IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 17, 2020, accepted November 28, 2020, date of publication December 3, 2020,
date of current version December 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3042183

Learning Multimodal Word Representations by Explicitly Embedding Syntactic and Phonetic Information

WENHAO ZHU, SHUANG LIU, CHAOMING LIU, XIAOYA YIN, AND XIAPING XV
Shanghai University, Shanghai 201900, China

Corresponding author: Shuang Liu (josby@foxmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61572434 and Grant 91630206, in part by the Shanghai Science and Technology Committee under Grant 19DZ2204800, and in part by the National Key Research and Development Program of China under Grant 2017YFB0701501.

ABSTRACT Word embedding (i.e., word representation) transforms words into computable mathematical expressions (usually vectors) according to semantics. Compared with human semantic representation, these purely text-based models are severely deficient because they lack perceptual information attached to the physical world. This observation promotes the development of multimodal word representation models. Multimodal models have been proven to outperform text-based models on learning semantic word representations, and almost all previous multimodal models only focus on introducing perceptual information. However, it is obvious that syntactic information can effectively improve the performance of multimodal models on downstream tasks. Therefore, this article proposes an effective multimodal word representation model that uses two gate mechanisms to explicitly embed syntactic and phonetic information into multimodal representations and uses supervised learning to train the model. We select Chinese and English as examples and evaluate the model using several downstream tasks. The results show that our approach outperforms the existing models. We have made the source code of the model available to encourage reproducible research.

INDEX TERMS Word representation, multimodal word representation, natural language processing, supervised learning.

I. INTRODUCTION

Word embedding is often used in natural language processing (NLP) tasks such as machine translation [59], text classification [1], and dialogue systems [50]. There are various word embedding models, such as word2vec [56], GloVe [20], etc. Well-performing word embedding should reflect semantics accurately. At present, most popular methods for learning word embeddings are based on the distributional hypothesis, which utilizes cooccurrence statistics from massive text datasets. However, the process of humans understanding semantics is known in psycholinguistics as language comprehension [16]. Humans are first stimulated by perceptual information (text, sound, etc.), then extract implicit syntactic information from the brain, and finally use the information from their brain to reprocess the perceptual information and understand semantics. Therefore, compared to human semantic representation, these purely text-based models are severely deficient because they lack perceptual information attached to the physical world. This observation has led to the development of multimodal word representation models that utilize both linguistic (e.g., text) and perceptual information (e.g., images and audio). Such models can learn better semantic word representations than text-based models, as evidenced by a range of evaluations [8], [33].

A typical example is that the meaning of concrete words, such as "*bird*" and "*thunder*" are mostly learned from perceptual experiences of seeing, touching and listening. In contrast, more abstract words, such as "*obscure*" and "*lovely*", are less associated with perceptual modalities and act as relatively fixed parts in the sentence structure. According to different types of words, information from different modalities contributes differently to the meaning of words, which has been found in cognitive psychology [21], [22] and computational experiments [10].

However, the existing multimodal models focus on the processing of perceptual information and ignore the introduction of syntactic information. Syntactic information, such as

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

part of speech, refers to the results obtained by dividing the combinatorial relations between words in a sentence according to specific standards. Recently, Vashishth *et al.* [57] and Wang *et al.* [52] completely addressed the word cooccurrence information collected implicitly based on the distributional hypothesis as syntactic information. This approach relies heavily on continuous context, which is the integrity of the training corpus. For target words with multiple parts of speech, if the contexts of a specific part of speech are abundant but their occurrences in training data are low or they do not appear, then the corresponding semantics will not exist in the embedding of the obtained word representation. For example, if the word representation for "*break*" does not have the semantic corresponding to the noun but only the semantic corresponding to the verb, then it is clearly not ideal. For low frequency words, it is more difficult to obtain syntactic information through the distributional hypothesis.

These factors inspire us to build a multimodal word representation model that can embed syntactic and perceptual information effectively, and the model is called MSP. To this end, two fusion mechanisms have been added to the MSP: a modality-specific gate and a language-specific gate. After constructing the perceptual and syntactic representations, the modality-specific gate uses the seq2seq neural network [2], [32], [35] to explicitly embed syntactic and phonetic information in word representations and train the model based on the supervised method. The second mechanism is a language-specific gate. It uses dynamic fusion methods [52] to assign fusion weights to each modality to increase the adaptability of MSP to different language groups. The reason is that in MSP, phonetic information acts as perceptual information while different languages have different emphases on phonetic information. For example, phonetic languages (such as English) are more dependent on phonetic information than ideographic languages (such as Chinese). In addition, extensive analysis was conducted to clarify the principles of the proposed method. In summary, we have two major contributions:

- We propose the multimodal word representation model called MSP. Compared with the existing word embedding models, MSP explicitly embeds syntactic and phonetic information in the model, simulates multimodal information fusion through two gate mechanisms, and obtains a multimodal word representation model with excellent performance through supervised training. The core idea of this model is that it uses supervised training to learn a set of general language information fusion rules.
- The use of syntactic information can significantly improve the performance of the multimodal word representation model. On various NLP tasks, we use multiple word representation models and pre-trained language models as baselines to compare the performance and set MSP- with no processing of syntactic information as a control. The task results confirm this conclusion.

II. RELATED WORKS

Researchers have been working on building multimodal representation models for many years, most of which can be divided into two types.

A. JOINT TRAINING MODELS

These models build multimodal representations with raw inputs of both linguistic and perceptual resources. The recently introduced work is an extension of the skip-gram model [56]. For instance, Hill *et al.* [10] propose a corpus fusion method that inserts the perceptual features of a word in the training corpus, which is then used to train the skip-gram model. Lazaridou *et al.* [31] proposed the MMSkip model, which injects visual information in the process of learning linguistic representations by adding a max-margin objective function to minimize the distance between linguistic vectors and visual vectors. The joint training methods implicitly propagate perceptual information to word representations and simultaneously learn multimodal representations. However, the abovementioned models do not introduce syntactic information. This weakens the effect of introducing perceptual information and consequently leads to only limited improvement. Vashishth *et al.* [53] incorporate syntactic and semantic information in word representations by using graph convolutional networks, and explicit embedded syntactic information effectively improves the performance of the model; however, this model does not introduce perceptual information.

B. SEPARATE TRAINING MODELS

These models independently learn linguistic and perceptual representations and integrate them afterwards. The simplest approach is concatenation, which fuses linguistic and visual vectors by concatenating them. Concatenation has been proven to be effective in learning multimodal models [8], [10], [11]. Variations of this method apply transformation and dimension reduction techniques, including the singular value decomposition (SVD) [8] and canonical correlation analysis (CCA), to the concatenation result [10]. In addition, Vashishth *et al.* [53] and Silberer *et al.* [54] use a stacked autoencoder to learn multimodal representations by embedding linguistic and visual inputs into a common space with the objective of reconstructing the individual inputs. However, the abovementioned methods can only generate multimodal representations of those words that have image information, thus drastically reducing the multimodal vocabulary. Wang *et al.* [52] build a multimodal model that can dynamically fuse semantic representations of different modalities according to different types of words. In the last two years, the research of constructing multimodal word representation using phonetic information has also been carried out. Zhu *et al.* [58] propose enhanced double-carrier word representation via phonetics and writing. It trained written embedding based on phonetic embedding and the final word representation fuses writing and phonetic embedding. Zhu *et al.* [63] use a synchronized way that adopts an

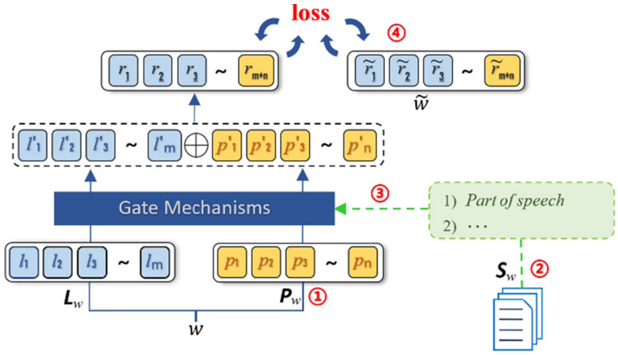
loss
④
r_1 r_2 r_3 ~ r_{m+n}
$\tilde{r}_1$ $\tilde{r}_2$ $\tilde{r}_3$ ~ $\tilde{r}_{m+n}$
$\widetilde{w}$
l'_1 l'_2 l'_3 ~ l'_m ⊕ p'_1 p'_2 p'_3 ~ p'_n
Gate Mechanisms
③
1) Part of speech
2) ···
l_1 l_2 l_3 ~ l_m
p_1 p_2 p_3 ~ p_n
S_w ②
L_w
P_w ①
w

FIGURE 1. The four numbers correspond to the four steps of the method. L_W, P_W and S_W are the linguistic representation, perceptual representation and syntactic information of the target word w, respectively. In the fourth step, w and $\widetilde{w}$ are semantic relational word pairs.

attention model to utilize both text and phonetic perceptual information in unsupervised learning tasks. In terms of the two types of models discussed in this section, MSP belongs to separate training model.

Based on the existing researches, the above methods are all effective methods to generate multimodal word representation. However, no matter the joint training model or the separate training model, most of them only focus on the introduction of a class of modality information during the learning process. In contrast, MSP uses gate mechanisms to introduce perceptual information and syntactic information in the one model.

III. PROPOSED METHOD

Fig. 1 shows the framework of our proposed MSP, which contains four stages:

- Build the perceptual representation — Language comprehension begins with receiving perceptual stimuli. Most linguists believe that sound is the primary perceptual form of language, so the model processes the phonetic feature of words and treats the result as a perceptual representation.
- Construct the syntactic information — Janda [27] have experimentally demonstrated that syntactic information plays an irreplaceable role in language comprehension. In MSP, for each word, we construct the probability distribution of the part of speech as the syntactic information.
- Modality-specific gates and language-specific gates are used to explicitly embed syntactic information in training and fuse the linguistic representation and perceptual representation. We employ the GloVe and word2vec vectors as our linguistic representations, which are trained using global word cooccurrence statistics.
- We design the objective function and train the MSP model using supervised learning.

A. CONSTRUCT PERCEPTUAL REPRESENTATION

The goal of this phase is to build the perceptual representation P_w. According to linguistics, different perceptual information of the word considers different information on concepts. For example, image may include information such as shape and color. By contrast, voice contain the concept of information is less, but the phonetic context and the text context can't be regarded as duplicated, they are a complementary relationship that provides a richer semantic for each other. For example, in the case of disambiguation, "*minute*" has two meanings. When the pronunciation of "*minute*" is ['mɪnɪt], it indicates a time unit, and when it is pronounced [maɪ'nju:t], it means tiny. For words with similar sounds and different meanings, the text can provide richer semantics for the model (such as *ship* and *sheep*), and the difference in their writing helps us distinguish the different meanings of the two words. Moreover, while every word has a corresponding pronunciation, images do not have this natural advantage. In this article, we choose sound, which is the primary perceptual stimulus, as the perceptual information; therefore, the model needs to obtain the phonetic representation of words. Specifically, the automatic segmentation of spoken words has been successfully trained and reported previously [3], [6]. The training audio corpus in the present work has been previously segmented into phonetic words. We use the Mel-scale Frequency Cepstral Coefficient (MFCC) method — a common approach to obtain the phonetic features of the audio — to convert the speech frames of words into vectors. Those vectors contain a considerable amount of noise, such as background noise and speaker characteristics; however, what we want to obtain is the phonetic structure [61], which is not changed by the environment or the speaker. To disentangle the phonetic structure and noise, we use an end-to-end approach to process phonetic vectors and obtain the results as perception representations [58].

B. CONSTRUCT SYNTACTIC REPRESENTATION

MSP uses part of speech (POS) information to construct syntactic representations. Part of speech is the most common syntactic structure. It is the result of the classification of words based on grammatical features (including syntactic functions and morphological changes) and helps people to collocate and understand the meanings of words. Modern English words can be divided into fourteen parts of speech, but only five are used most often — nouns, verbs, prepositions, adverbs and adjectives. In this model, GCNW uses WordNet to structure syntactic information. WordNet is an English dictionary based on cognitive linguistics in which the relationship between words is human annotated [14]. It can label the POS tag of a word in each specific context. Handling polysemy is the key to constructing POS features. The problem of obtaining the POS tag can be formulated as $p = F(w, c)$, where F is the mapping function that obtains the corresponding POS tag p based on the target word w and specific context c.

First, we use WordNet to label the POS tag of each word in the corpus. Note that the same word may be labeled differently in different contexts. Next, for target word w in the corpus, we count the occurrence Occ_p^w of each POS p.

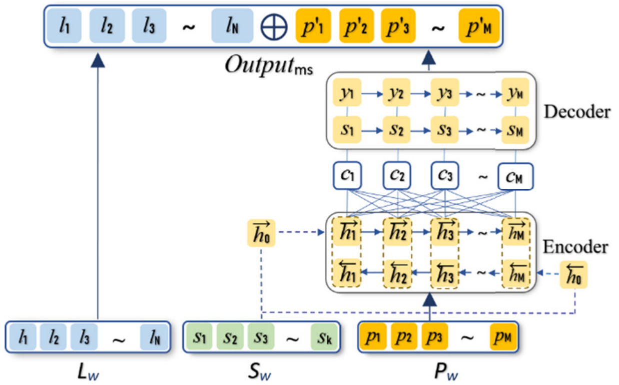
l_1 l_2 l_3 ~ l_N ⊕ p'_1 p'_2 p'_3 ~ p'_M
$Output_{ms}$
y_1 y_2 y_3 ~ y_M
s_1 s_2 s_3 ~ s_M
Decoder
c_1 c_2 c_3 ~ c_M
$\vec{h}_0$
$\vec{h}_1$ $\vec{h}_2$ $\vec{h}_3$ ~ $\vec{h}_M$
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ ~ $\overleftarrow{h}_M$
$\overleftarrow{h}_0$
Encoder
l_1 l_2 l_3 ~ l_N
S_1 S_2 S_3 ~ S_k
p_1 p_2 p_3 ~ p_M
L_w
S_w
P_w

FIGURE 2. Overview of the modality-specific gate, where L_W, P_W and S_W represent the linguistic representation, perceptual representation and syntactic feature, respectively, of the target word w.

In equation (1), m is the total number of times that word w has occurred in the corpus,

$$Occ_p^w = \sum\nolimits_{i=0}^{m} T(F(w,c))T(F(w,c)) = \begin{cases} 1, & F(w,c) = p \\ 0, & F(w,c) \neq p \end{cases} \tag{1}$$

Then, Occ_p^w is normalized; thus, the probability distribution of part of speech of the word w is obtained,

$$Pro_p^w = \frac{Occ_p^w}{\sum_p' Occ_{p'}^w} p, p' \in \{noun, verb, prep, adv, adj\} \tag{2}$$

Finally, we treat the probability distribution of the POS as the syntactic information of word w and construct it into a feature vector that is used in the next phase.

C. GENERATE REPRESENTATION IN MSP

In this phase, the model explicitly uses two fusion mechanisms, fusing linguistic representation and perceptual representation, to introduce syntactic information in training.

1) MODALITY-SPECIFIC GATE

To simulate the role of syntactic information in language comprehension, namely, the reprocessing of perceptual information, we add a modality-specific gate to the model. The modality-specific gate is basically a seq2seq model based on the attention mechanism [2], [55], which is a training method that transforms sequences in different domains. As shown in Fig. 2, the seq2seq model consists of two parts — an encoder and a decoder. The encoder generates intermediate semantic c using the hidden state h of the bidirectional RNN. $\vec{h}_i$ represents the hidden layer state of the forward RNN, $\overleftarrow{h}_i$ represents the hidden layer state of the reverse, and the two are spliced to obtain h_i, namely, $h_i = \left[\vec{h}_i : \overleftarrow{h}_i\right]$. The decoder uses long short-term memory networks (LSTM) [39] to decode c to obtain output sequence y. For the output sequence $[y_1, y_2 \ldots, y_{i-1}]$ and the current i^{th} dimension input X, y_i can be expressed as:

$$y_i = p(y_i \mid y_1, \ldots, y_{i-1}, X) = g(y_{i-1}, s_i, c_i) \tag{3}$$

For the M-dimension phonetic representation used as the input, y_i is determined by three factors as $g(y_{i-1}, s_i, c_i)$ the hidden state s_i at the i^{th} dimension, the intermediate semantic vector c_i, and the output y_{i-1} at the i-1$^{\text{th}}$ dimension, where s_i is related to the hidden state s_{i-1}, and c_i is obtained by equation (4). In equation (4), e_{ij} is the alignment model in the attention mechanism and is used to measure the influence of the j^{th} dimension information of the input sequence on the i^{th} dimension information of the output sequence.

$$c_i = \sum_{j=1}^{M} \frac{exp(e_{ij})}{\sum_{k=1}^{M} exp(e_{ik})} \cdot h_j \tag{4}$$

The encoder needs to initialize the parameters during training at which time the effect of the syntactic information is reflected. The model uses the syntactic feature vector of word w to initialize the parameters $\vec{h}_i$ and $\overleftarrow{h}_i$ in training. The network output y_i of the end-to-end type is the probability distribution. Softmax calculations are performed on each dimension of the sequence $[y_1, y_2, \ldots, y_M]$, and p' whose dimension is equal to the input phonetic representation is obtained. Finally, a linguistic representation and p' are concatenated to obtain $Output_{\text{ms}}$.

2) LANGUAGE-SPECIFIC GATE

In linguistics, languages can be divided into ideographic languages and phonological languages according to the dependence of text and sound. Ideographic languages (Chinese, etc.) focus more on text than phonological languages (English, etc.). The use of neural networks to dynamically fuse different modalities has been proven to be effective [52]. Based on this observation, in order to improve the applicability of the model, we add the language-specific gate to assign weights for the linguistic representation and the perceptual representation. In the joint training phase, the model uses a neural network to simulate the current language's dependence on different modalities, and the weighted linguistic representation and weighted phonetic representation will be concatenated to obtain $Output_{\text{ls}}$.

3) JOINT TRAINING

In this phase, the model will integrate the outputs of the two gates. According to the literature [8], [52], dynamic weighted fusion is an effective method. Thus, we add a set of variable weights $\{w_{ms}, w_{ls}\}$ to the network to weight the outputs and superimpose the results to generate $Output_{\text{MSP}}$ as equation (5).

$$Output_{\text{MSP}} = \sum_{i \in A} Output_i \cdot w_i, \quad A \in \{ms, ls\} \tag{5}$$

To train the model, WordNet is introduced as the training dataset. WordNet can search the synonym set corresponding to the target word according to semantic conditions, and the semantic similarity is also human annotated. In the joint training phase, according to equation (6), the model first calculates the mean cosine similarity between MSP representations corresponding to words in the synonym set.

$$\begin{aligned} Similarity\,(w, sw) &= \cos(\theta) = \frac{\vec{w} \cdot \vec{sw}}{\left\|\vec{w}\right\| \times \left\|\vec{sw}\right\|} \\ &= \frac{\sum_i^n w_i \cdot sw_i}{\sqrt{\sum_i^n (w_i)^\wedge 2} \times \sqrt{\sum_i^n (sw_i)^\wedge 2}} \end{aligned} \tag{6}$$

Then, according to the training objective, the model minimizes the loss, namely, the difference between the mean cosine similarity and the human-annotated similarity. The model performs iterative training, during which the MSP representations will be updated with the network.

Suppose the dictionary contains M words, each word w corresponds to N synonyms $\widetilde{w}$, and the human-annotated similarity between w and $\widetilde{w}$ is Sim(w, $\widetilde{w}$). To train the model and learn the network parameters, we minimize the objective function as follows:

$$ss = \sum_m^M \sum_n^N \left\|Similarity\,(w_m, \widetilde{w}_n) - Sim\,(w_m, \widetilde{w}_n)\right\|^\wedge 2 \tag{7}$$

Although WordNet provides a set of annotated synonyms for almost all words, this does not mean that all words can find a synonym set. For some unqualified words, the model deletes them before training.

IV. TASK EVALUATION

A. BASELINE ALGORITHMS

Word2vec is the most common word representation model. It includes two training modes, CBOW and skip-gram. In the tasks, we compare MSP with word2vec implemented with the CBOW structure. **GloVe** [20] is another efficient word representation model that incorporates global word cooccurrence information. **DFM** [52] is a multimodal model that uses three novel dynamic fusion methods to assign importance weights to each modality, and the weights are learned under the weak supervision of word association pairs. **DCWE** [58] is enhanced double-carrier word representation model via phonetics and writing, and it trained written representation based on phonetic representation and the final word representation fuses text and phonetic embedding. **DPWR** [63] is trained in a synchronized way that adopts an attention model to utilize both linguistic and phonetic information in unsupervised learning tasks. **SynGCN** [57] incorporates syntactic and semantic information in word embeddings by using graph convolutional networks. **GloVe-ph** is a multiple information connection model that directly concatenates the linguistic representation and the perceptual representation. **MSP** is the multimodal word representation model generated by the method described in this article in which the linguistic representation is represented by GloVe. **MSP-w2v** changes the linguistic representation in MSP from GloVe to word2vec. **MSP-** removes the modality-specific gate in MSP to verify the effectiveness of the method described in this article. We also compare the pre-trained language models, including ELMo and BERT, on tasks; however, considering the constraints of the pre-trained language model on task types, they are only used for text classification task. **ELMo** [36] is a pretrained language model that trains a model with multiple BiLSTM layers, and the output of the model is a sentence representation. **BERT** [19] is a pretrained transformer network model. In the comparative experiment, the model consists of 12 layers, 768 hidden layers, 12 heads, and 110 M parameters.

B. EXPERIMENTAL SETUP

For the English linguistic representation, we use the 300-dimensional GloVe and word2vec, which are trained on the Common Crawl corpus consisting of 840 B tokens and a vocabulary of 2.2 M words. For the Chinese linguistic representation, we also use the 300-dimensional GloVe and word2vec, and those vectors are trained on the Wikipedia data set and web news corpus and use *Jieba*[1] for word segmentation. The dimension of the perceptual representation in the MSP is set to 100. To control the dimensions, other word representation models used for comparison are also retrained according to the dimensions of the MSP. The MSP model is implemented by using TensorFlow. We set the initial learning rate to 0.02 and the batch size to 100, and we randomly initialize the parameters of the model according to a normal distribution. We set the minimum word frequency to 5 by default. If a word appears in the document less than 5 times, it is discarded. The related data and code will be posted on GitHub for replication[2].

We use four intrinsic and two extrinsic evaluation methods to evaluate MSP. Intrinsic evaluation methods include concept categorization task, word similarity task, word analogy task and part of speech tagging task. Those methods focus on measuring lexical internal pattern information, such as semantic information. However, a language model that performs well in an intrinsic evaluation does not necessarily produce similar performance in an extrinsic evaluation. Therefore, this chapter added text classification task and text similarity task as extrinsic evaluation methods to verify the applicability of MSP to different types of tasks.

C. CONCEPT CATEGORIZATION TASK

1) DATASET AND EVALUATION CRITERION

Concept categorization involves grouping nominal concepts into natural categories. For instance, *computers* and *phones* should belong to the electronic products class. In our experiments, we evaluate the models on the *AP* (Almuhareb, 2006), *Battig* (Baroni and Lenci, 2010), *BLESS* (Baroni and Lenci, 2011), and *ESSLI* (Baroni *et al.*, 2008) datasets. We calculate the classification accuracy $\sigma\%$ to evaluate the models, and a higher accuracy corresponds to a better model.

[1] https://github.com/fxsjy/jieba

[2] https://github.com/JayJosby/MSP

TABLE 1. Task results σ% of the concept categorization task.

Model	*AP*	*Battig*	*BLESS*	*ESSLI*
Word2vec	61.22	59.52	60.56	55.78
GloVe	62.98	61.15	62.37	57.02
DFM	67.12	61.24	65.44	60.03
DCWE	68.71	62.80	65.92	59.89
DPWR	65.99	63.47	65.74	60.02
SynGCN	68.03	65.66	65.75	60.57
GloVe-ph	54.27	51.57	57.66	47.78
MSP-w2v	68.44	65.91	64.84	59.26
MSP-	61.89	59.34	61.72	49.96
MSP	**69.97**	**66.08**	**67.23**	**61.28**

TABLE 2. Word similarity datasets.

Language	English				Chinese	
Dataset	*WordSim353*	*MC30*	*MTurk287*	*MTurk771*	*WS240*	*WS296*
Size	353	30	287	771	240	296

2) RESULTS AND DISCUSSION

Table 1 lists the results of the concept categorization task. Overall, we found that MSP is superior to existing word representation methods in all four data sets, and MSP-w2v also performs well. On average, we obtain an approximately 1.4% absolute increase in performance on the concept categorization task compared to the best performing baseline. The concept classification task needs to calculate the topic similarity (topically related words) between different words rather than the functional similarity (in place substitutable words). The supervised learning method used by MSP in the training captures the topic similarity of words by utilizing the synonymous relationship between words, which provides advantages for the performance of the model on the task.

D. WORD SIMILARITY TASK

1) DATASET AND EVALUATION CRITERION

We used *WordSim-353* (L. Finkelstein, 2010), *MC30* (S. Hassan and R. Mihalcea. 2009), *Mturk287* (G. Halawi *et al.*, 2012), *Mturk771* (G. Halawi *et al.*, 2012), *WS-240* and *WS-296* as the evaluation datasets. All datasets contained a list of word pairs along with human-annotated similarities. Table 2 lists the information of those datasets.

The task uses the cosine similarity between a pair of word representations as the similarity of semantics and employs the Pearson correlation ρ to evaluate the relation between the human-annotated semantic similarity and the cosine similarity. A larger ρ indicates a higher correlation and a better model.

2) RESULTS AND DISCUSSION

The results are listed in Table 3 and Table 4. For English, when the Pearson coefficient ρ is the evaluation criterion, MSP and MSP-w2v perform the best for all four datasets at 1.1∼5.9% higher than the state-of-the-art baseline models. For Chinese, MSP performs the best for both datasets. These results show that MSP generated better performances than the existing models. However, because the word similarity information is introduced into the objective function, the results of the word similarity task cannot be used alone to prove the good performance of MSP. The addition of the word similarity task is intended to validate the applicability of the model over different language sets.

TABLE 3. The results of the word semantic similarity task in English.

Model	*WordSim353*	*MC30*	*MTurk287*	*MTurk771*
Word2vec	0.657	0.784	0.716	0.675
GloVe	0.671	0.751	0.727	0.672
DFM	0.687	0.789	0.731	0.697
DCWE	0.690	0.793	0.729	0.723
DPWR	0.679	0.799	0.735	0.710
SynGCN	0.674	0.797	0.720	0.702
GloVe-ph	0.566	0.587	0.576	0.547
MSP-w2v	0.699	0.798	0.736	0.737
MSP-	0.669	0.776	0.712	0.688
MSP	**0.709**	**0.810**	**0.742**	**0.761**

TABLE 4. The results of the word semantic similarity task in Chinese.

Model	*WordSim-240*	*WordSim-296*
Word2vec	0.539	0.539
GloVe	0.546	0.574
DFM	-	-
DCWE	0.571	0.580
DPWR	0.569	0.574
SynGCN	0.580	0.577
GloVe-ph	0.524	0.506
MSP-w2v	0.573	0.590
MSP-	0.534	0.547
MSP	**0.595**	**0.603**

Further analysis shows that the task performances are much lower than those of the text-based models when the linguistic and perceptual representations are directly concatenated. This indicates that the direct concatenating representations increase the information of the word representation, but this approach is not applicable to the subsequent tasks.

E. WORD ANALOGY TASK

1) DATASET AND EVALUATION CRITERION

This task is to predict word b_2 given three words a_1, a_2, and b_1 such that the relation b_1: b_2 is the same as the relation a_1: a_2. We compare models on *SemEval-2012* (Jurgens *et al.*, 2012) and *MSR* (Mikolov *et al.*, 2013c) using the Pearson correlation.

2) RESULTS AND DISCUSSION

The evaluation results on the word analogy task are summarized in Table 5. Overall, we find that MSP outperforms all the existing word representation models.

Compared to the best performing baseline model, on average, MSP obtains an approximately 3.6% increase in performance. The results demonstrate that the learned

TABLE 5. The results of the word analogy task.

Model	*SemEval-2012*	*MSR*
Word2Vec	0.487	0.521
GloVe	0.520	0.549
DFM	0.535	0.545
DCWE	0.529	0.571
DPWR	0.540	0.568
SynGCN	0.529	0.564
GloVe-ph	0.423	0.480
MSP-w2v	0.537	0.566
MSP-	0.493	0.535
MSP	**0.556**	**0.617**

TABLE 6. The results of the part of speech tagging task.

Model	*TreeBank*
word2vec	94.2
GloVe	94.1
DFM	95.0
DCWE	94.9
DPWR	94.7
SynGCN	96.1
GloVe-ph-	87.8
MSP-w2v	95.9
MSP-	94.8
MSP	**96.4**

representations from MSP more effectively capture the semantic and syntactic properties of words.

F. PART-OF-SPEECH TAGGING TASK

1) DATASET AND EVALUATION CRITERION

Part-of-speech (POS) tagging aims at associating with each word, a unique tag describing its syntactic role. For evaluating word representation models, we use Lee *et al.*'s LSTM model [64] on *Treebank* POS dataset (Marcus *et al.*, 1994) and evaluate performance with tagging accuracy.

2) RESULTS AND DISCUSSION

Table 6 shows the experimental results of part-of-speech tagging task. Compared with the existing word representation models, MSP has a better performance — MSP gets an excellent result like grammar enhancement model SynGCN, which is 2.2% more accurate than the text-based word representation models and 1.5% more accurate than the multimodal models. The introduction of syntactic information effectively improves the performance of multimodal model.

Combining the results of other intrinsic evaluation tasks, it can be concluded that the word representation generated by the MSP model contain more semantic and syntactic information, and that such information can be used in relevant downstream tasks.

G. TEXT CLASSIFICATION TASK

1) DATASET AND EVALUATION CRITERION

We also perform a text classification task to check our method's applicability. The task is based on several public datasets, including *scale*, *IMDB,* and *Yelp reviews*. The *scale* v1.0 dataset, which we obtained from (Pang and Lee, 2005), is used as the evaluation dataset; and this dataset contains 5004 samples with review texts labeled with 1-4 stars. The *IMDB* data set contains 50,000 film reviews, including 25,000 opinion-filled reviews for training and 25,000 reviews for testing; and these data set can be used for classification. We also use *Yelp reviews* as a dataset, which we obtained from (Zhang *et al.*, 2015). This dataset contains 1,569,264 samples of review texts labeled with 1-4 stars. For the text classification task, we use the mean of the word representations to represent a sentence or document. The text classifier was trained with *LIBLINEAR*[3] [65]. For the corpus that does not distinguish between the training and testing sets, 75% of the characters are selected as the training set, and the remaining 25% are used for testing. We calculate the classification accuracy $\sigma\%$ to evaluate the models

TABLE 7. Accuracy $\sigma\%$ of the text classification with the MSP and word representation models.

Model	*scale*	*IMDB*	*Yelp Reviews*
Word2vec	60.61	60.51	60.76
GloVe	60.98	63.13	61.36
DFM	69.39	67.27	71.40
DCWE	72.72	67.74	73.56
DPWR	70.90	69.29	74.01
SynGCN	74.54	70.66	73.24
GloVe-ph	56.12	52.97	59.66
MSP-w2v	76.71	69.91	76.84
MSP-	66.35	61.34	69.72
MSP	**77.95**	**72.08**	**78.23**

TABLE 8. Accuracy $\sigma\%$ of the text classification with the MSP and pre-trained language models.

Model	*scale*	*IMDB*	*Yelp Reviews*
ELMo	76.54	70.25	75.44
Bert	77.39	**72.97**	76.15
MSP	**77.95**	72.08	**78.23**

2) RESULTS AND DISCUSSION

Table 7 and Table 8 list the results of the text classification task. Compared to other baseline word representation models, MSP performs the best for all datasets, which shows that MSP not only significantly improves the model performance, but it is also applicable to different downstream tasks. Moreover, other models with embedded syntactic information, such MSP-w2v and SynGCN, also perform well. This shows the effectiveness of the introduced syntactic information for this type of task. When compared to the pre-trained language models, the difference between other models' and MSP's performance on the text classification task is slight. However, BERT and other language models are only applicable to tasks

[3] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

TABLE 9. The results of the text similarity task.

Model	*STS*	*SICK*
Word2vec	0.472	0.566
GloVe	0.478	0.559
DFM	0.477	0.570
DCWE	0.480	0.573
DPWR	0.471	0.569
SynGCN	0.491	0.578
GloVe-ph	0.448	0.530
MSP-w2v	0.489	0.581
MSP-	0.471	0.556
MSP	**0.495**	**0.588**

with larger granularity, such as those at the sentence level; and they require extremely large numbers of parameters and training costs. Therefore, MSP has its own advantages in this application.

H. TEXT SIMILARITY TASK

1) DATASET AND EVALUATION CRITERION

The content of text similarity task is to calculate the similarity s_1 of a pair of sentences, and then measure the performance of the model by comparing the difference between the similarity s_1 and the similarity s_2 of manual annotation.

We superimpose the word vectors in the sentence, express the average vector as the sentence representation, and take the cosine similarity between the two sentence vectors as the similarity s_1. Pearson correlation coefficient is used to calculate the correlation between s_1 and the s_2. We experimented with the *SICK* and *STS* datasets. The *SICK* data set contained 9,927 pairs of sentences (4,500 pairs of training sets /4,927 pairs of test sets /500 pairs of validation sets). The *STS* data set consists of 8,628 sentence pairs, divided into training sets (5,749 of training sets /1,500 of test sets /1,379 of verification sets).

2) RESULTS AND DISCUSSION

Table 9 list the results of the text similarity task. According to the results, MSP performs best across all data sets. Compared with text-based word representation and multimodal word representation without introduction of syntactic information, the results obtained by MSP are improved by 0.016 and 0.012 respectively.

Based on the results of the extrinsic evaluation methods in this chapter, it can be concluded that MSP not only performs well in the intrinsic evaluation method, but also gets similar results in the extrinsic evaluation, which indicates that MSP not only can effectively improve the internal mode information represented by words, but also has good applicability for different types of tasks.

V. MODEL ANALYSIS

Compared with the existing word embedding models, MSP achieves a great improvement. Its gate mechanisms effectively integrate multimodal information, which is reflected by its good performance. The MSP consistently performs better than the MSP- model on all task results; and when MSP removed the modality-specific gate, the performance of the model experienced a significant decrease but was still higher than that of GloVe-Ph. This suggests that after the removal of the modality-specific gate, the model loses the reinforcement effect of syntactic information. However, language-specific gates still play a role in adjusting the weights of the modality; and without this mechanism, MSP would completely degenerate to GloVe-ph.

TABLE 10. The weights of the modalities.

	Text	Perception
MSP (English)	0.8225	0.1775
MSP (Chinese)	0.8976	0.1024

For the text classification task, when compared to other text-based models and multimodal models, MSP is still better than MSP- and has the best performances in three datasets. Moreover, the improvement effect is better than those for the other tasks, indicating that the introduced syntactic information plays a role in making MSP more suitable for tasks that utilized syntactic information.

The applicability of MSP to different languages is also quantitatively analyzed. Table 10 presents the combination weights of the linguistic and perceptual representations learned in language-specific gates for English and Chinese. The ratio between the linguistic information and perceptual information was 0.8225:0.1775 for English and 0.8976:0.1024 for Chinese. Linguistic representation has a higher weight for both languages, which indicates that text is more important for carrying information. However, phonetic languages such as English have a stronger dependence on phonetic information than ideographic languages such as Chinese, which is in line with the linguistic viewpoint. The above results indicate the following:

- MSP is a word embedding model with better comprehensive performance because the MSP includes extra multimodal information and uses effective mechanisms to process that information. This is demonstrated in a series of tasks.
- Adding syntactic information can effectively improve the performance of the model. Similar to perceptual information, syntactic information is also needed for building multimodal representations and can effectively improve the performance of the model on downstream tasks.
- MSP is applicable to different languages. The learned weights show clear differences between phonetic and ideographic languages.

VI. CONCLUSION

Based on the observation that almost all previous multimodal models only focus on introducing perceptual information and ignore syntactic information, we propose the new multimodal word representation model MSP. MSP uses two fusion mechanisms to embed explicit syntactic information and phonetic information and uses supervised training to learn performance-enhancing multimodal word representations. Experimental evaluations show that our proposed model achieves substantial gains on all benchmarks. Qualitative analysis further proved the validity and applicability of MSP.

As one of the main research directions related to the development of language representations, the performance of multimodal models depends not only on the source of the perceptual information but also on the method used to incorporate that information. Such an incorporation method should not be limited to the incorporation of only two kinds of information and should also be capable of incorporating information from more than two modes. Future work includes exploring better representations of semantic words by combining information from other modalities. We believe that the multimodal model is of great significance in promoting the development of applications related to natural language processing.

REFERENCES

[1] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016, *arXiv:1607.01759*. [Online]. Available: http://arxiv.org/abs/1607.01759

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Conf. NIPS Adv.*, San Diego, CA, USA, 2017, pp. 5998–6008.

[3] A. J. Anderson, D. Kiela, S. Clark, and M. Poesio, "Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 17–30, Dec. 2017.

[4] A. K Vijayakumar, R. Vedantam, and D. Parikh, "Sound-Word2 Vec: Learning word representations grounded in sounds," 2017, *arXiv:1703.01720*. [Online]. Available: http://arxiv.org/abs/1703.01720

[5] A. Neelakantan, J. Shankar, A. Passos, and A. McCallum, "Efficient non-parametric estimation of multiple embeddings perword in vector space," in *Proc. EMNLP*, 2014, pp. 1059–1069.

[6] D. Kiela and S. Clark, "Learning neural audio embeddings for grounding semantics in auditory perception," *J. Artif. Intell. Res.*, vol. 60, pp. 1003–1030, Dec. 2017.

[7] D. Kiela and S. Clark, "Multi- and cross-modal semantics beyond vision: Grounding in auditory perception," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2461–2470.

[8] E. Bruni, N. K. Tran, and M. Baroni, "Multimodal distributional semantics," *J. Artif. Intell. Res.*, vol. 49, pp. 1–47, Jan. 2014.

[9] F. Hill and A. Korhonen, "Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Stroudsburg, PA, USA, 2014, pp. 255–265.

[10] F. Hill, R. Reichart, and A. Korhonen, "Multi-modal models for concrete and abstract concept meaning," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 285–296, Dec. 2014.

[11] G. Collell, T. Zhang, and M.-F. Moens, "Imagined visual representations as multimodal embeddings," in *Proc. 31st AAAI Conf. Artif. Intell.*, Menlo Park, CA, USA, 2017, pp. 4378–4384.

[12] G. Halawi, G. Dror, E. Gabrilovich, and Y. Koren, "Large-scale learning of word relatedness with constraints," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2012, pp. 1406–1414.

[13] G. Segal, "Representing representations," in *Language and Thought*, P. Carruthers and J. Boucher, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[14] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[15] G. K. Pullum and W. A. Ladusaw, *Phonetic Symbol Guide*. Chicago, IL, USA: Univ. Chicago Press, 1996.

[16] Andrews, "Language comprehension as structure building," *J. Pragmatics*, vol. 26, no. 3, p. 436, 1996.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] J. De Villiers and P. de Villiers, "Linguistic determinism and the understanding of false beliefs," in *Children's Reasoning and the Mind*, P. Mitchell and K. Riggs, Eds. New York, NY, USA: Psychology Press, 2000.

[19] 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

[20] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Stroudsburg, PA, USA, 2014, pp. 1532–1543.

[21] J. Wang, J. A. Conder, D. N. Blitzer, and S. V. Shinkareva, "Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies," *Hum. Brain Mapping*, vol. 31, no. 10, pp. 1459–1468, Oct. 2010.

[22] J. R. Binder, L. L. Conant, C. J. Humphries, L. Fernandino, S. B. Simons, M. Aguilar, and R. H. Desai, "Toward a brain-based componential semantic representation," *Cognit. Neuropsychol.*, vol. 33, nos. 3–4, pp. 130–174, May 2016.

[23] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Charagram: Embedding words and sentences via character n-grams," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1504–1515.

[24] J. P. Turian, L.-A. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," in *Proc. ACL*, 2010, pp. 384–394.

[25] K. Levin, A. Jansen, and B. Van Durme, "Segmental acoustic indexing for zero resource keyword search," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA, Apr. 2015, pp. 5828–5832.

[26] U. Khandelwal, H. He, P. Qi, and D. Jurafsky, "Sharp nearby, fuzzy far away: How neural language models use context," 2018, *arXiv:1805.04623*. [Online]. Available: https://arxiv.org/abs/1805.04623

[27] L. A. Janda, "Cognitive linguistics," *SSRN Electron. J.*, vol. 3, pp. 129–141, 2009.

[28] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," in *Proc. 10th Int. Conf. World Wide Web*, New York, NY, USA, 2001, pp. 406–414.

[29] L. Qiu, Y. Cao, Z. Nie, Y. Yu, and Y. Rui, "Learning word representation considering proximity and ambiguity," in *Proc. AAAI Conf. Artif. Intell.*, Menlo Park, CA, USA, 2014, pp. 1572–1578.

[30] L. Talmy, "Cognitive linguistics," in *Encyclopedia of Language & Linguistics*, K. Brown, Ed. Amsterdam, The Netherlands: Elsevier, 2006.

[31] A. Lazaridou, N. T. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," in *Proc. NAACL*, 2015, pp. 153–163.

[32] T. Liu, K. Wang, L. Sha, B. Chang, and Z. Sui, "Table-to-text generation by structure-aware seq2seq learning," 2017, *arXiv:1711.09724*. [Online]. Available: https://arxiv.org/abs/1711.09724

[33] M. Andrews, G. Vigliocco, and D. Vinson, "Integrating experiential and distributional data to learn semantic representations," *Psychol. Rev.*, vol. 116, no. 3, pp. 463–498, 2009.

[34] M. Hiscock, "Imagery and verbal processes," *PsycCRITIQUES*, vol. 19, p. 487, 1974.

[35] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," 2017, *arXiv:1708.06426*. [Online]. Available: https://arxiv.org/abs/1708.06426

[36] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 2227–2237.

[37] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. NAACL*, Stroudsburg, PA, USA, 2018, pp. 2227–2237.

[38] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proc. Annu. Meeting ACL*, Stroudsburg, PA, USA, 2014, pp. 302–308.

[39] O. Melamud, J. Goldberger, and I. Dagan, "Context2vec: Learning generic context em-bedding with bidirectional LSTM," in *Proc. InCoNLL*, 2016, pp. 1–11.

[40] P. Jin and Y. Wu, "SemEval-2012 task 4: Evaluating Chinese word similarity," in *Proc. Joint Conf. Lexical Comput. Semantics*, Stroudsburg, PA, USA, 2013, pp. 374–377.

[41] 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: https://arxiv.org/abs/1802.05365

[42] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[43] R. Varley, P. Carruthers, and J. Boucher, Eds., *Language and Thought*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[44] R. Varley, M. Siegal, and S. C. Want, "Severe impairment in grammar does not preclude theory of mind," *Neurocase*, vol. 7, no. 6, pp. 489–493, Jan. 2001.

[45] R. C. Schank and R. P. Abelson, "Scripts, plans, and knowledge," in *Proc. 4th Int. Joint Conf. Artif. Intell.* San Francisco, CA, USA: Morgan Kaufmann, 1975, pp. 151–157.

[46] R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Hove, U.K.: Psychology Press, 2013.

[47] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Aug. 2008.

[48] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in *Proc. 15th Conf. ISCA*, Amsterdam, The Netherlands, 2014, pp. 1–5.

[49] S. Hassan and R. Mihalcea, "Cross-lingual semantic relatedness using encyclopedic knowledge," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2009, pp. 1192–1201.

[50] S. Ryu, S. Kim, J. Choi, H. Yu, and G. G. Lee, "Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems," *Pattern Recognit. Lett.*, vol. 88, pp. 26–32, Mar. 2017.

[51] S. Wang, J. Zhang, and C. Zong, "Associative multichannel autoencoder for multimodal word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Stroudsburg, PA, USA, 2018, pp. 115–124.

[52] S. Wang, J. Zhang, and C. Zong, "Learning multimodal word representation via dynamic fusion methods," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Menlo Park, CA, USA, 2018, pp. 5973–5980.

[53] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proc. ACL*, 2014, pp. 721–732.

[54] C. Silberer, V. Ferrari, and M. Lapata, "Visually grounded meaning representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2284–2297, Dec. 2017.

[55] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[56] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781

[57] S. Vashishth, M. Bhandari, P. Yadav, P. Rai, C. Bhattacharyya, and P. Talukdar, "Incorporating syntactic and semantic information in word embeddings using graph convolutional networks," 2018, *arXiv:1809.04283*. [Online]. Available: https://arxiv.org/abs/1809.04283

[58] W. Zhu, X. Jin, S. Liu, Z. Lu, W. Zhang, K. Yan, and B. Wei, "Enhanced double-carrier word embedding via phonetics and writing," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 19, no. 2, pp. 1–18, 2019.

[59] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: https://arxiv.org/abs/1609.08144

[60] X. Liu, "Contrastive study on similarities and differences between Chinese and English characters," in *Proc. ICCESE*. Paris, France: Atlantis Press, 2017, pp. 1–4.

[61] Y.-C. Chen, S.-F. Huang, C.-H. Shen, H.-Y. Lee, and L.-S. Lee, "Phonetic-and-semantic embedding of spoken words with applications in spoken content retrieval," 2018, *arXiv:1807.08089*. [Online]. Available: http://arxiv.org/abs/1807.08089

[62] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 649–657.

[63] W. Zhu, X. Xu, K. Yan, S. Liu, and X. Yin, "A synchronized word representation method with dual perceptual information," *IEEE Access*, vol. 8, pp. 22335–22344, 2020.

[64] K. Lee, L. He, and L. Zettlemoyer, "Higher-order coreference resolution with coarse-to-fine inference," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 2. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 687–692.

[65] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

WENHAO ZHU was born in 1979. He received the bachelor's, master's, and Ph.D. degrees from Zhejiang University in 2002, 2006, and 2009, respectively. From 2012 to 2013, he was a Visiting Scholar with the Computer Laboratory, University of Cambridge, for one year. He is currently an Associate Professor with the School of Computer Engineering and Science, Shanghai University, China. His research interests are in the areas of text representation, information extraction, and web data mining.

SHUANG LIU is currently pursuing the master's degree with Shanghai University. His main research fields include artificial intelligence, natural language processing, and machine learning.

CHAOMING LIU is currently pursuing the master's degree with Shanghai University. His main research fields include artificial intelligence, natural language processing, and machine learning.

XIAOYA YIN is currently pursuing the master's degree with Shanghai University. Her main research fields include artificial intelligence, natural language processing, and machine learning.

XIAPING XV is currently pursuing the master's degree with Shanghai University. Her main research fields include artificial intelligence, natural language processing, and machine learning.