IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 6, 2020, accepted September 18, 2020, date of publication September 23, 2020,
date of current version October 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3025958

A Quantum Entanglement-Based Approach for Computing Sentence Similarity

YAN YU[1,2], DONG QIU[1], AND RUITENG YAN[1]
[1]School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2]College of Mobile Telecommunications, Chongqing University of Posts and Telecom, Chongqing 401520, China

Corresponding author: Dong Qiu (dongqiumath@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 11671001, and in part by the Doctor Training Program, Chongqing University of Posts and Telecommunications, China, under Grant BYJS201915.

ABSTRACT It is important to learn directly from original texts in natural language processing (NLP). Many deep learning (DP) models needing a large number of manually annotated data are not effective in deriving much information from corpora with few annotated labels. Existing methods using unlabeled language information to provide valuable messages consume considerable time and cost. Our provided sentence representation based on quantum computation (called Model I) needs no prior knowledge except word2vec. To reduce some semantic noise caused by the tensor product on the entangled words vector, two improved models (called Model II and Model III) are proposed to reduce the dimensions of the sentence embedding stimulated by Model I. The provided models are evaluated in the STS tasks of 2012, 2014, 2015 and 2016, for a total of 21 corpora. Experimental results show that using quantum entanglement and dimensionality reduction in sentence embedding yields state-of-the-art performances on semantic relations and syntactic structures. Compared to the Pearson correlation coefficient (Pcc) and mean squared error (MSE), the results of 16 out of 16 corpora are better than the results of the comparative methods.

INDEX TERMS Quantum computation, text representation, sentence similarity, tensor product, dimensionality reduction.

I. INTRODUCTION

Semantic textual similarity (STS) is a task that measures the degree of the semantic similarity between two sentences. There are many applications in natural language processing (NLP) that refer to the textual similarity in semantics, such as document summarization, semantic search, question answering, document classification, and natural language inference (NLI). The main challenge of STS in recent years is how to mine more semantic information to make the calculational results infinitely close to those of humans. The only criterion for evaluating the computational results is the degree of the approximation to human-made scores. The closer the calculated result is to the human-made score, the more general the model is. With the development and improvement of word embedding, existing text analysis methods are continuously emerging, which is mainly based on word representations. Some methods and the calculation results on semantic analysis based on word vectors are collected in [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

Compared to the mentioned models, two problems should be considered. First, only the semantic information of the word is considered, but the influence between words is ignored. Second, the relationship between words is considered in the methods based on dependency trees but with complex computing processing. To solve the problems mentioned above, our proposed methods consider the influence between words and integrate the theoretical knowledge on quantum entanglement into the textual representation. In the models, we use the tensor product to extend the dimensionality of the sentence representation with more semantic information. Because of the entanglement between the adjoining words, the impact of the continuous synonyms in any sentence can widen the semantic differences in the sentence pair.

For the sake of the influence of continuous synonyms, we provide two approaches of dimensionality reduction to sentence representation. The one approach introduces sentence level improvement on the sentence representation based on quantum entanglement, in that we directly decrease the dimensionality of the sentence embedding. In the other method, we give an entangled words level advancement to

the sentence representation-based quantum entanglement by decreasing some dimensions with relatively small terms of the entangled words vectors; as a consequence, the ultimate dimensionality of the sentence representation is also reduced. Our proposed methods are composed of the three models mentioned above. Experimental results for STS of the years 2012, 2014, 2015 and 2016 on the similarity in sentence pairs demonstrate the high performances of our proposed approaches.

In brief, the innovations are as follows. First, introducing quantum mechanics methods, two continuous notional words are entangled together with the numerical computation method of the tensor product. Taking the entangled words pair as a whole can mine more semantic information. Then, by means of the physical ideas on extracting the primary factors and ignoring the secondary factor, two models of dimensionality reduction are proposed to optimize the model of the sentence representation based on quantum entanglement, which is different from the dimensionality reduction ideas of DP. Last, the experimental results of the proposed models are excellent, and the algorithms are very simple with no need for any prior knowledge except for word2vec.

The paper is organized as follows. Section II summarizes some related literature on quantum computation and sentence similarity. Section III explains our proposed models in detail. The detailed comments on the different combinations of the proposed models are explained in Section IV. Section V demonstrates the experimental results and lists the comparison to other methods. In Section VI, some conclusions are drawn.

II. RELATED WORKS

The main idea of this paper is to improve the sentence representation based on quantum entanglement with dimensionality reduction. In this section, we review some related works on quantum computation and sentence similarity.

A. QUANTUM COMPUTATION

In recent years, the integration of quantum computation with other disciplines has become increasingly popular, such as the application of machine learning in quantum computation [2]–[7], the introduction of quantum theory into artificial intelligence [8]–[11], the application of quantum theory in information science [12]–[17], quantum chemistry [18], and quantum annealing algorithms [19]. In [2], experimental machine learning of quantum states was possible to efficiently learn and classify, which indicates that the classification of quantum states can be achieved with limited resources. J. Venderley *et al.* established a machine-learning-based approach that can enable rapid exploration of large phase spaces [6]. In [14], efficient verification protocols for any stabilizer state were given. C. Guo *et al.* introduced a machine learning model in which matrix product operators were trained to implement sequence-to-sequence prediction to predict the next sequence [20]. In [3], the number of neural network features for machine learning was shown, which could naturally be mapped into the quantum optical domain by introducing the quantum optical neural network. A framework that captures entanglement distillation in the presence of natural correlations arising from memory channels was introduced by Waeldchen *et al.* [21].

There are very few works integrating natural language processing with quantum computation [22]. In [22], P. Zhang *et al.* designed a sentence representation using quantum language for description. The sentence embedding represented by Dirac symbols was input into deep neural networks to compute the similarity of the question answer sentence pair.

B. SENTENCE SIMILARITY

The key to the improvement of many top-level applications is the development of supporting technologies. In the big data era, it is important to advance the accuracy rate of text similarity, as the computation of text similarity is a significant part of NLP. The common techniques on word embedding [23] or sentence embedding [24] include GloVe [25], PSL [26], ST [27], SCBOW [28], PROJ [29], PP-tf-idf [30], DAN [31], LSTM [32], and RNN [29]. The similarity in semantics can be applied to many NLP fields [33]–[35]. Z.-T. Guan *et al.* proposed a cross-lingual multikeyword ranked search scheme based on the open multilingual WordNet with flexible keyword and language preference settings [33]. A. J. M. Traina *et al.* provided technologies and tools to meet the variety and veracity characteristics of big and complex data and consider the semantic information of data [34]. In [35], a method for querying relational databases with keywords to simplify access to these data is proposed. An algorithm using latent Dirichlet allocation (LDA) and OpenAI-GPT to generate negative examples is introduced to multilingual STS [36]. Latent semantic analysis and LDA are compared to identify the unit [37]. Quan *et al.* [1] provide an efficient framework for the sentence similarity merging the attention weight mechanism with a constituency tree and give comparison experimental results to other methods. All the results of these classic methods are collected from [1], as shown in Table 2.

The main methods for computing sentence similarity are text embedding or neural network models [38], [39], [41], [42]. A multitask learning approach for understanding the relationship between two sentences is reported by Choi and Lee [43]. A. Skabar and K. Abdalgader provide an algorithm that is based on fuzzy relations to identify overlapping clusters of semantically related sentences [44]. A text expansion and deep model-based approach for service recommendation is proposed, which can bridge the vocabulary gap between services and user queries with the collective semantic similarity of sentences and descriptions [39]. An interactive self-attentive Siamese neural network is used to verify the effectiveness of the interactive self-attention [40]. With the development of capsule networks, the text representation preprocessed by neural networks can achieve out-of-state results as the input of classification and machine translation. ELMo

transfers the top LSTM layer into a linear combination of the vectors stacked above each input word for each end task, with marked improvement [45]. Due to the powerful pretraining function of the transformer, some new models calculating classification and language inference have achieved state-of-the-art results, such as XLNet [46], BERT and its variants [47]–[50], and UNILM [51], [52]. Minaee *et al.* provided a comprehensive review of deep learning-based text classification [53]. However, few studies have examined semantic similarity [54].

Compared to the works mentioned above, existing computations of sentence similarity mainly focus on the similarity between word representations and do not consider the similarity in semantics. Considering only the similarity between words, the calculated sentence similarity is far below the human-made score when several synonyms are included in the sentence pair. Some methods combine other datasets or contexts to infer the implied semantics of sentences with complex computational processes or large corpora. Our provided model (called Model I) that integrates quantum entanglement into sentence representations can explain the modification between words and express more semantic information. Considering that the tensor product can expand the semantic difference between sentences when several synonyms are included in the sentence pair, our proposed two advanced models (called Model II and Model III) based on Model I reduce the dimensions of the sentence representation to decrease the semantic difference between sentences. We use the different combinations of the three models to optimize our method to mine more semantic information in the sentence pair. Our provided method can compute the sentence similarity of all the sentence pairs with at least two notional words in each sentence. Moreover, it integrates quantum entanglement with sentence embedding and utilizes dimensionality reduction to reduce the semantic error in the sentence pair, with experimental results achieving state-of-the-art performances on 21 corpora.

III. APPROACHES

In this section, we first provide a sentence embedding based on quantum computation (Model I) and then construct two advanced models (Model II and Model III) to reduce the semantic error caused by the tensor product. Considering the dimensionality of sentence representation expanded to d^2 (the dimension of word2vec is d) by the tensor product, the high dimensionality of the sentence representation may give rise to some unnecessary semantic errors that cause the similarities of sentence pairs to decline. The advanced models have different effects on the sentence similarity in semantics. Model II that considers the overall influence on sentence pairs reduces the dimensionality from the sentence level. Model III that considers the local information of the sentence pair declines the dimension of the sentence embedding on the entangled words level. For the combination of the models, our provided method first considers the effect of Model II on the sentence pair with a large similarity and then introduces Model III to decrease the semantic errors of the sentence pair with a small similarity.

A. MODEL I

1) EXTRACT WORDS

Remove the function words to extract the notional words from the sentence. Store the notional words in an array A with the original sequences of the words in the sentence,

$$A = \{w_1, w_2, \cdots, w_i, \cdots, w_n\}, \tag{1}$$

where w_i is the ith notional word in the sentence and n is the total number of notional words.

2) NORMALIZE WORD VECTOR

$$|w_i\rangle = \frac{\vec{s_i}}{|\vec{s_i}|}, \tag{2}$$

where $\vec{s_i}$ is the vector of the ith word and $|\vec{s_i}|$ is the module of $\vec{s_i}$. $|w_i\rangle$ representing the normalized word vector is called a ket. A ket in quantum mechanics expresses a state in the Hilbert space, which is a column vector. The dimension of the word vector is d, so $|w_i\rangle$ is also a d dimensional column vector.

3) ENTANGLED WORDS VECTOR

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

$$B = \{(w_1w_2), (w_2w_3), (w_3w_4), \cdots, (w_{n-1}w_n)\}. \tag{3}$$

The definition of the entangled words vector is

$$|w_i\rangle|w_{i+1}\rangle = |w_iw_{i+1}\rangle = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{bmatrix} \otimes \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{bmatrix}, \tag{4}$$

where w_i is the ith word in array A, w_{i+1} is the right adjoint word of w_i, and $\otimes$ denotes the tensor product.
The definition of the tensor product of $\vec{a}$ and $\vec{b}$ is

$$\vec{a}\vec{b} = \vec{a} \otimes \vec{b} = \begin{bmatrix} a_1b_1 \\ a_2b_1 \\ \vdots \\ a_ib_j \\ \vdots \\ a_nb_m \end{bmatrix}. \tag{5}$$

V and W are Hilbert spaces with m and n dimensions, respectively, then $V \otimes W$ is an mn dimensional vector space.
Thus, the entangled words vector is

$$|w_iw_{i+1}\rangle = \begin{bmatrix} u_1v_1 \\ u_2v_1 \\ \vdots \\ u_iv_j \\ \vdots \\ u_dv_d \end{bmatrix}. \tag{6}$$

$|w_i\rangle$ and $|w_{i+1}\rangle$ are d dimensional vectors; then, $|w_i w_{i+1}\rangle$ is a d^2 dimensional vector.

4) SENTENCE REPRESENTATION

The sentence representation is defined as

$$|T\rangle = \sum_{i=1}^{n-1} |w_i w_{i+1}\rangle, \tag{7}$$

where we set all the entangled coefficients to 1 to simplify the sentence embedding.

5) SENTENCE SIMILARITY

The direction cosine of the two sentence representations is defined as the sentence similarity of the sentence pair. Hence, the sentence similarity is

$$cos(|T_1\rangle, |T_2\rangle) = \frac{\langle T_1|T_2\rangle}{||T_1\rangle| \cdot ||T_2\rangle|}, \tag{8}$$

where $\langle T_1|T_2\rangle$ denotes the inner product of $\langle T_1|$ and $|T_2\rangle$, $||T_1\rangle|$ and $||T_2\rangle|$ are the norms of $|T_1\rangle$ and $|T_2\rangle$, respectively, and $\langle T_1|$ is the conjugate transpose of $|T_1\rangle$.

B. ADVANCED MODELS

A vector with n^2 dimensions can be performed by the tensor product on two vectors with n dimensions, so the tensor can describe the states of the objects more detail than vectors. Consequently, using tensors to analyze the semantic relations between words may expand the differences between sentences, such as the following sentence pair.

S_a: *A man is playing on a guitar and singing*,

S_b: *A woman is playing an acoustic guitar and singing*.

The two sentences are different though they have many of the same words. The human-made score is only 0.44 (divided by 5). Compared to the two sentences, there is only one word 'man' in S_a different from S_b and the word 'acoustic' is absent. If we entangle the adjacent notional words together, there are 3 out of 5 words entanglement pairs, which is more than two. Due to the dimensionality expansion of the tensor product, the calculational similarities of the sentence pairs from the semantic analysis are apparently lower than human-made scores. The main reason is the tensor product expressing the semantic relations of the words in a very particular way. To reduce the impact of secondary factors, we provide two methods with dimensionality reduction: Model II and Model III.

1) MODEL II

In Model II, the dimension of the sentence representation is decreased to D_1, which is smaller than d^2 (d is the dimensionality of the word2vec). The main idea is to extract the top D_1 values and reorder them by their original indexes from the sentence representation based on quantum entanglement. The algorithm is illustrated in Algorithm 1.

Algorithm 1 An Improved Sentence Level Dimension Reduction Model Based on Quantum Entanglement

Input: array A, B, word2vec, sentence embedding dimension D, D_1.

Output: $cos\theta$

1: Input one sentence of the sentence pair, extract all the notional words and store in an array $A = \{w_1, w_1, \ldots, w_n\}$
2: Entangle the two adjacent notional words together to form the array $B = \{(w_1w_2), (w_2w_3), \ldots, (w_{n-1}w_n)\}$
3: Obtain the entangled words representation by the tensor product: $|w_i w_{i+1}\rangle$
4: Generate the sentence representation by linear superpositions of all the entangled words representations with D dimensions: $|T\rangle = \sum_{i=1}^{n-1} |w_i w_{i+1}\rangle$
5: Reduce the dimensionality of the sentence representation to D_1: remove the $D - D_1$ dimensions with smaller absolute values to achieve the sentence embedding as $|T_1^{\backslash}\rangle$
6: input the other sentence, repeat Step 1 to 5 to receive the sentence embedding $|T_2^{\backslash}\rangle$
7: Compute the direction cosine between the sentence pair: $cos\theta = \frac{\langle T_1^{\backslash}|T_2^{\backslash}\rangle}{||T_1^{\backslash}\rangle| \times ||T_2^{\backslash}\rangle|}$
8: **return** $cos\theta$

2) MODEL III

In Model III, the dimension of the entangled words representation declines to D_2, which is smaller than d^2 (d is the dimensionality of the word2vec). The main idea is to extract the top D_2 values and reorder them by their original indexes. Then, the modified entangled words vector is substituted into the sentence representation; as a consequence, the sentence embedding is modified to a vector with D_2 dimensions. We show the simulation steps in Algorithm 2.

C. CORRELATION OF THE THREE MODELS

If the dimensions of Model II and Model III are not reduced, namely, $D_1 = D_2 = D$, the three models are equivalent. When $D_1 < D$ or $D_2 < D$, the dimension reduction in Model II or Model III is effective. Then, they are different from Model I. The dimension reduction in Model II is on the sentence level, considering the global semantics of sentences. The dimension reduction in Model III is from the entangled words level, given the local semantics of sentences. Therefore, they have different priorities. Moreover, the granularity of the two advanced models is different. Specifically, Model II and Model III cannot be converted to each other, and all three models can be calculated separately on the similarity of any sentence pair.

IV. EXPERIMENTS

The difference between Model II and Model III is obvious. Model II decreases the dimensions of the sentence representation on the sentence level, but Model III decreases the

Algorithm 2 An Improved Entangle Words Level Dimension Reduction Model Based on Quantum Entanglement

Input: array A, B, word2vec, entangled words embedding dimension D, D_2.

Output: $cos\theta$

1: Input one sentence of the sentence pair, extract all the notional words and store in an array $A = \{w_1, w_1, \ldots, w_n\}$
2: Entangle the two adjacent notional words together to form the array $B = \{(w_1w_2), (w_2w_3), \ldots, (w_{n-1}w_n)\}$
3: Obtain the entangled words representation by the tensor product: $|w_iw_{i+1}\rangle$
4: Decrease the dimensions of the entangled words representation to D_2: remove the $D - D_2$ dimensions with smaller absolute values to achieve the entangled words embedding as $|w_i^{\backslash}w_{i+1}^{\backslash}\rangle$
5: Generate the sentence representation by linear superpositions of all the entangled words representations with D_2 dimensions: $|T_1^{\backslash\backslash}\rangle = \sum_{i=1}^{n-1} |w_i^{\backslash}w_{i+1}^{\backslash}\rangle$
6: input the other sentence, repeat Step 1 to 5 to receive the sentence embedding $|T_2^{\backslash\backslash}\rangle$
7: Compute the direction cosine between the sentence pair: $cos\theta = \frac{\langle T_1^{\backslash\backslash}|T_2^{\backslash\backslash}\rangle}{||T_1^{\backslash\backslash}\rangle|\times||T_2^{\backslash\backslash}\rangle|}$
8: **return** $cos\theta$

dimensions of the sentence representation on the level of the entangled words. The two methods also have different effects on sentence similarity. Consequently, we utilize the different models to discover the different influences on the semantic analysis and syntax structures. Model II focuses on the selection of overall sentence attributes, which is suitable for the semantic analysis of the sentence pair with a large similarity. However, Model III focuses on the characteristic distribution of the entangled words and describes the semantics in a more detailed way. Moreover, it can grasp the main influential factors of the entangled words while ignoring the secondary factors and is suitable for the sentence pair with low similarity. Thus, we use Model II and Model III to optimize the sentence representation differently. Subscripts 1 and 2 identify the physical quantities of Model II and Model III, respectively.

A. COMBINATION OF MODEL I AND MODEL II

We first define three variables: σ_1, E_1 and λ_1. σ_1 means the threshold of the human-made score y_1 of the sentence pair. The relative error E_1 is defined as

$$E_1 = \frac{|y_1 - S_1|}{y_1}, \tag{9}$$

where S_1 is the calculational value of the sentence similarity modeled by Model I. λ_1 is defined as the threshold value of the relative error. When $y_1 > \sigma_1$ and $E_1 > \lambda_1$, Model II is used to compute the similarity of the sentence pair once more. For the other sentence pairs, we keep the results of Model I to calculate the Pearson correlation coefficient (Pcc) and mean squared error (MSE).

When all the sentence pairs are calculated by Model I, Pcc and MSE are defined as follows.

$$Pcc = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\delta_x \delta_y} = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2}\sqrt{N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}}, \tag{10}$$

$$MSE = \frac{\sum_{i=1}^{N}\sqrt{(x_i - y_i)^2}}{N}. \tag{11}$$

where x_i denotes the experimental results, y_i is the human score of the ith sentence pair and N is the total sentence pairs in a corpus. If sentences in one corpus are computed by two different models, we change the standard deviation of the calculated sentence similarities as follows, denoted by δ_x.

$$\delta_x = \sqrt{\frac{(N_0 - 1)\delta_0^2 + (N_1 - 1)\delta_1^2}{N_0 + N_1 - 2}}, \tag{12}$$

where $N_0 + N_1 = N$. N_0 sentences pairs are computed by Model I with the standard deviation of δ_0, and the other N_1 sentences are calculated by Model II with the standard deviation of δ_2. We replace δ_x in Equation (10) with Equation (12) to obtain the expression of Pcc.

$$MSE = \frac{\sum_{i=1}^{N_0}\sqrt{(x_i - y_i)^2} + \sum_{j=1}^{N_1}\sqrt{(x_j - y_j)^2}}{N}, \tag{13}$$

where x_i is the calculated similarity of the sentence pair modeled by Model I and the total text modeled by Model I is N_0, x_j is the calculated similarity of the sentence pair modeled by Model II, and the total text modeled by Model II is N_1.

B. COMBINATION OF MODEL I, MODEL II AND MODEL III

Four variables σ_2, γ_2, E_2 and λ_2 are introduced. σ_2, E_2, y_2 and λ_2 are defined as the same means of σ_1, E_1, y_1 and λ_1. γ_2 is the minimum of the human-made score y_2 of the sentence similarity. When $y_1 > \sigma_1$ and $E_1 > \lambda_1$, Model II is introduced. When $\gamma_2 < y_2 < \sigma_2$ and $E_2 > \lambda_2$, Model III is introduced. Where λ_1 and λ_2 can be selected as the different values, σ_1 and σ_2 cannot also be the same, and γ_2 cannot be equal to 0. The sentence pairs not satisfying the two conditions mentioned above are modeled by Model I. Therefore, all the sentence pairs are computed not more than twice. When calculating Pcc and MSE, we replace the similarity of the sentence pair computed by Model I with the values recalculated either by Model II or by Model III. Consequently, each calculational similarity of the sentence pair is used one time. There are N_1 sentence pairs recomputed by Model II and N_2 pairs recalculated by Model III, so the other $N_0 = N - N_1 - N_2$ sentences are modeled by Model I. The standard deviation of the calculated similarities of sentence

TABLE 1. Datasets for the SemEval Semantic Textual Similarity Tasks (year 2012, 2014, 2015, 2016). Note that, the figures in bracket refer to the number of sentence pairs in the corpus.

STS'12	STS'14	STS'15	STS'16
MSRvid (750)	deft-forum (450)	answers-forums (375)	plagiarism (230)
SMTeuroparl (459)	deft-news (300)	answers-students (750)	postediting (244)
OnWN (750)	headlines (750)	belief (375)	answer-answer (254)
MSRpar (750)	images (750)	images (750)	headlines (249)
SMTnews (399)	tweet-news (750)	headlines (750)	question-question (209)
	OnWN (750)		

TABLE 2. Some collected Pearson correlation in each dataset from 'An Efficient Framework for Sentence Similarity Modeling' [1]. The figures in bold refer to the maximum Pearson correlation of each corpus in [1].

Year	Dataset	PROJ	PP-tfidf	DAN	*RNN*	LSTM	ST	GloVe	PSL	iRNN	SCBOW	ACVT
2012	MSRpar	0.44	0.47	0.40	0.19	0.09	0.17	0.48	0.42	0.43	0.44	**0.58**
	MSRvid	0.74	0.79	0.70	0.67	0.71	0.42	0.64	0.60	0.73	0.45	**0.83**
	SMTeuroparl	0.49	**0.52**	0.44	0.41	0.44	0.35	0.46	0.42	0.47	0.45	0.43
	OnWN	0.70	**0.73**	0.66	0.63	0.56	0.30	0.55	0.63	0.70	0.64	0.70
	SMTnews	0.63	**0.66**	0.60	0.51	0.51	0.31	0.50	0.57	0.58	0.39	0.54
2014	deft-forum	0.51	**0.54**	0.49	0.42	0.46	0.13	0.27	0.37	0.49	0.41	0.48
	deft-news	0.72	**0.74**	0.72	0.54	0.39	0.24	0.68	0.67	0.72	0.59	**0.74**
	headlines	0.71	0.71	0.69	0.58	0.51	0.38	0.60	0.65	0.70	0.64	**0.72**
	images	0.78	0.81	0.77	0.68	0.63	0.51	0.61	0.62	0.78	0.65	**0.81**
	OnWN	0.80	0.81	0.76	0.68	0.62	0.23	0.58	0.61	0.79	0.61	**0.87**
	tweet-news	0.76	**0.77**	0.74	0.58	0.48	0.40	0.51	0.65	**0.77**	0.73	0.75
2015	answers-forums	0.65	0.68	0.63	**0.73**	0.51	0.36	0.31	0.39	0.67	0.22	0.69
	answers-students	0.78	**0.79**	0.78	0.65	0.56	0.33	0.63	0.69	0.78	0.37	**0.79**
	belief	0.75	**0.78**	0.72	0.52	0.53	0.25	0.41	0.53	0.76	0.48	0.70
	images	0.80	**0.84**	0.78	0.71	0.64	0.18	0.68	0.70	0.81	0.26	0.82
	headlines	0.75	0.77	0.74	0.65	0.57	0.44	0.62	0.68	0.75	0.22	**0.79**

pairs are defined as δ_x.

$$\delta_x = \sqrt{\frac{(N_0-1)\delta_0^2 + (N_1-1)\delta_1^2 + (N_2-1)\delta_2^2}{N_0+N_1+N_2-3}}, \quad (14)$$

where δ_0, δ_1 and δ_2 are the standard deviation of the similarities of sentence pairs modeled by Model I, Model II and Model III, respectively. δ_x in Equation (10) is replaced by Equation (10) to achieve Pcc. The MSE is changed as follows.

$$\begin{aligned} MSE &= \frac{\sum_{i=1}^{N_0+N_1+N_2}\sqrt{(x_i-y_i)^2}}{N_0+N_1+N_2} \\ &= \frac{\sum_{i=1}^{N}\sqrt{(x_i-y_i)^2}}{N}. \end{aligned} \quad (15)$$

V. EXPERIMENTAL RESULTS

A. DATASETS

In the study, the public word2vec lib is assigned, and every word is a 300-dimensional vector [55]. Particularly, the sentence pair is deleted if either sentence has fewer than two notional words. Experiments are carried out with the corpora released by SemEval Semantic Textual Similarity Tasks, including the years 2012, 2014, 2015 and 2016, as shown in Table 1 [56]. We only study the semantic textual similarity of English in corpora and ignore the other language contents. The corpora of the four years are generated as following. E. Agirre *et al.* presented a pilot on semantic textual similarity and provided five corpora including Microsoft research paraphrases, videos and statistical machine translations [57]. In 2014, E. Agirre *et al.* added OntoNotes-WordNet sense mappings, news headlines and new genres to the previous corpus [58]. Sentence pairs from headlines, image descriptions and committed belief annotations and answer pairs from tutorial dialog systems and QA websites were introduced in 2015 [59]. The SemEval-2016 task involves plagiarism detection, postedited machine translations, questions-answers and article headlines of news [60]. These corpora consist of sentence pairs and their textual similarities ranging from 0.0 to 5.0. To compare with experimental scores, we divide each human-made score by 5.

B. EXPERIMENTAL SETTINGS

In this subsection, we discuss the results of the designed experiments in various aspects and then compare them with the methods that used word/sentence embedding, as illustrated in Table 2 [1]. 'ACVT' is the proposed method, and the bold-type figures are the best values for every corpus, as indicated in Table 2. The adjustment processes of the parameters of λ_1, σ_1, γ_2, λ_2 and σ_2 are as follows. First, for combination of Model I and Model II, we adjust the parameters λ_1 and σ_1 to obtain the best Pcc of each corpus. Second, for the combination of Model I, Model II and Model III, we first adjust the parameters λ_1 and σ_1 to the optimum values and obtain the best Pcc value by adjusting the parameters γ_2, λ_2 and σ_2. The last Pcc is the optimal value, and the MSE of the last Pcc is set as the optimum MSE value.

C. COMPARING WITH WORD EMBEDDING-BASED METHODS

Table 3 lists the comparison of the results of the proposed methods to the selected results from Table 2 [1]. As illustrated by Table 3, 'ACVT' denotes the proposed method in [1], and 'Best Values' denotes the maximum Pcc for each corpus collected from Table 2. 'Model I and II' denotes the combined

TABLE 3. Comparison our methods with other methods of Pearson correlation in each dataset. 'ACVT' means the experimental results of the model provided in [1], 'Best Values' is the best result of each corpus compared with all the results in Table 2. The figures in bold refer to the maximum Pearson correlation of each corpus.

Year	Dataset	ACVT	Best Values	Our methods	
				Model I and II	Model I, II and III
2012	MSRpar	0.58	0.58	0.56	**0.59**
	MSRvid	0.83	0.83	0.88	**0.90**
	SMTeuroparl	0.43	0.52	0.69	**0.69**
	OnWN	0.70	0.73	0.83	**0.84**
	SMTnews	0.54	0.66	0.77	**0.78**
2014	deft-forum	0.48	0.54	0.63	**0.66**
	deft-news	0.74	0.74	0.76	**0.78**
	headlines	0.72	0.72	0.79	**0.81**
	images	0.81	0.81	0.86	**0.87**
	OnWN	0.87	0.87	0.91	**0.92**
	tweet-news	0.75	0.77	0.82	**0.82**
2015	answers-forums	0.69	0.69	0.85	**0.86**
	answers-students	0.79	0.79	0.85	**0.86**
	belief	0.70	0.78	0.86	**0.87**
	images	0.82	0.84	0.87	**0.89**
	headlines	0.79	0.79	0.82	**0.85**

Model I and Model II method. 'Model I, II and III' denotes the proposed combination of Model I, Model II and Model III. In Table 3, we discover that all Pccs calculated by the proposed methods for each dataset exceed the best values collected from Table 2 except for Model I and II on STS'12.MSRpar. For STS'12.MSRpar, Pcc calculated by our first method is only less than 'Best Values' (the relative percentage is just 1.7) by 0.02, but Pcc of our second method increases by 0.01 to 0.59.

For the methods of Model I and II, the corpus with the greatest improvement rate is STS'12.SMTeuroparl with an improvement rate to 32.7% relative to 'Best Values' and an improvement rate of 60.5% relative to 'ACVT'. The corpus with the second-highest improvement rate is STS'15.answers-forum with an improvement rate of 23.2%. The datasets with the third-highest improvement rate are STS'12.SMTnews and STS'14.deft-forum, with an improvement rate to 16.7%. In addition, the datasets with improvement rates over 10% are STS'12.OnWN and STS'15.belief, with improvement rates of 13.7% and 10.3%, respectively. Moreover, Pcc of STS'14.headlines is 0.79, which is more 0.07 higher than 'Best Values', with an improvement rate of 9.7%, which is slightly lower than 10%. Contrasting the influences of the different combinations on Pcc of all the corpora, the maximum absolute improvement is 0.17 for STS'12.SMTeuroparl, and the next is STS'15.answers-forums, which achieved 0.15. There are four corpora with Pcc improvement exceeding 0.1.

Considering the combination of Model I, Model II and Model III, the results of our proposed method are higher than 'Best Values' for all the STS datasets. The corpus with the greatest improvement rate is also STS'12.SMTeuroparl with an improvement rate of 32.7% relative to 'Best Values' and an improvement rate of 60.5% relative to 'ACVT'. The corpus with the second-highest improvement rate is STS'15.answers-forum with an improvement rate of 24.6%. The datasets with the next highest improvement rate are STS'12.SMTnews and STS'14.deft-forum, with improvement rates of 22.2% and 18.2%, respectively. Additionally, the datasets with improvement rates over 10% are STS'12.OnWN, STS'14.headlines and STS'15.belief, with an improvement rates of 15.1%, 12.5% and 11.5%, respectively. There are 7 out of 16 datasets that improved over 10%, including one corpus that improved over 30% and two corpora that improved over 20% with rates of 24.6% and 22.2%, respectively. Compared to the effects of the two proposed methods, the maximum absolute growth is 0.17 for STS'12.SMTeuroparl and the next is STS'15.answers-forums achieving 0.16. In addition, there are three corpora with Pcc that increased by over 0.1.

Fig. 1 and Fig. 2 illustrate the comparison of the calculated results of different years of STS with the histogram. The height of the histogram of the proposed methods for each corpus is more than that of the best results collected from other studies listed in Table 2. Excluding MSRpar shown in Fig. 1, the differences in the heights of the histograms for all the other corpora are excellent. Consequently, it is generally accepted that our proposed methods significantly improve Pcc of every STS dataset, which demonstrates that the proposed methods are effective and valuable.

D. OVERALL RESULTS

Table 4 evaluates the influence of different combinations of different models. It is evident that every result calculated by the combination of Model I, Model II and Model III is no less than the result of the corresponding corpus computed by Model I and II. Four corpora increase by 0.03, five corpora increase by 0.02, ten corpora increase by 0.01, and two corpora remain unchanged. Moreover, for the combination of Model I and Model II, the corpus with the maximum Pcc is STS'14.OnWN, which reaches 0.91. There are fifteen out of

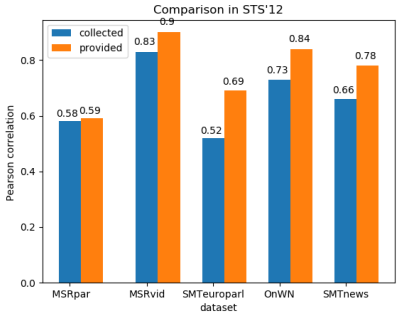
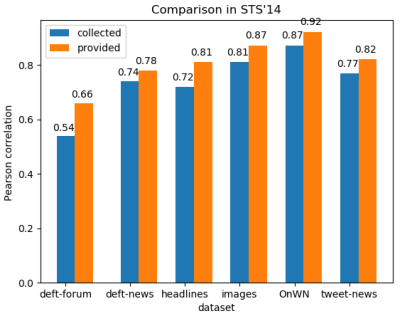
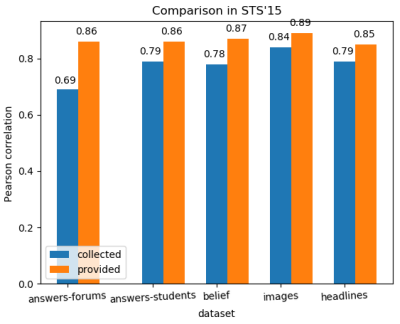


FIGURE 1. Comparison of Pearson correlation with other methods in STSí12 dataset. 'collected' refers the best value collected from Table 2, and 'proposed' indicates the experimental result of our proposed method for each corpus.



(a)



(b)

FIGURE 2. Comparison of Pearson correlation with other methods in STS dataset year of 2014 and 2015. 'collected' refers the best value collected from Table 2, and 'proposed' indicates the experimental result of our proposed method for each corpus.

twenty-one corpora with Pcc exceeding 0.8 and three corpora close to 0.8, which are 0.77, 0.76 and 0.78. Subsequently, for the combination of Model I, Model II and Model III, the maximum Pcc reaches 0.92 from STS'14.OnWN. Pccs of STS'12.MSRvid and STS'14.OnWN exceed 0.9. It is important that the two corpora with Pccs exceeding 0.9 contain 750 sentence pairs. There are sixteen Pccs out of twenty-one that are higher than 0.8, and two corpora are close to 0.8, which are the same value of 0.78. It is amazing to find that all Pccs of STS'15 and STS'16 surpass 0.8, which means that the influences of the proposed models are predominant.

The MSE of different corpora influenced by the different models can be observed in Table 5. MSE is a measure reflecting the degree of the difference between the estimated values

TABLE 4. Pearson correlation of our methods in each dataset.

Year	Dataset	Model I and II	Model I, II and III
2012	MSRpar	0.56	0.59
	MSRvid	0.88	0.90
	SMTeuroparl	0.69	0.69
	OnWN	0.83	0.84
	SMTnews	0.77	0.78
2014	deft-forum	0.63	0.66
	deft-news	0.76	0.78
	headlines	0.79	0.81
	images	0.86	0.87
	OnWN	0.91	0.92
	tweet-news	0.82	0.82
2015	answers-forums	0.85	0.86
	answers-students	0.85	0.86
	belief	0.86	0.87
	images	0.87	0.89
	headlines	0.82	0.85
2016	answer-answer	0.84	0.85
	headlines	0.81	0.82
	plagiarism	0.89	0.90
	postediting	0.87	0.88
	question-question	0.83	0.86

TABLE 5. Mean squared error of our methods in each dataset.

Year	Dataset	Model I and II	Model I, II and III
2012	MSRpar	0.028	0.026
	MSRvid	0.022	0.019
	SMTeuroparl	0.015	0.016
	OnWN	0.033	0.025
	SMTnews	0.025	0.021
2014	deft-forum	0.048	0.043
	deft-news	0.035	0.032
	headlines	0.041	0.038
	images	0.024	0.022
	OnWN	0.024	0.022
	tweet-news	0.033	0.031
2015	answers-forums	0.036	0.025
	answers-students	0.036	0.025
	belief	0.020	0.020
	images	0.027	0.024
	headlines	0.040	0.033
2016	answer-answer	0.044	0.036
	headlines	0.048	0.044
	plagiarism	0.025	0.023
	postediting	0.031	0.029
	question-question	0.032	0.029

and the measured values. In this work, MSE indicates the fitting degree between the calculated values and human-made scores of sentence similarities from the semantic analysis. The smaller the MSE is, the higher the fitting degree. All the values in Table 5 are less than 0.05, the minimum MSE is just 0.015 and the maximum MSE is only 0.048. Data from Models I and II illustrate that the MSE of the corpora STS'14.deft-forum and STS'16.headlines are 0.048 achieving the peak though they are infinitesimal. There are five corpora of which the MSE exceeds 0.04 and seven corpora between 0.03 and 0.039. Compared to the corpora of STS'12, it is evident from the results that the proposed method performs dominantly because all the MSEs of datasets are less than 0.035. The dataset with the lowest MSE is STS'12.SMTeuroparl, just 0.015. As detailed in Table 5 from the combination of Model I, Model II and Model III, the minimum MSE is 0.016, which is more 0.001 than that of Model I and II.

The maximum MSE is only 0.044, which is less 0.004 than that of Model I and II. MSEs of the two corpora STS'12.MSRvid and STS'12.SMTeuroparl are insufficient to 0.02, with 0.019 and 0.016, respectively. There are fourteen out of twenty-one corpora with MSEs below 0.03, which demonstrates that the performances of our proposed methods are perfect. Compared to the results of the two proposed methods, except for dataset STS'12.SMTeuroparl, all MSEs of the corpora are decreased by the introduction of Model III. The MSEs of STS'12.OnWN and STS'16.answer-answer are reduced by 0.008. The greatest reduction in MSEs is 0.011 for STS'15.answers-students.

Comparing Table 4 to Table 5, except for STS'12.SMTeuroparl, all Pccs of the other corpora increase and the MSEs are reduced by introducing Model III, as illustrated in Fig. 3 and Fig. 4. For STS'12.SMTeuroparl, the Pcc is unchanged but the MSE increases by 0.001 after the introduction of Model III. For the first three corpora with the greatest changeable values of MSE, Pcc increases by only 0.01. The MSEs of the corpora with Pccs exceeding 0.9 are small with values of 0.022, 0.023 and 0.019. However, we cannot reach such a conclusion as the higher Pcc with the lower MSE. For example, the Pcc of STS'12.MSRpar is 0.59 with the value of 0.026 for MSE, but the Pcc of STS'16.headlines is 0.82 with the value of 0.044 for MSE. The main reason is that the combination of Model I, Model II and Model III decreases the difference between the calculated values and the human-made values of some sentence pairs. The great reduction in MSE can be explained by the characteristics of Model II and Model III. Moreover, comparing the variation tendencies between the Pcc and MSE of the same corpus in Table 4 and Table 5, the conclusion is obtained that the Pcc and MSE can analyze the semantic information of sentence pairs from different standpoints.

E. DETAILED RESULTS

In this subsection, we discuss the Pcc and MSE of some specific corpora by adjusting the parameters of Model II and Model III. The adjustment process is as follows. First, the similarities S_1 of all the sentence pairs in the corpus are calculated by Model I. Second, Equation (9) is used to compute the similarity error E_1. When the sentence similarity error E_1 satisfies the condition $E_1 > \lambda_1$, Model II is selected to recalculate the sentence similarity. Subsequently, the combination of Model I and Model II is formed. Third, the parameters of Model II σ_1, λ_1 and D_1 are regulated to achieve the optimal values of Pcc and MSE. Finally, Model III is introduced to optimize the combination of Model I and Model II, which is called the combination of Model I, Model II and Model III. choose some appropriate values for γ_2 and σ_2 according to the value of σ_1. When the sentence similarity and the similarity error both satisfy the following condition: $E_2 > \lambda_2$ and $\gamma_2 < y_2 < \sigma_2$, Model III is used to compute the sentence similarity. The parameters of Model III σ_2, γ_2, λ_2 and D_2 are adjusted to optimize the values of the Pcc and MSE. The values of σ_1 and σ_2 satisfy the condition of $\sigma_1 > \sigma_2$, so all the

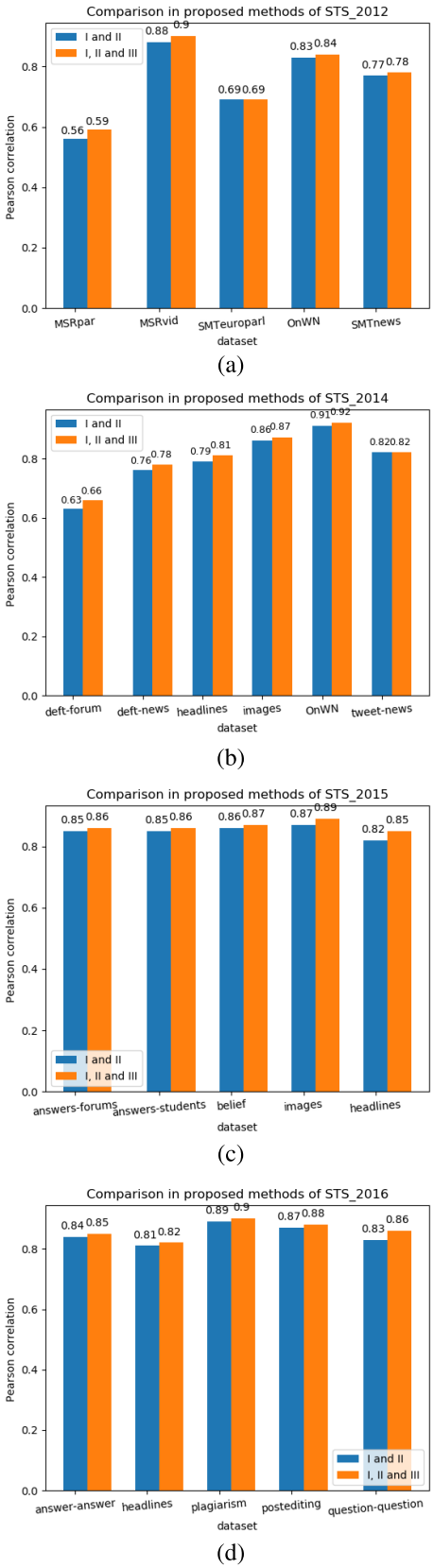
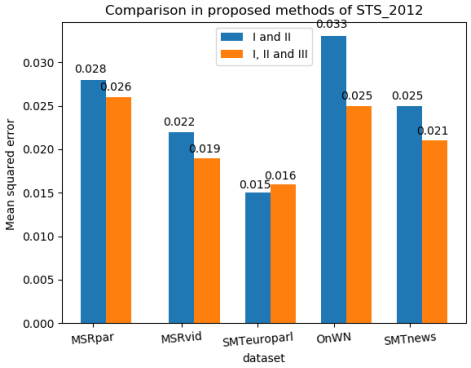
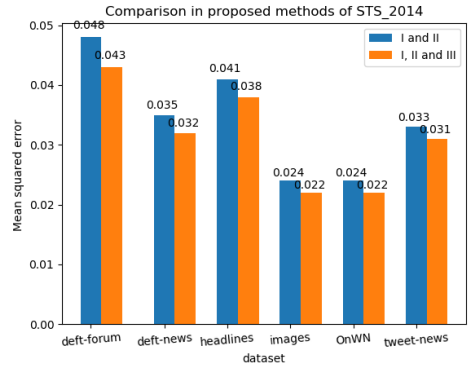


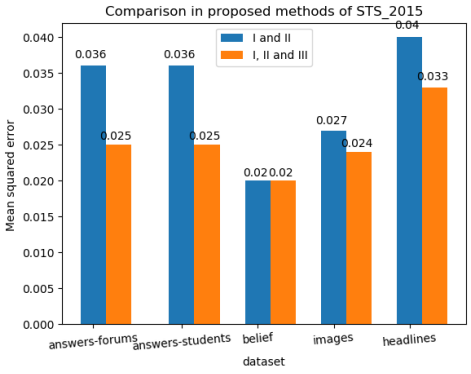
FIGURE 3. Comparison of Pearson correlation between the two combinations of the proposed models. 'I and II' indicates the combination of Model I and Model II, and 'I, II and III' refers the combination of all the proposed models.



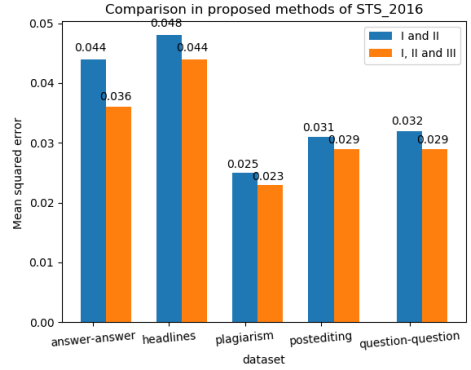
(a)



(b)



(c)



(d)

FIGURE 4. Comparison of MSE between the two combinations of the proposed models. 'I and II' indicates the combination of Model I and Model II, and 'I, II and III' refers the combination of the three proposed models.

sentence pairs are recalculated by either Model II or Model III one time. The similarities of all sentence pairs in each corpus are used just one time for the calculations of Pcc and MSE. If the sentence pair is recalculated by Model II or Model III, the sentence similarity computed by Model I is replaced by the recalculated value when computing Pcc and MSE. All the parameters of the combination of Model I, Model II and Model III are adjusted after the adjustments of the parameters of Model II in the following tables.

1) EXPERIMENTAL RESULTS OF STS'12.SMTeuroparl

Table 6 exhibits the influence of γ_2, λ_2 and D_2 on STS'12. SMTeuroparl, which is comprised of 459 sentence pairs. A large number of the annotated similarities of sentence pairs are greater than 0.8, and no sentence pairs with annotated similarities are less than 0.3. As a result, when changing γ_2 from 0.2 to 0.3, the Pcc and MSE both remain invariant. Attributed to the large number of human-made scores over 0.8, the effect of Model II is better than that of Model III. Therefore, the Pcc of the combination of Model I and Model II is higher than that of the combination of the provided three models. In the majority of cases, MSEs decrease apparently by the introduction of Model III, as shown in Table 6. The detailed comparable charts affected by λ_2 and D_2 are expressed in Fig. 5(a) and Fig. 6(a). Both λ_2 and D_2 can markedly alter the Pcc of the datasets.

2) EXPERIMENTAL RESULTS OF STS'14.input.images

Table 7 explains the comparison of the combination of Model I and Model II to the combination of the three provided models in STS'14.input.images consisting of 750 sentence pairs with sentence similarities ranging from 0.0 to 1.0. Compared to the Pcc and MSE, all the values of MSE are very small but with high Pccs, which illustrates that Model III can decrease the semantic error apparently for the sentence pair.

3) EXPERIMENTAL RESULTS OF STS'15.input.images

The effect of the parameters on STS'15.input.images is given in Table 8, and the detailed comparison charts of Pcc and MSE are interpreted in Fig. 5(c) and Fig. 6(c), respectively. The corpus STS'15.input.images is composed of 750 sentence pairs with human-made sentence similarities from 0.0 to 1.0. There are some long sentences with the number of notional words over 10 in the dataset. The longer the sentence is, the greater the sentence semantic differences are. The impacts of the entangled words operated by the tensor product are marked in semantics; as a consequence, Model III plays an irreplaceable role in the reduction in semantic noise. As Fig. 5(c) and Fig. 6(c) show, the influences of D_2 are more obvious than those of λ_2, which can be attributed to the considerably transformed Pcc and MSE.

4) EXPERIMENTAL RESULTS OF STS'16.ANSWER-ANSWER

Table 9 shows the influences of different combinations of different models on STS'16.input.answer-answer, which contains 259 sentence pairs with the annotated sentence

TABLE 6. Pearson correlation and mean squared error are influenced by the different models and the parameters of the same combinations of models in the dataset: STS'12.SMTeuroparl.txt.

Models	Parameters	Pearson Correlation	Mean Squared Error
I and II	$\sigma_1 = 0.7, \lambda_1 = 0.2, D_1 = 10000$	0.690	0.0151
I, II and III	$0.3 < y_2 < 0.7, \lambda_2 = 0.4, D_2 = 85000$	0.688	0.0163
	$0.3 < y_2 < 0.7, \lambda_2 = 0.4, D_2 = 75000$	0.687	0.0094
	$0.3 < y_2 < 0.7, \lambda_2 = 0.4, D_2 = 65000$	0.643	0.0085
	$0.2 < y_2 < 0.7, \lambda_2 = 0.4, D_2 = 75000$	0.687	0.0094
	$0.2 < y_2 < 0.7, \lambda_2 = 0.3, D_2 = 75000$	0.672	0.0092
	$0.2 < y_2 < 0.7, \lambda_2 = 0.2, D_2 = 75000$	0.667	0.0092

TABLE 7. Pearson correlation and mean squared error are influenced by the different models and parameters of the same combinations of models in the dataset: STS'14.input.images.txt.

Models	Parameters	Pearson Correlation	Mean Squared Error
I and II	$\sigma_1 = 0.5, \lambda_1 = 0.25, D_1 = 25000$	0.860	0.0239
I, II and III	$0.1 < y_2 < 0.5, \lambda_2 = 0.3, D_2 = 85000$	0.883	0.0200
	$0.1 < y_2 < 0.5, \lambda_2 = 0.4, D_2 = 85000$	0.882	0.0200
	$0.2 < y_2 < 0.5, \lambda_2 = 0.4, D_2 = 85000$	0.873	0.0217

TABLE 8. Pearson correlation and mean squared error are influenced by the different models and parameters of the same combinations of models in the dataset: STS'15.input.images.txt.

Models	Parameters	Pearson Correlation	Mean Squared Error
I and II	$\sigma_1 = 0.3, \lambda_1 = 0.35, D_1 = 37000$	0.8724	0.0267
I, II and III	$0.1 < y_2 < 0.3, \lambda_2 = 0.4, D_2 = 80000$	0.8727	0.0271
	$0.1 < y_2 < 0.3, \lambda_2 = 0.3, D_2 = 80000$	0.8723	0.0275
	$0.1 < y_2 < 0.3, \lambda_2 = 0.3, D_2 = 85000$	0.8875	0.0244

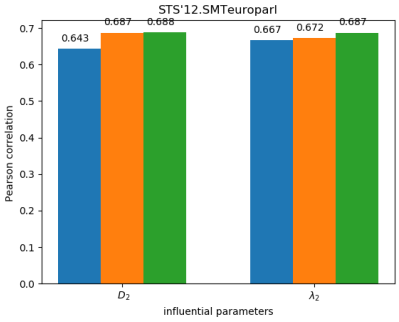
TABLE 9. Pearson correlation and mean squared error are influenced by the different models and parameters of the same combinations of models in the dataset: STS'16.answer-answer.txt.

Models	Parameters	Pearson Correlation	Mean Squared Error
I and II	$\sigma_1 = 0.4, \lambda_1 = 0.3, D_1 = 10000$	0.838	0.044
I, II and III	$0.15 < y_2 < 0.3, \lambda_2 = 0.3, D_2 = 85000$	0.846	0.036
	$0.15 < y_2 < 0.5, \lambda_2 = 0.3, D_2 = 85000$	0.847	0.036
	$0.0 \leq y_2 < 0.5, \lambda_2 = 0.3, D_2 = 85000$	0.859	0.033
	$0.0 \leq y_2 < 0.5, \lambda_2 = 0.3, D_2 = 75000$	0.774	0.066
	$0.1 < y_2 < 0.5, \lambda_2 = 0.3, D_2 = 75000$	0.842	0.042
	$0.1 < y_2 < 0.5, \lambda_2 = 0.3, D_2 = 80000$	0.856	0.036

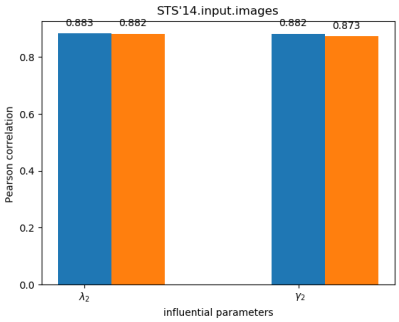
similarities of 0.0, 0.2, 0.4, 0.6, 0.8 and 1.0. The human-made score of the sentence similarity is just one of the six values. The lengths of the two sentences are different in some sentence pairs. For example, one sentence of some sentence pairs in the corpus is very short with just two national words, but the other sentence is more than five notional words. In the corpus, many high-frequency words are not considered in the provided methods according to the abstraction laws of words. With some colloquial words and short sentence pairs, the Pcc and MSE are sensitive to changes in the parameters, as listed in Fig. 5(d) and Fig. 6(d). When $\sigma_1 = 0.4$, $\lambda_1 = 0.3$, $D_1 = 10000$, $\lambda_2 = 0.3$ and $D_2 = 85000$, comparing Pcc and MSE of $0.15 < y_2 < 0.3$ to that of $0.15 < y_2 < 0.5$, the diversifications are insignificant with only 0.001, as detailed in the first columns in Fig. 5(d) and Fig. 6(d). Compared to the second columns in Fig. 5(d) and Fig. 6(d), when D_2 is changed from 85000 to 75000 on the condition of $\sigma_1 = 0.4$, $\lambda_1 = 0.3$, $D_1 = 10000$, $0.0 \leq y_2 < 0.5$ and $\lambda_2 = 0.3$, the Pcc declines by 0.085 varying from 0.859 to 0.774, and the MSE is approximately doubled with the variation from 0.033 to 0.066. When we only alter γ_2 from 0.0 to 0.1, Pcc and MSE, as explained in the third columns in Fig. 5(d) and Fig. 6(d), respectively, vary dramatically but are smaller than the influence of D_2 shown in the second columns in Fig. 5(d) and Fig. 6(d), respectively. When $\sigma_1 = 0.4$, $\lambda_1 = 0.3$, $D_1 = 10000$, $\lambda_2 = 0.3$ and $D_2 = 75000$, comparing the Pcc and MSE of $0.0 \leq y_2 < 0.5$ to that of $0.1 \leq y_2 < 0.5$, respectively, the alterations are apparent with the value of Pcc changing from 0.774 to 0.842 and the MSE changing from 0.066 to 0.042, as displayed in the last columns in Fig. 5(d) and Fig. 6(d), respectively.

5) SUMMARY

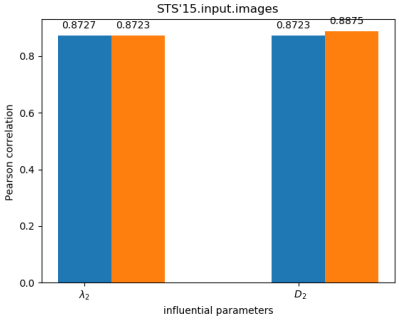
In summary, compared to the Pcc and MSE of the four corpora influenced by Model II and Model III from Table 6 to Table 9, the effect of Model III on long sentence pairs with low sentence similarities is better than the effect of Model II,



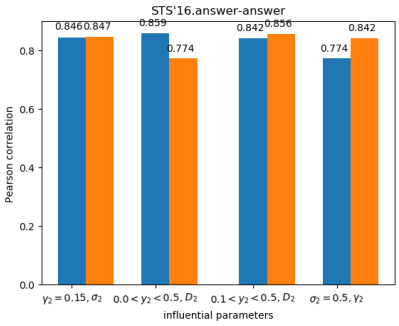
(a)



(b)

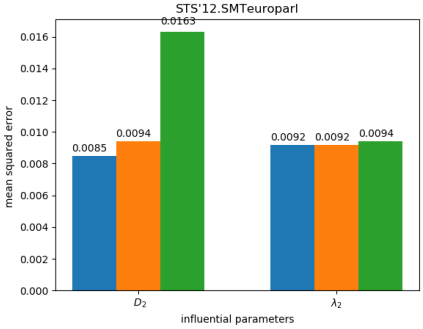


(c)

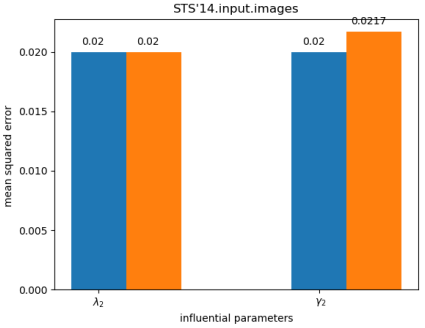


(d)

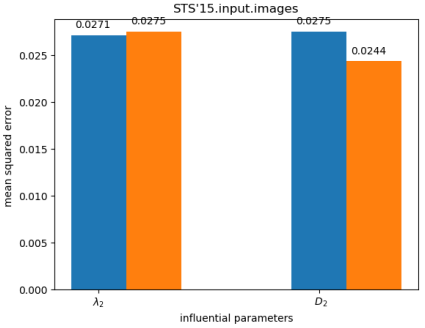
FIGURE 5. Pearson correlation influenced by different parameters.



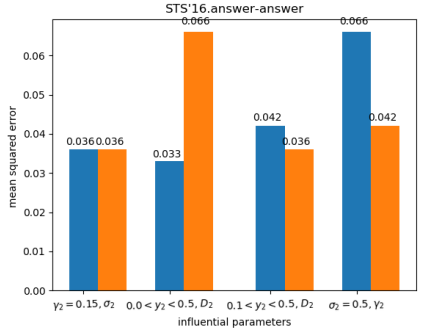
(a)



(b)



(c)



(d)

FIGURE 6. Mean squared error influenced by different parameters.

but for the sentence pairs with high similarities, the effect of Model II is better. The impact of Model III on the corpus consisting of more sentences with more notional words is clearer. Take STS'14.input.images as an example. All Pccs of the combination of Model I, Model II and Model III are much higher than that of the combination of Model I and Model II, and all the MSEs of the combination of Model II, Model II and Model III are lower than that of the combination of Model I and Model II. The influence of parameters on the data with a small number of sentence pairs is more obvious, as evident from STS'16.answer-answer.

VI. CONCLUSION

This study demonstrates how to integrate quantum theory into text embedding to construct sentence representations based on quantum entanglement. Considering that the dimension expansion of the entangled words vector caused by the tensor product may introduce some semantic noise, our models on dimensionality reduction are reported, which incorporate the physical idea of identifying the principal contradictions and ignoring the secondary contradictions. The Pcc and MSE of each corpus are obtained and compared with the results of other models. Experiments are implemented on 21 datasets, including the SemEval Semantic Textual Similarity Tasks (years 2012, 2014, 2015, 2016). It is clear from the above discussions that 16 out of 16 datasets outperform the comparative methods significantly and need no prior knowledge except for word2vec. The data from the experiments indicate the advantage of our approaches in that sentence embedding based on quantum computation taking dimensionality reduction into account can efficiently mine semantic information without complex computing processes. For future work, we attempt to extend the current framework to some study on the semantic structure of sentences considering the different weights of words.

ACKNOWLEDGMENT

The authors would like to thank the referees for their valuable remarks.

REFERENCES

[1] Z. Quan, Z.-J. Wang, Y. Le, B. Yao, K. Li, and J. Yin, "An efficient framework for sentence similarity modeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 853–865, Apr. 2019.

[2] J. Gao, L.-F. Qiao, Z.-Q. Jiao, and, "Experimental machine learning of quantum states," *Phys. Rev. Lett.*, vol. 120, May 2018, Art. no. 240501.

[3] G. R. Steinbrecher, J. P. Olson, D. Englund, and J. Carolan, "Quantum optical neural networks," *NPJ Quantum Inf.*, vol. 5, no. 1, Dec. 2019, Art. no. 60.

[4] M. Ohzeki, S. Okada, M. Terabe, and S. Taguchi, "Optimization of neural networks via finite-value quantum fluctuations," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 9950.

[5] G. Carleo, Y. Nomura, and M. Imada, "Constructing exact representations of quantum many-body systems with deep neural networks," *Nature Commun.*, vol. 9, no. 1, Dec. 2018, Art. no. 5322.

[6] J. Venderley, V. Khemani, and E.-A. Kim, "Machine learning out-of-equilibrium phases of matter," *Phys. Rev. Lett.*, vol. 120, no. 25, Jun. 2018, Art. no. 257204.

[7] M. Y. Niu, I. L. Chuang, and J. H. Shapiro, "Qudit-basis universal quantum computation using (2) interactions," *Phys. Rev. Lett.*, vol. 120, no. 16, Apr. 2018, Art. no. 160502.

[8] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, pp. 195–202, Sep. 2017.

[9] C. Chen, D. Dong, H.-X. Li, J. Chu, and T.-J. Tarn, "Fidelity-based probabilistic Q-Learning for control of quantum systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 920–933, May 2014.

[10] Z. Li, X. Liu, N. Xu, and J. Du, "Experimental realization of a quantum support vector machine," *Phys. Rev. Lett.*, vol. 114, no. 14, Apr. 2015, Art. no. 140504.

[11] P. Rebentrost, M. Mohseni, and S. Lloyd, "Quantum support vector machine for big data classification," *Phys. Rev. Lett.*, vol. 113, no. 13, Sep. 2014, Art. no. 130503.

[12] R. Chao and B. W. Reichardt, "Quantum error correction with only two extra qubits," *Phys. Rev. Lett.*, vol. 121, no. 5, Aug. 2018, Art. no. 050502.

[13] D. Poulin, A. Kitaev, D. S. Steiger, M. B. Hastings, and M. Troyer, "Quantum algorithm for spectral measurement with a lower gate count," *Phys. Rev. Lett.*, vol. 121, no. 1, Jul. 2018, Art. no. 010501.

[14] S. Pallister, N. Linden, and A. Montanaro, "Optimal verification of entangled states with local measurements," *Phys. Rev. Lett.*, vol. 120, no. 17, Apr. 2018, Art. no. 170502.

[15] X.-D. Cai, D. Wu, Z.-E. Su, M.-C. Chen, X.-L. Wang, L. Li, N.-L. Liu, C.-Y. Lu, and J.-W. Pan, "Entanglement-based machine learning on a quantum computer," *Phys. Rev. Lett.*, vol. 114, no. 11, Mar. 2015, Art. no. 110504.

[16] C. Rouze and N. Datta, "Finite blocklength and moderate deviation analysis of hypothesis testing of correlated quantum states and application to classical-quantum channels with memory," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 593–612, Jan. 2018.

[17] Murphy, Y.-Z. Niu, and, "Qubit-basis universal quantum computation using interactions," *Phys. Rev. Lett.*, vol. 120, Oct. 2018, Art. no. 160502.

[18] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, and L. Cronin, "Controlling an organic synthesis robot with machine learning to search for new reactivity," *Nature*, vol. 559, no. 7714, pp. 377–381, Jul. 2018.

[19] A. Mott, J. Job, J.-R. Vlimant, D. Lidar, and M. Spiropulu, "Solving a higgs optimization problem with quantum annealing for machine learning," *Nature*, vol. 550, no. 7676, pp. 375–379, Oct. 2017.

[20] C. Guo, Z. Jie, W. Lu, and D. Poletti, "Matrix product operators for sequence-to-sequence learning," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 98, no. 4, Oct. 2018, Art. no. 042114

[21] S. Waeldchen, J. Gertis, E. T. Campbell, and J. Eisert, "Renormalizing entanglement distillation," *Phys. Rev. Lett.*, vol. 116, no. 2, Jan. 2016, Art. no. 020502.

[22] P. Zhang, J.-B. Niu, Z. Su, and, "End-to-end quantum-like language models with application to question answering," in *Proc. 32nd AAAI Conf. Artif. Intel. (AAAI)*, 2018, pp. 2–7.

[23] X. Yang and K. Mao, "Task independent fine tuning for word embeddings," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 885–894, Apr. 2017.

[24] R. L. Cilibrasi and P. M. B. Vitanyi, "The Google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.

[25] J. Pennington, M. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Int. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[26] J. Wieting, M. Bansal, K. Gimpel, and D. Roth, "From paraphrase database to compositional paraphrase model and back," *Trans. Assoc. Comput. Linguistics*, vol. 3, pp. 98–104, Oct. 2015.

[27] R. Kiros, "Skip-thought vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 3294–3302.

[28] T. Kenter, A. Borisov, and M. D. Rijke, "Siamese CBOW: Optimizing word embeddings for sentence representations," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 941–951.

[29] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Towards universal paraphrastic sentence embeddings," in *Proc. Int. Conf. Learn. Represent.*, vol. 2016, pp. 1–19.

[30] S. Wang, J. Zhang, and C. Zong, "Learning sentence representation with guidance of human attention," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 4137–4143.

[31] M. Iyyer, V. Manjunatha, J. Boyd-Graber, H. D. Iii, "Deep unordered composition rivals syntactic methods for text classification," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics*, 2015, pp. 1681–1691.

[32] F. A. Gers, and N. N. Schraudolph, "Learning precise timing with LSTM recurrent networks," *J. Mach. Learn. Res.*, vol. 3, no. 1, pp. 115–143, 2003.

[33] Z. Guan, X. Liu, L. Wu, J. Wu, R. Xu, J. Zhang, and Y. Li, "Cross-lingual multi-keyword rank search with semantic extension over encrypted data," *Inf. Sci.*, vol. 514, pp. 523–540, Apr. 2020.

[34] A. J. M. Traina, S. Brinis, and G. V. Pedrosa, "Querying on large and complex databases by content: Challenges on variety and veracity regarding real applications," *Inf. Syst.*, vol. 86, pp. 10–27, Oct. 2019.

[35] M. S. Ramada, J. C. da Silva, and P. de Sá Leitáo-Jánior, "From keywords to relational database content: A semantic mapping method," *Inf. Syst.*, vol. 88, Feb. 2020, Art. no. 101460.

[36] A. Mahtab, D. Chahna, M. Robert, and, "Multilingual semantic textual similarity using multilingual word representations," in *Proc. Int. Conf. Semantic Comput.* 2020, pp. 194–198, doi: 10.1109/ICSC.2020.00040.

[37] M. Matej and P. Jaroslav, "Accuracy of unit under test identification using latent semantic analysis and latent Dirichlet allocation," in *Proc. IEEE 15th Int. Sci. Conf. Inform.*, 2019, pp. 161–166, doi: 10.1109/Informatics47936.2019.9119262.

[38] S. Abdul-Rauf, H. Schwenk, P. Lambert, and M. Nawaz, "Empirical use of information retrieval to build synthetic data for SMT domain adaptation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 745–754, Apr. 2016.
[39] M. Shi, Y. Tang, and J. Liu, "Functional and contextual attention-based LSTM for service recommendation in mashup creation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 5, pp. 1077–1090, May 2019.
[40] Z. Li, H. Lin, W. Zheng, M. M. Tadesse, Z. Yang, and J. Wang, "Interactive self-attentive siamese network for biomedical sentence similarity," *IEEE Access*, vol. 8, pp. 84093–84104, 2020.
[41] J. Liang, L. Jiang, L. Cao, Y. Kalantidis, L.-J. Li, and A. G. Hauptmann, "Focal visual-text attention for memex question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1893–1908, Aug. 2019.
[42] R. Zhao and K. Mao, "Fuzzy Bag-of-Words model for document representation," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 794–804, Apr. 2018.
[43] H. Choi and H. Lee, "Multitask learning approach for understanding the relationship between two sentences," *Inf. Sci.*, vol. 485, pp. 413–426, Jun. 2019.
[44] A. Skabar and K. Abdalgader, "Clustering sentence-level text using a novel fuzzy relational clustering algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 62–75, Jan. 2013.
[45] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365
[46] Z. Yang, Z. Dai, Y. Yang, and, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5733–5763.
[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[48] Z. Lan, M. Chen, S. Goodman, and, "ALBERT: A lite BERT for self-supervised learning of language representations," Tech. Rep., 2019.
[49] G. Lample and A. Conneau, "Cross-lingual language model pretraining," 2019, *arXiv:1901.07291*. [Online]. Available: http://arxiv.org/abs/1901.07291
[50] Y.-H. Liu, M.Ott, N. Goyal, and, "RoBERTa: A robustly optimized BERT pretraining approach," Tech. Rep., 2019.
[51] L. Dong, N. Yang, W. Wang, and, "Unified language model pre-training for natural language understanding and generation," Tech. Rep., 2019.
[52] H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, S. Piao, J. Gao, M. Zhou, and H.-W. Hon, "UniLMv2: Pseudo-masked language models for unified language model pre-training," 2020, *arXiv:2002.12804*. [Online]. Available: http://arxiv.org/abs/2002.12804
[53] S. Minaee, N. Kalchbrenner, E. Cambria, and, "Deep learning based text classification: A comprehensive review," Tech. Rep., 2020.
[54] A. Radford, J. Wu, R. Child, and, "Language models are unsupervised multitask learners," *Open AI Blog.*, vol. 1, no. 8, 2019, pp. 1–9.
[55] [Online]. Available: http://code.google.com/archive/p/word2vec
[56] [Online]. Available: http://groups.google.com/group/STS-semeval
[57] E. Agirre, D. Cer, M. Diab, and, "SemEval-2012 Task 6: A pilot on semantic textual similarity," in *Proc. Conf. Lexical Comput. Semantics*, 2012, pp. 385–393.
[58] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe, "SemEval-2014 task 10: Multilingual semantic textual similarity," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 81–91.
[59] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, and J. Wiebe, "SemEval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability," in *Proc. 9th Int. Workshop Semantic Eval. (SemEval)*, 2015, pp. 252–263.
[60] E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe, "SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 497–511.

YAN YU received the bachelor's degree in physics from Neijiang Normal University, in 2006, and the master's degree in theoretical physics from Guangxi Normal University, in 2012. She is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Chongqing University of Posts and Telecommunications. She is also a Researcher with the College of Mobile Telecommunications, Chongqing University of Posts and Telecom. Her main research interests include machine learning, nature language processing, and quantum computation.

DONG QIU was born in Suining, Sichuan, China, in 1977. He received the bachelor's degree in mathematics from the Sichuan Institute of Education, in 2004, and the Ph.D. degree in mathematics from the University of Electronic Science and Technology of China, in 2009. He is currently a Professor with the Chongqing University of Posts and Telecommunications, Chongqing, China, where he is also a Deputy Director with the Institute of Applied Mathematics. His research interests include fuzzy sets and natural language processing.

RUITENG YAN received the bachelor's degree in electronic information science and technology from the College of Mobile Telecommunications, Chongqing University of Posts and Telecom, in 2014. He is currently pursuing the master's degree with the School of Computer Science and Technology, Chongqing University of Posts and Telecommunications. His main research interests include machine learning, nature language processing, and sentiment analysis.