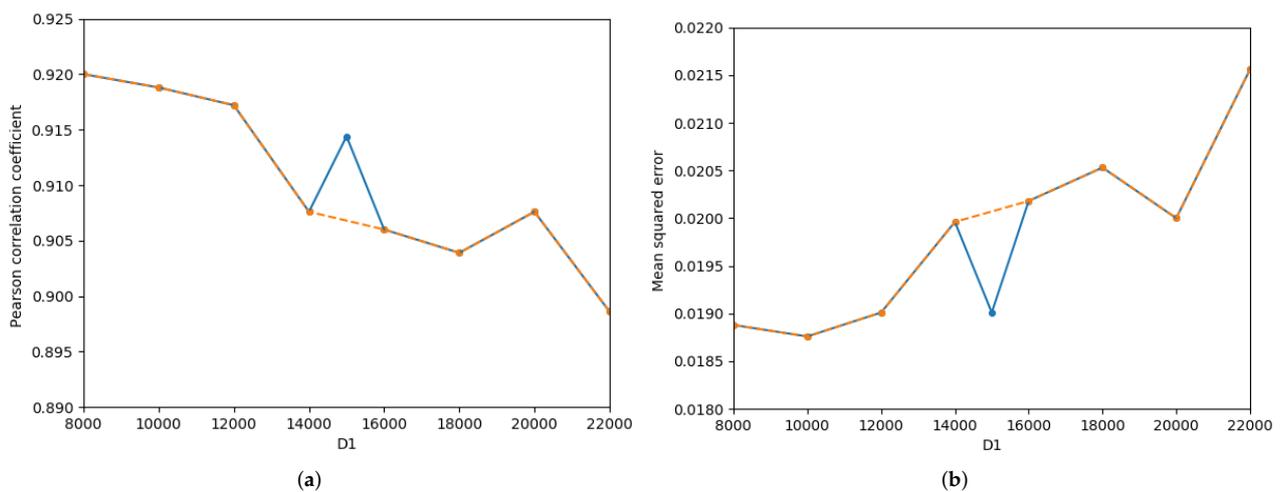


Figure 5. Cont.

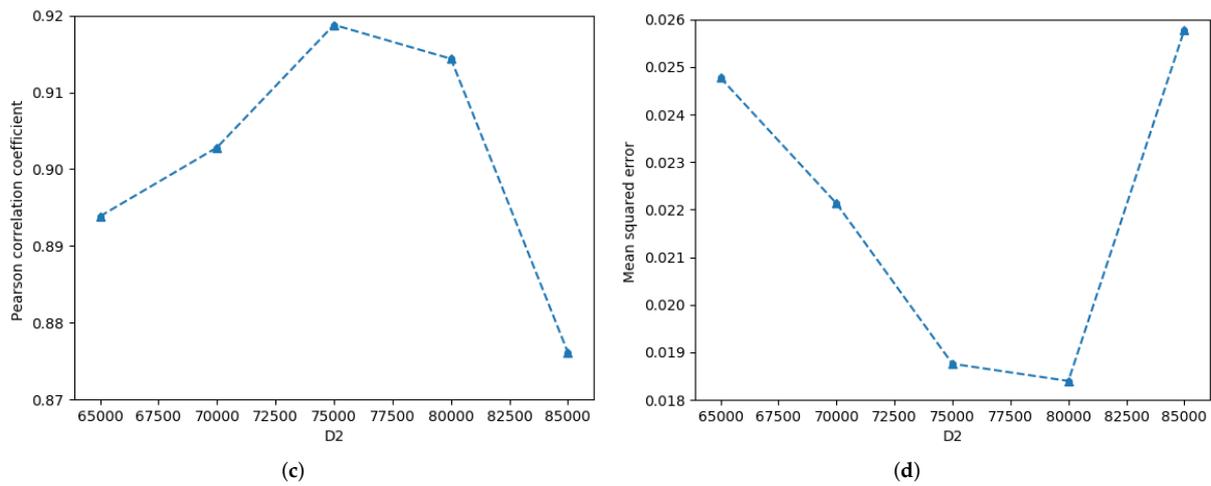


Figure 5. Detailed influence of the D_1 and D_2 on semantics of STS'15.images.txt. (a) Pearson correlation coefficient- D_1 , (b) Mean squared error- D_1 , (c) Pearson correlation coefficient- D_2 , (d) Mean squared error- D_2 .

5.4.2. Influence on STS'15.Headlines

The effects of dimension D_1 of dimensionality reduction at the sentence level and dimension D_2 of the dimensionality reduction at the entangled word level on sentence semantics in STS'15.headlines are described in Figure 6. With the exception of $D_1 = 15,000$, with increasing D_1 , the Pcc first increases and then linearly decreases, while the MSE first decreases and then gradually increases, as shown by the orange curves in Figure 6a,b. However, considering all of the experimental results including $D_1 = 15,000$, the change curves of the Pcc and MSE both show large fluctuations, as presented by the blue curves in Figure 6a,b. Moreover, the Pcc of $D_1 = 15,000$ is higher than all of the Pcc values, and the MSE of $D_1 = 15,000$ achieves the minimum value. Comparisons of Figure 6c,d show that with the increasing of D_2 , the Pcc first increases slowly and then decreases sharply, while the MSE decreases slightly and then increases rapidly.

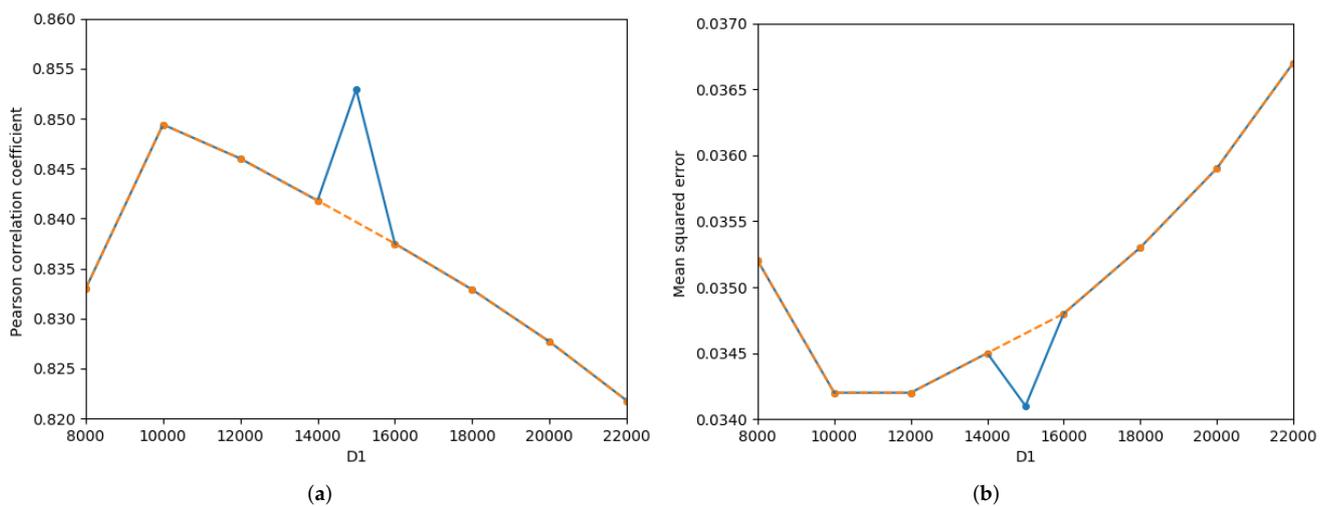


Figure 6. Cont.

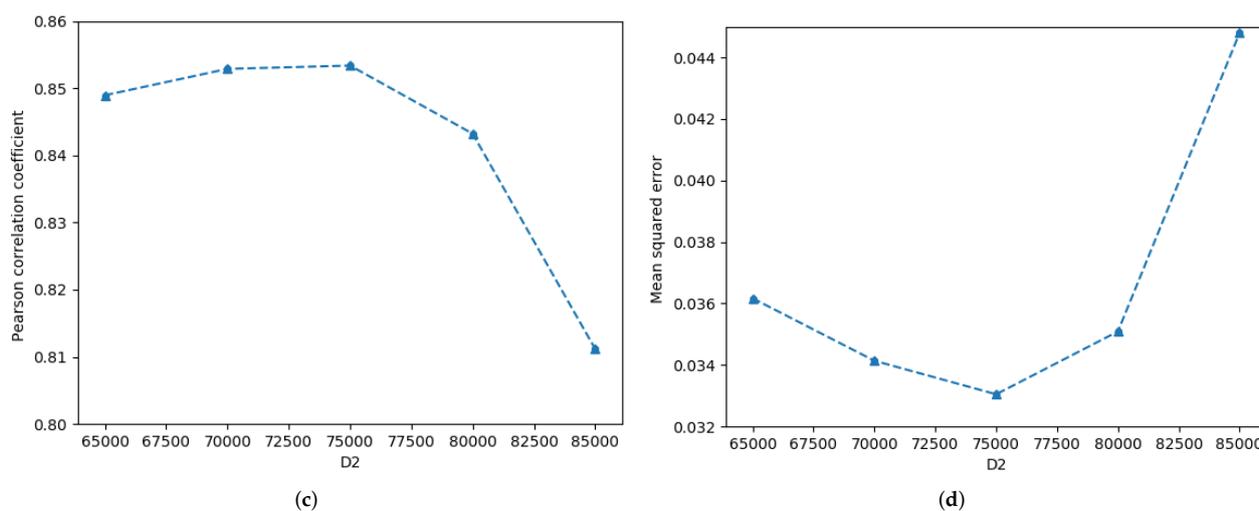


Figure 6. Detailed influence of the D_1 and D_2 on semantics of STS'15.headlines.txt. (a) Pearson correlation coefficient- D_1 , (b) Mean squared error- D_1 , (c) Pearson correlation coefficient- D_2 , (d) Mean squared error- D_2

5.4.3. Summary

Comparisons of the orange curves to the blue curves in Figures 5a,b and 6a,b show that the change trend of the orange curves is smoother than that of the blue curves. The blue curves only consider one more point $D_1 = 15,000$ than the orange curve. Hence, the optimum value is related to the interval variables. A larger interval variable corresponds to greater fluctuations of the the curve, as observed by the comparison of the orange curves and the blue curves of Figures 5a,b and 6a,b. A smaller interval variable makes it more likely that the optimal value will be obtained. This result shows that the optimal solution obtained when the interval variable is large may not even be the local optimal value.

6. Conclusions and Future Works

This paper proposes a quantum entangled word representation based on syntax trees to represent sentences. When the sentence structure is complex, the two words that have a direct modified relationship are not necessarily in close proximity. Introducing quantum entanglement between words that have long-range dependencies enables remote words to also directly establish modified relationships. Combining the attention mechanism based on the dependency tree with the quantum entanglement coefficient, the entanglement coefficient between words is related not only to the PoS combination of the two words and the distribution of the two words in the dictionary but also to the modified relationship between the words. Utilizing the dependency trees of sentences to establish long-distance connections between words only considering the entangled words reduces semantic errors. Moreover, the use of the dependency tree-based attention weight can reduce the influence of adjacent entangled words without directly modifying the sentence semantics, thereby more accurately expressing the sentence semantics. We also discuss the impact of PoS combination, tree depth difference, and dimensionality reduction of entangled words on the sentence semantics. As the maximum of Pcc and the minimum value of MSE obtained here are not the optimal solutions, and may not even be the local optimal solutions, in future works, we mainly consider how to obtain the optimal value easily and effectively by introducing the theory of convex optimization to generalize this model. In this model, the semantic expansion of associated words is obtained by the tensor product of word vectors of the two related words, which is not applicable to the short sentences consisting of only one content word.

Author Contributions: Y.Y. conceptualization and project administration; D.Q., data curation and formal analysis; R.Y., software and validation. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant no. 12171065 and 11671001) and the Doctor Training Program of Chongqing University of Posts and Telecommunications, China (Grant no. BYJS201915).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, Y.; Song, D.; Li, X.; Zhang, P.; Wang, P.; Rong, L.; Yu, G.; Wang, B. A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Inf. Fusion* **2020**, *62*, 14–31. [[CrossRef](#)]
2. Zhang, P.; Niu, J.; Su, Z.; Wang, B.; Ma, L.; Song, D. End-to-end quantum-like language models with application to question answering. In Proceedings of the 32nd Conference on Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, Baltimore, MD, USA, 9–11 November 2018; pp. 5666–5673.
3. Zhang, Y.; Song, D.; Zhang, P.; Li, X.; Wang, P. A quantum-inspired sentiment representation model for twitter sentiment analysis. *Appl. Intell.* **2018**, *49*, 3093–3108. [[CrossRef](#)]
4. Yu, Y.; Qiu, D.; Yan, R. A quantum entanglement-based approach for computing sentence similarity. *IEEE Access* **2020**, *8*, 174265–174278. [[CrossRef](#)]
5. Yu, Y.; Qiu, D.; Yan, R. Quantum entanglement based sentence similarity computation. In Proceedings of the 2020 IEEE International Conference on Progress in Informatics and Computing (PIC2020), Online, 18–20 December 2020; pp. 250–257.
6. Zhang, Y.; Wang, Y.; Yang, J. Lattice LSTM for chinese sentence representation. *IEEE ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1506–1519. [[CrossRef](#)]
7. Liu, D.; Fu, J.; Qu, Q.; Lv, J. BFGAN: Backward and forward generative adversarial networks for lexically constrained sentence generation. *IEEE ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2350–2361. [[CrossRef](#)]
8. Wang, B.; Kuo, C. SBERT-WK: A sentence embedding method by dissecting bert-based word models. *IEEE ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2146–2157. [[CrossRef](#)]
9. Hosseinalipour, A.; Gharehchopogh, F.; Masdari, M.; Khademi, A. Toward text psychology analysis using social spider optimization algorithm. *Concurr. Comp.-Pract. E* **2021**, *33*, e6325. [[CrossRef](#)]
10. Hosseinalipour, A.; Gharehchopogh, F.; Masdari, M.; Khademi, A. A novel binary farmland fertility algorithm for feature selection in analysis of the text psychology. *Appl. Intell.* **2021**, *51*, 4824–4859. [[CrossRef](#)]
11. Osmani, A.; Mohasefi, J.; Gharehchopogh, F. Enriched latent Dirichlet allocation for sentiment analysis. *Expert Syst.* **2020**, *37*, e12527. [[CrossRef](#)]
12. Huang, X.; Peng, Y.; Wen, Z. Visual-textual hybrid sequence matching for joint reasoning. *IEEE Trans. Cybern.* **2020**, *51*, 5692–5705. [[CrossRef](#)]
13. Dai, D.; Tang, J.; Yu, Z.; Wong, H.; You, J.; Cao, W.; Hu, Y.; Chen, C. An inception convolutional autoencoder model for chinese healthcare question clustering. *IEEE Trans. Cybern.* **2021**, *51*, 2019–2031. [[CrossRef](#)] [[PubMed](#)]
14. Yin, C.; Tang, J.; Xu, Z.; Wang, Y. Memory augmented deep recurrent neural network for video question answering. *IEEE Trans. Neural Netw. Learn Syst.* **2020**, *31*, 3159–3167. [[CrossRef](#)] [[PubMed](#)]
15. Mohammadzadeh, H.; Gharehchopogh, F. A multi-agent system based for solving high-dimensional optimization problems: A case study on email spam detection. *Int. J. Commun. Syst.* **2021**, *34*, e4670. [[CrossRef](#)]
16. Li, X.; Jiang, H.; Kamei, Y.; Chen, X. Bridging semantic gaps between natural languages and apis with word embedding. *IEEE Trans. Softw. Eng.* **2020**, *46*, 1081–1097. [[CrossRef](#)]
17. Osmani, A.; Mohasefi, J.; Gharehchopogh, F. Sentiment classification using two effective optimization methods derived from the artificial bee colony optimization and imperialist competitive algorithm. *Comput. J.* **2022**, *65*, 18–66. [[CrossRef](#)]
18. Li, L.; Jiang, Y. Integrating language model and reading control gate in BLSTM-CRF for biomedical named entity recognition. *IEEE ACM Trans. Comput. Biol. Bioinform.* **2020**, *17*, 841–846. [[CrossRef](#)] [[PubMed](#)]
19. Khataei, M.H.; Gharehchopogh, F.; Majidzadeh, K.; Sangar, A. A new hybrid based on long Short-term memory network with spotted Hyena optimization algorithm for multi-label text classification. *Mathematics* **2022**, *10*, 488. [[CrossRef](#)]
20. Choi, H.; Lee, H. Multitask learning approach for understanding the relationship between two sentences. *Inf. Sci.* **2019**, *485*, 413–426. [[CrossRef](#)]
21. Zhang, L.; Luo, M.; Liu, J.; Chang, X.; Yang, Y.; Hauptmann, A. Deep top-k ranking for image-sentence matching. *IEEE Trans. Multimed.* **2020**, *22*, 775–785. [[CrossRef](#)]

22. Huang, F.; Zhang, X.; Zhao, Z.; Li, Z. Bidirectional spatial-semantic attention networks for image-text matching. *IEEE Trans. Image Process.* **2019**, *28*, 2008–2020. [[CrossRef](#)]
23. Ma, Q.; Yu, L.; Tian, S.; Chen, E.; Ng, W. Global-local mutual attention model for text classification. *IEEE ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2127–2139. [[CrossRef](#)]
24. Xu, X.; Wang, T.; Yang, Y.; Zuo, L.; Shen, F.; Shen, H. Cross-modal attention with semantic consistence for image-text matching. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 5412–5425. [[CrossRef](#)]
25. Vaswani, A.; Shazeer, N.; Parmar, N. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
26. Guo, Q.; Qiu, X.; Xue, X.; Zhang, Z. Low-rank and locality constrained self-attention for sequence modeling. *IEEE ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2213–2222. [[CrossRef](#)]
27. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q. Xlnet: Generalized autoregressive pretraining for language understanding. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; pp. 5754–5764.
28. Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; Hon, H. Unified language model pre-training for natural language understanding and generation. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; pp. 13042–13054.
29. Bao, H.; Dong, L.; Wei, F.; Wang, W.; Yang, N.; Liu, X.; Wang, Y.; Gao, J.; Piao, S.; Zhou, M.; Hon, H. Unilmv2: Pseudo-masked language models for unified language model pre-training. In Proceedings of the 37th International Conference on Machine Learning, (ICML 2020), Online, 13–18 July 2020; pp. 642–652.
30. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACLHLT 2019), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
31. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A lite BERT for self-supervised learning of language representations. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, 26–30 April 2020; pp. 1–16.
32. Conneau, A.; Lample, G. Cross-lingual language model pretraining. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; pp. 7057–7067.
33. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018), New Orleans, LA, USA, 1–6 June 2018; pp. 2227–2237.
34. Gharehchopogh, F. Advances in tree seed algorithm: A comprehensive survey. *Arch. Comput. Methods Eng.* **2022**, 1–24. [[CrossRef](#)]
35. Wang, J.; Yu, L.; Lai, K.; Zhang, X. Treestructured regional CNN-LSTM model for dimensional sentiment analysis. *IEEE ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 581–591. [[CrossRef](#)]
36. Shen, M.; Kawahara, D.; Kurohashi, S. Dependency parser reranking with rich subtree features. *IEEE ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1208–1218. [[CrossRef](#)]
37. Luo, H.; Li, T.; Liu, B.; Wang, B.; Unger, H. Improving aspect term extraction with bidirectional dependency tree representation. *IEEE ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1201–1212. [[CrossRef](#)]
38. Zhang, J.; Zhai, F.; Zong, C. Syntax-based translation with bilingually lexicalized synchronous tree substitution grammars. *IEEE Trans. Speech Audio Process.* **2013**, *21*, 1586–1597. [[CrossRef](#)]
39. Chen, W.; Zhang, M.; Zhang, Y. Distributed feature representations for dependency parsing. *IEEE ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 451–460. [[CrossRef](#)]
40. Geng, Z.; Chen, G.; Han, Y.; Lu, G.; Li, F. Semantic relation extraction using sequential and treestructured LSTM with attention. *Inf. Sci.* **2020**, *509*, 183–192. [[CrossRef](#)]
41. Fei, H.; Ren, Y.; Ji, D. A tree-based neural network model for biomedical event trigger detection. *Inf. Sci.* **2020**, *512*, 175–185. [[CrossRef](#)]
42. Cao, Q.; Liang, X.; Li, B.; Lin, L. Interpretable visual question answering by reasoning on dependency trees. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 887–901. [[CrossRef](#)] [[PubMed](#)]
43. Wu, Y.; Zhao, S.; Li, W. Phrase2vec: Phrase embedding based on parsing. *Inf. Sci.* **2020**, *517*, 100–127. [[CrossRef](#)]
44. Widdows, D.; Cohen, T. Graded semantic vectors: An approach to representing graded quantities in generalized quantum models. In Proceedings of the Quantum Interaction—9th International Conference (QI 2015), Filzbach, Switzerland, 15–17 July 2015; Volume 9535, pp. 231–244.
45. Aerts, D.; Sozzo, S. Entanglement of conceptual entities in quantum model theory (qmod). In Proceedings of the Quantum Interaction—6th International Symposium (QI 2012), Paris, France, 27–29 June 2012; Volume 7620, pp. 114–125.
46. Nguyen, N.; Behrman, E.; Moustafa, M.; Steck, J. Benchmarking neural networks for quantum computations. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 2522–2531. [[CrossRef](#)]

47. Sordoni, A.; Nie, J.; Bengio, Y. Modeling term dependencies with quantum language models for IR. In Proceedings of the 36th International ACM SIGIR conference on research and development in Information Retrieval (SIGIR'13), Dublin, Ireland, 28 July–1 August 2013; pp. 653–662.
48. Cohen, T.; Widdows, D. Embedding probabilities in predication space with hermitian holographic reduced representations. In Proceedings of the Quantum Interaction—9th International Conference (QI 2015), Filzbach, Switzerland, 15–17 July 2015; Volume 9535, pp. 245–257.
49. Yuan, K.; Xu, W.; Li, W.; Ding, W. An incremental learning mechanism for object classification based on progressive fuzzy three-way concept. *Inf. Sci.* **2022**, *584*, 127–147. [[CrossRef](#)]
50. Xu, W.; Yuan, K.; Li, W. Dynamic updating approximations of local generalized multigranulation neighborhood rough set. *Appl. Intell.* **2022**. [[CrossRef](#)]
51. Xu, W.; Yu, J. A novel approach to information fusion in multi-source datasets: A granular computing viewpoint. *Inf. Sci.* **2017**, *378*, 410–423. [[CrossRef](#)]
52. Xu, W.; Li, W. Granular computing approach to two-way learning based on formal concept analysis in fuzzy datasets. *IEEE Trans. Cybern.* **2016**, *46*, 366–379. [[CrossRef](#)]
53. Hou, Y.; Zhao, X.; Song, D.; Li, W. Mining pure high-order word associations via information geometry for information retrieval. *ACM Trans. Inf. Syst.* **2013**, *31*, 1–12. [[CrossRef](#)]
54. Xie, M.; Hou, Y.; Zhang, P.; Li, J.; Li, W.; Song, D. Modeling quantum entanglements in quantum language models. In Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015), Buenos Aires, 25–31 July 2015; pp. 1362–1368.
55. Aerts, D.; Beltran, L.; Bianchi, M.; Sozzo, S.; Veloz, T. Quantum cognition beyond hilbert space: Fundamentals and applications. In Proceedings of the Quantum Interaction—10th International Conference (QI 2016), San Francisco, CA, USA, 20–22 July 2016; Volume 10106, pp. 81–98.
56. Zhang, Y.; Song, D.; Zhang, P.; Wang, P.; Li, J.; Li, X.; Wang, B. A quantum-inspired multimodal sentiment analysis framework. *Theor. Comput. Sci.* **2018**, *752*, 21–40. [[CrossRef](#)]
57. Zhang, Y.; Li, Q.; Song, D.; Zhang, P.; Wang, P. Quantum-inspired interactive networks for conversational sentiment analysis. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019), Macao, China, 10–16 August 2019; pp. 5436–5442.
58. Aerts, D.; Arguelles, J.; Beltran, L.; Distrito, I.; Bianchi, M.; Sozzo, S.; Veloz, T. Context and interference effects in the combinations of natural concepts. In Proceedings of the Modeling and Using Context—10th International and Interdisciplinary Conference (CONTEXT 2017), Paris, France, 20–23 July 2017; Volume 10257, pp. 677–690.
59. Galofaro, F.; Toffano, Z.; Doan, B. A quantum-based semiotic model for textual semantics. *Kybernetes* **2018**, *47*, 307–320. [[CrossRef](#)]
60. Agirre, E.; Cer, D.; Diab, M.; Gonzalez-Agirre, A. Semeval-2012 task 6: A pilot on semantic textual similarity. In Proceedings of the 6th International Workshop on Semantic Evaluation, Montreal, QC, Canada, 7–8 June 2012; pp. 385–393.
61. Agirre, E.; Banea, C.; Cardie, C.; Cer, D.; Diab, M.T.; Gonzalez-Agirre, A.; Guo, W.; Mihalcea, R.; Rigau, G.; Wiebe, J. Semeval-2014 task 10: Multilingual semantic textual similarity. In Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland, 23–24 August 2014; pp. 81–91.
62. Agirre, E.; Banea, C.; Cardie, C.; Cer, D.; Diab, M.; Gonzalez-Agirre, A.; Guo, W.; Lopez-Gazpio, I.; Maritxalar, M.; Mihalcea, R.; et al. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In Proceedings of the 9th International Workshop on Semantic Evaluation, Denver, CO, USA, 4–5 June 2015; pp. 252–263.
63. Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; Specia, L. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation, Vancouver, BC, Canada, 3–4 August 2017; pp. 1–14.
64. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), Online, 10–11 November 2021; pp. 6894–6910.
65. Zhang, Y.; He, R.; Liu, Z.; Lim, K.; Bing, L. An unsupervised sentence embedding method by mutual information maximization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), Online, 16–20 November 2020; pp. 1601–1610.
66. Li, B.; Zhou, H.; He, J.; Wang, M.; Yang, Y.; Li, L. On the sentence embeddings from pre-trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), Online, 16–20 November 2020; pp. 9119–9130.
67. Schick, T.; Schütze, H. Generating datasets with pretrained language models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), Online, 10–11 November 2021; pp. 6943–6951.
68. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), Hong Kong, China, 2–7 November 2019; pp. 3982–3992.
69. Quan, Z.; Wang, Z.; Le, Y. An efficient framework for sentence similarity modeling. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 853–865. [[CrossRef](#)]
70. Wang, S.; Zhang, J.; Zong, C. Learning sentence representation with guidance of human attention. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI2017), Melbourne, Australia, 19–25 August 2017; pp. 4137–4143.