Natural Language Processing Journal 5 (2023) 100038

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

Out-of-vocabulary word embedding learning based on reading comprehension mechanism

Zhongyu Zhuang [a], Ziran Liang [a], Yanghui Rao [a,*], Haoran Xie [b], Fu Lee Wang [c]

[a] Sun Yat-sen University, Guangzhou, China
[b] Lingnan University, Hong Kong, China
[c] Hong Kong Metropolitan University, Hong Kong, China

ARTICLE INFO

Keywords:
Out-of-vocabulary words
Word embedding
Reading comprehension mechanism

ABSTRACT

Currently, most natural language processing tasks use word embeddings as the representation of words. However, when encountering out-of-vocabulary (OOV) words, the performance of downstream models that use word embeddings as input is often quite limited. To solve this problem, the latest methods mainly infer the meaning of OOV words based on two types of information sources: the morphological structure of OOV words and the contexts in which they appear. However, the low frequency of OOV words themselves usually makes them difficult to learn in pre-training tasks by general word embedding models. In addition, this characteristic of OOV word embedding learning also brings the problem of context scarcity. Therefore, we introduce the concept of "similar contexts" based on the classical "distributed hypothesis" in linguistics, by borrowing from the human reading comprehension mechanisms to make up for the deficiency of insufficient contexts in previous OOV word embedding learning work. The experimental results show that our model achieved the highest relative scores in both intrinsic and extrinsic evaluation tasks, which demonstrates the positive effect of the "similar contexts" introduced in our model on OOV word embedding learning.

1. Introduction

Most models for natural language processing (NLP) tasks represent each word in a sentence as a fixed-length vector, which is named word embedding (Mikolov et al., 2013). However, when these models are applied to downstream tasks, they often encounter some words that have not appeared in the vocabulary, which are referred to as out-of-vocabulary (OOV) words. Since these words have not been seen during training, the model generally cannot learn the embeddings of such words correctly, resulting in the performance of the model in downstream tasks not being very ideal (Adams et al., 2017). Therefore, how to tackle OOV words has become one of the largest challenges that need to be overcome in NLP tasks.

For OOV word embedding learning, one possible solution is to assign a unique random vector to each OOV word or to use a unified random vector to represent all OOV words. However, this simple strategy has a very significant drawback: although it enables OOV words to have corresponding vectors as word embeddings, it almost does not capture complex semantic relationships and provide essential information about OOV words for downstream tasks. This does not necessarily lead to better performance in the end, and it may even have the opposite effect in some cases.

To solve this problem, most recent work has proposed methods based on the morphological structure (form) and contexts in which OOV words appear to learn embedding vectors for OOV words and alleviate the data sparsity dilemma in OOV word embedding learning. The typical idea of methods based on OOV word morphology is to fully utilize finer grained sub-word information to provide an initialization embedding for OOV words with limited co-occurrence information. One of the most prominent works in this direction is FastText (Bojanowski et al., 2017), which predicts OOV words by introducing character-level n-gram subword information. In FastText, the embedding of an OOV word can be obtained by adding up n-gram vectors. However, such morphology-based models often require pre-training from scratch and occupy a large amount of computing resources and time. Therefore, MIMICK (Pinter et al., 2017), BoS (Zhao et al., 2018), and KVQ-FH (Sasaki et al., 2019) have been proposed, which only use the morphological structure of words and generate vectors for unseen words by learning from pre-trained embeddings.

The main drawback of morphology-based OOV word embedding methods is their inability to handle OOV words with different meanings in different contexts when neglecting the contexts in which these words appear. This is due to the fact that these methods generate a fixed representation determined by the word's morphological structure. In light

* Corresponding author.
E-mail addresses: zhuangzhy@mail2.sysu.edu.cn (Z. Zhuang), liangzr5@mail2.sysu.edu.cn (Z. Liang), raoyangh@mail.sysu.edu.cn (Y. Rao), hrxie2@gmail.com (H. Xie), pwang@hkmu.edu.hk (F.L. Wang).

https://doi.org/10.1016/j.nlp.2023.100038
Received 2 August 2023; Received in revised form 27 September 2023; Accepted 25 October 2023

of this consideration, Hu et al. (2019) propose HiCE to include OOV word contexts. The structure of HiCE consists of three parts: context encoder, character CNN, and aggregator. It captures the morphological information and context information of the current OOV word through context encoder and character CNN respectively, and then uses an aggregator to infer the embedding of each OOV word by combining the above two kinds of information.

Although we can try to solve the problem of the polysemy phenomenon by introducing the contexts of an OOV word itself, this method still faces a major challenge: the frequency of words that have not appeared in the vocabulary is often low in downstream tasks. Therefore, simply introducing the contexts of an OOV word itself is not enough to provide enough semantic information. To address this challenging issue, we attempt to learn OOV word embeddings by imitating three strategies in human reading comprehension mechanism and use the classical "distributed hypothesis" (Harris, 1954) based on similar contexts to learn relevant embeddings for OOV words and compensate for the lack of context information in previous OOV word embedding learning models.

Currently, research on human reading comprehension mechanism has been very rich, and the current widely accepted one is the top-down human reading theory (Angosto et al., 2013). The top-down reading theory holds that to interpret a piece of information, people need to start from the meaning of a paragraph, and infer according to the meaning of sentences and words, rather than focusing only on words in the bottom-up reading theory. In this way, the reader's grasp of sentence meaning, understanding of various suffixes, and prior knowledge becomes crucial.

In the top-down reading process, humans can use their cognitive skills to implement three strategies to understand unknown words. These three strategies are: synonym substitution, word form correction, and word meaning inference (Tunmer and Nicholson, 2011; Tunmer and Hoover, 2019; Maluf and Cardoso-Martins, 2013). According to the "distributed hypothesis", words that appear in similar contexts tend to have similar semantics. When readers encounter new words in the reading process, they will try to find those known words that have similar contexts with the current new words to infer the meaning of new words; and when readers encounter words that have similar morphological structures with the words they know, they will speculate whether there is a spelling error in the current word and decode the current word; in addition, readers will also use the meaning of sentences where the current word has appeared to guess the meaning of the current word, or use some prefixes or suffixes unique to word formation to reason (Maluf and Cardoso-Martins, 2013; Gülçehre et al., 2016; Taylor et al., 2011).

The existing OOV word embedding learning models only use the last two strategies in the top-down reading process and ignore the role of the "synonym substitution" strategy, rendering the semantic information related to OOV words cannot be mined to the maximum extent. Therefore, we introduce the concept of "similar contexts" and learn the relevant embedding of OOV words based on the "distributed hypothesis" (Harris, 1954) by using similar contexts to compensate for the lack of context information in previous OOV word embedding learning methods.

The main contributions of our work can be summarized as follows: First, we are not just integrating the morphological structure of words and direct contextual information without deeply mining the semantic information that can be provided in the contexts, but rather proposing and introducing the concept of "similar contexts", and using the famous "distributed hypothesis" as the theoretical basis to search for known words with "similar contexts" based on the contexts of OOV words to help us infer more reasonable word embeddings for OOV words, which addresses the issue of insufficient context information in OOV word embedding learning models. Second, we design three corresponding strategies by mimicking the three strategies used by humans in reading comprehension, propose an OOV word embedding learning framework based on human reading comprehension mechanism, and further explore the context information of OOV words using the "similar contexts" we proposed. Third, we compare the performance of our model with other baseline models in multiple tasks, and the experimental results validate the feasibility and effectiveness of our method.

In the following, Section 2 discusses related work; Section 3 introduces our method; Section 4 presents our experiments; Section 5 is the conclusion and future work.

2. Related work

2.1. Word embedding learning based on OOV word morphological structure

Considering that the meaning of words is often related to the morphological structure of words such as word prefixes and suffixes, some methods embed character-level features into word embeddings during training (Wieting et al., 2016; Kim et al., 2018; Edizel et al., 2019; Zhang et al., 2019; Bojanowski et al., 2017). One major disadvantage of these methods is that they often require pre-training from scratch and occupy a large amount of computing resources or memory. For example, FastText uses about 2 million n-gram characters to generate embeddings for OOV words. Some simpler models have been proposed that generate embeddings for OOV words using only the surface form of words by imitating well-trained word embeddings (Pinter et al., 2017; Zhao et al., 2018; Sasaki et al., 2019; Fukuda et al., 2020). However, the word formation may be complex and highly internally structured (Anderson, 1992), and the performance of such methods is often not particularly ideal.

To balance the complexity and performance of the model and solve the scalability problem, LOVE (Learning Out-of-Vocabulary Embeddings) (Chen et al., 2022) uses WordPiece (Wu et al., 2016) to obtain both the character sequence and subwords of words, thus avoiding the highly redundant problem caused by character-level n-grams used in FastText. In addition, in order to alleviate the deterioration of model performance caused by character-level perturbations (Liang et al., 2018a; Belinkov and Bisk, 2018; Sun et al., 2020; Jin et al., 2020), LOVE also uses possible common character-level errors for data augmentation to generate corresponding positive samples, which are used as inputs to the encoder to obtain corresponding embeddings by pulling them closer to the embeddings corresponding to word prototypes. Considering the improvement of negative samples on model performance, LOVE also introduces infoNCE loss (Wang and Isola, 2020) for contrastive learning to push negative sample pairs away from each other. Experimental results show that LOVE performs well in terms of model complexity, performance and scalability. However, these methods still rely solely on the morphological structure of OOV words to infer their meanings, without using the semantic information of each OOV word.

2.2. Word embedding learning based on OOV word contexts

Methods based on the morphological structure of words have shown excellent performance in OOV word embedding learning. However, when dealing with polysemy or some special proper names, the performance of methods based solely on morphological structure will decline. In order to alleviate this problem, methods such as Comick (Garneau et al., 2019) and HiCE (Hu et al., 2019) have been proposed. However, such models often only utilize the context of OOV words themselves, since the frequency of OOV words in downstream tasks is often low, the additional semantic information brought by these methods is still insufficient.

3. Our method

3.1. MIMICK-like module

For a given pre-trained word embedding and OOV words, the core idea of a MIMICK-like model is to imitate the embedding space of the background word embedding model through morphological structure or contexts of OOV words. Specifically, for a given size of vocabulary $\mathcal{V}$, its corresponding word embedding matrix is $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times m}$, for word w, its word embedding is $\mathbf{u}_w \in \mathbb{R}^m$, the goal of a MIMICK-like model is to estimate an embedding $\mathbf{v}_w \in \mathbb{R}^m$ for any word $w \notin \mathcal{V}$, and the training objective is to minimize the expected distance between $\mathbf{u}_w$ and $\mathbf{v}_w$, as follows:

$$\mathcal{L}_d = \frac{1}{|\mathcal{V}|} \sum_{w \in \mathcal{V}} \mathcal{D}(\mathbf{u}_w, \mathbf{v}_w), \tag{1}$$

where we treat in-vocabulary word w as an OOV word during training and infer the corresponding embedding based on its morphological structure and contextual information, and $\mathcal{D}(\cdot)$ is a distance function, which can be the Euclidean or the cosine distance function. Furthermore, the predicted embedding $\mathbf{v}_w$ is generally obtained by the following formula:

$$\mathbf{v}_w = \psi(\zeta(w)). \tag{2}$$

In the above, $\zeta(\cdot)$ maps w to a list of subunits according to the word's morphological structure, and then feeds the sequence as input to function $\psi(\cdot)$, which generates the predicted embedding. $\psi(\cdot)$ can be CNN, RNN or some simple summation functions. As shown in Chen et al. (2022), the effectiveness of Positional Attention Module (PAM) was validated to be much better than CNN, RNN, and Self-Attention (SA) in generating the predicted embedding. Accordingly, we use PAM as the function $\psi(\cdot)$. After training, the model can induce embeddings for any word.

3.2. Reading comprehension mechanism

Another problem that traditional methods based on OOV word contexts need to face is that most OOV words have low frequency, so their contexts are difficult to provide enough effective information. To this end, we propose a framework based on human reading comprehension mechanism to achieve embedding learning of OOV words, in order to make up for the deficiency of insufficient context information. This mechanism mainly has three strategies:

1. **Synonym Substitution:** According to the classic "distributed hypothesis" in linguistics, words with similar contexts tend to have similar semantics. Based on this theory, people often infer the meaning of unknown words by looking for known words with similar contexts.
2. **Word Form Correction:** In this process, when people encounter words with typos during the reading process, they often look for known words with similar word forms and automatically correct the former to the latter.
3. **Word Meaning Inference:** This strategy indicates inferring the meaning of words through the morphological information (prefixes, roots, suffixes, etc.) of unknown words themselves and context information.

By imitating this reading comprehension mechanism, we propose an OOV word embedding learning model based on three corresponding strategies. These three strategies are similarity, decoding and prediction. We will explain these three strategies in detail in the following text. Fig. 1 shows the word embedding learning framework proposed by us based on reading comprehension mechanism.

3.2.1. Similarity

This strategy corresponds to the "synonym substitution" strategy in the reading comprehension mechanism. In this strategy, we infer the meaning of unknown words by using known words with similar contexts to make up for the deficiency of insufficient context information in previous OOV word embedding learning work.

First of all, for word $w \in \mathcal{V}$, in addition to obtaining its corresponding embedding $\mathbf{u}_w \in \mathbb{R}^m$ in traditional methods, we also need to obtain its context embedding $\mathbf{c}_w \in \mathbb{R}^m$. We introduce HiCE (Hu et al., 2019) to obtain the context embedding $\mathbf{c}_w$ of words themselves. The specific method is: in each round of the few shot learning task, we randomly select K sentences from $\mathcal{S} = \{S_1, \ldots, S_q\}$, where $\mathcal{S}$ is the set of the contexts where the word w has appeared. Then, we mask w and remove, if any, other OOV words from $\mathcal{S}$ to construct a context $\mathcal{S}^K = \{s_k\}_{k=1}^K$, where s_k is the kth sentence where the word w is located. With respect to the extreme situation where all words in all sentences that an OOV word has appeared are OOV words that makes contextual information unusable, our model can still infer an embedding based on the morphological structure of OOV word itself. Afterwards, we input s_k to the underlying context encoder E and obtain the encoding vector X, which is:

$$X = \mathrm{Concat}(E(s_1), \ldots, E(s_K)). \tag{3}$$

The formula for context embedding $\mathbf{c}_w$ is given below:

$$\mathbf{c}_w = \mathrm{FFN}(\mathrm{SA}(X)), \tag{4}$$

where SA is the self-attention layer, and FFN is the feed-forward network. For OOV word $w' \notin \mathcal{V}$, we obtain its context embedding $\mathbf{c}_{w'} \in \mathbb{R}^m$ ($\mathbf{c}_{w'} = \bar{\mathbf{v}}_{\mathrm{context}}$ in Section 3.2.3) by the same method as obtaining $\mathbf{c}_w$. When inferring OOV word embedding, we look for n (set to 5 in the experiment) embeddings $\mathbf{c}'_{w_j}$ $(j = 1, \ldots, n)$ in $\mathbf{W}'$ that are most similar to $\mathbf{c}_{w'}$ and obtain the word embedding $\mathbf{u}'_{w_j}$ $(j = 1, \ldots, n)$ corresponding to these embeddings. The formula of obtaining similar embedding $\bar{\mathbf{v}}_{\mathrm{sim}}$ is defined as follows:

$$\bar{\mathbf{v}}_{\mathrm{sim}} = \sum_{j=1}^{n} \rho_j \cdot \mathbf{u}'_{w_j}, \tag{5}$$

where $\rho_j = \mathrm{softmax}(\mathrm{sim}(\mathbf{c}'_{w_j}, \mathbf{c}_w))$ and $\mathrm{sim}(\cdot)$ represents the cosine similarity.

3.2.2. Decoding

This strategy corresponds to the "word form correction" strategy in the reading comprehension mechanism. In this strategy, we add positive and negative samples during training to improve the model's ability to recognize incorrect word forms.

To alleviate the deterioration of model performance caused by slight perturbations, we refer to LOVE's method and introduce data augmentation and contrastive learning ideas. The specific method is: generate five types of positive samples for existing data by swapping, discarding, inserting, replacing four character-level methods and synonym replacement to enrich the number of training samples. The first four augmentation methods imitate adversarial attacks (Schick and Schütze, 2019), i.e., we assume that some OOV words are formed by words in the vocabulary being attacked by the first four types, and many OOV words in real text are actually caused by people typing incorrectly. Accordingly, we can generate reliable embeddings for the words with typos (OOV words) through these four types of augmentation strategies. Furthermore, we add the synonym replacement strategy to ensure that words with different morphological structures but similar semantics are still embedded close in the embedding space, which prevents our model from overfitting the morphological structures of words. At the same time, negative sample examples with similar surface forms but different meanings are introduced for data enhancement. Considering that the MSE used by traditional MIMICK-like models can only pull the distance between positive samples closer, we introduce the

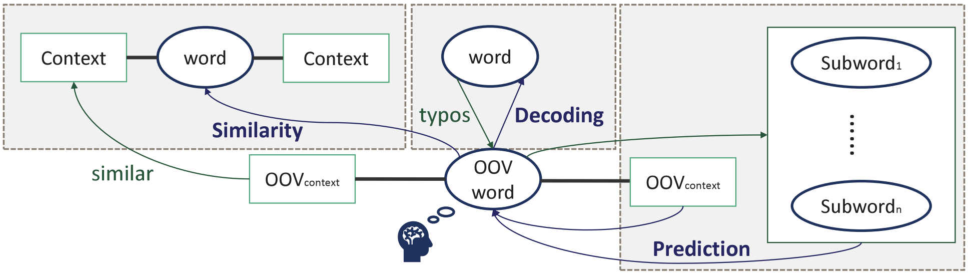

Fig. 1. Out-of-vocabulary word embedding learning based on reading comprehension mechanism.

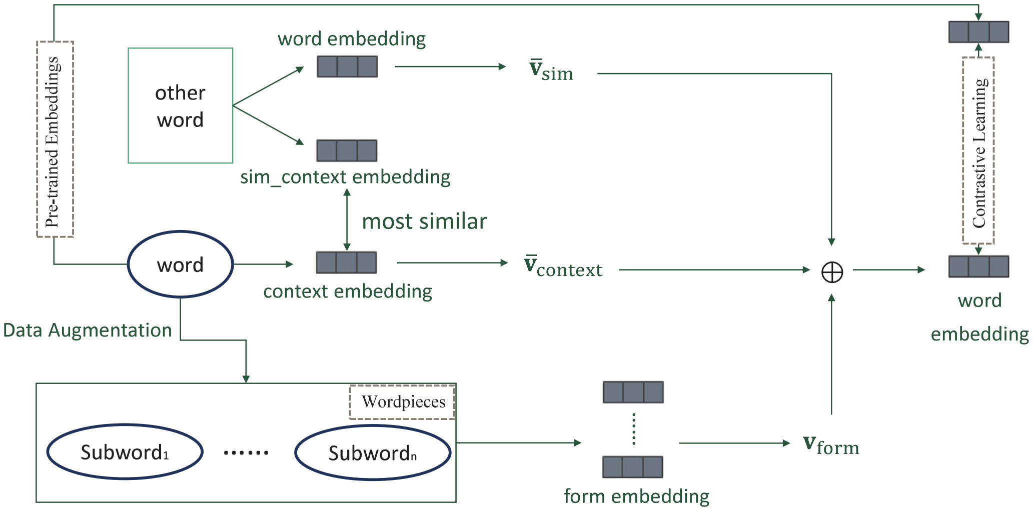

Fig. 2. Overall framework of the model.

infoNCE loss function in contrastive learning (Wang and Isola, 2020), which focuses on two indicators of Alignment and Uniformity. These two indicators are responsible for pulling positive samples closer and pushing negative samples away respectively. We will elaborate on this part in Section 3.3. So far, we can obtain word embeddings based on morphological structure.

3.2.3. Prediction

This strategy corresponds to the "word meaning inference" strategy in the reading comprehension mechanism. In this strategy, similar to the method of obtaining $\mathbf{c}_w$ in Section 3.2.1, we obtain context embedding $\bar{\mathbf{v}}_{\text{context}}$ by:

$$\bar{\mathbf{v}}_{\text{context}} = \mathbf{c}_{w'}. \tag{6}$$

After obtaining $\bar{\mathbf{v}}_{\text{context}}$, it is weighted and added with $\mathbf{v}_{\text{form}}$ obtained in Section 3.2.2 to correct the problem of one word with multiple meanings that cannot be recognized by morphology-based methods alone. Finally, the predicted embedding $\mathbf{v}_{w'}$ is obtained by:

$$\mathbf{v}_{w'} = \alpha \cdot \bar{\mathbf{v}}_{\text{context}} + \beta \cdot (1-\alpha) \cdot \bar{\mathbf{v}}_{\text{sim}} + (1-\beta)(1-\alpha) \cdot \mathbf{v}_{\text{form}}, \tag{7}$$

where $\alpha = \sigma(a \cdot \text{cxt}_{length} + b)$, in which, a and b are learnable parameters. β is a hyper parameter (set to 0.75 in the experiment), cxt_{length} is the contextual length of the word w', and $\sigma(\cdot)$ is the sigmoid function.

Fig. 2 shows the overall framework of our model.

3.3. Loss function

In this section, we focus on the loss function $\mathcal{L}(\cdot)$. Traditional MIMICK-like models usually use mean squared error (MSE) to try to give similar embeddings to words with similar surface forms. However, MSE can only pull positive sample pairs closer together, but cannot push negative sample pairs further apart and may even pull them closer together to make the loss as small as possible. To address this problem, Wang and Isola (2020) proposed infoNCE loss, which optimizes two properties: Alignment and Uniformity. Alignment describes the distance between positive sample pairs:

$$\mathcal{L}_{\text{align}} \triangleq \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \mathcal{D}(\mathrm{u}_x, \mathrm{u}_y), \tag{8}$$

where p_{pos} represents the distribution of positive sample pairs. Uniformity is used to measure whether the learned representation is uniformly distributed on the hypersphere.

$$\mathcal{L}_{\text{uniform}} \triangleq \log \mathop{\mathbb{E}}_{(x,y)\overset{i.i.d}{\sim} p_{\text{data}}} e^{-t\cdot\mathcal{D}(\mathrm{u}_x,\mathrm{u}_y)}, \tag{9}$$

where p_{data} is the data distribution and $t > 0$. These two properties are consistent with our expected vocabulary representation: positive sample words should be closer together, while negative sample words should be far apart from each other and finally scattered on the hypersphere. Our final loss function $\mathcal{L}(\cdot)$ is obtained by adding $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{uniform}}$ together, as follows:

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}. \tag{10}$$

4. Experiments

4.1. Datasets and experimental settings

We mainly use two tasks to evaluate word representations, i.e., intrinsic and extrinsic evaluation tasks. Intrinsic evaluation directly measures the syntactic or semantic relationship between words, such as word similarity in phrases. Extrinsic evaluation measures the performance of word embeddings as input features for downstream tasks such as text classification and named entity recognition (NER). In addition,

Table 1
The results of intrinsic evaluation on the Chimeras dataset. The highest score in each column is highlighted in bold, while the second highest score is highlighted with an underline.

Models	2 sent.	4 sent.	6 sent.	AVG
HiCE	36.18	38.41	41.18	38.59
LOVE	40.19	38.99	40.54	39.91
Ours w/o $\bar{\mathbf{v}}_{\text{sim}}$	39.80	38.79	40.97	39.85
Ours	**41.43**	**40.83**	**43.20**	**41.82**

we also evaluate the performance of the model by conducting ablation experiments.

Considering that our model uses contextual information to learn the embedding of OOV words, and the Chimeras (Lazaridou et al., 2017) dataset provides a small amount of relevant contexts (2, 4, or 6 sentences) for each OOV word. Therefore, we used the Chimeras dataset for intrinsic evaluation tasks.

For extrinsic evaluation tasks, we used four extrinsic datasets (two text classification tasks and two NER tasks): MR (Pang and Lee, 2005), SST2 (Socher et al., 2013), CoNLL-03 (Sang and Meulder, 2003), and BC2GM (Smith et al., 2008).

In ablation experiments, we used six datasets (six word similarity tasks): RareWord (Luong et al., 2013), MEN (Donig et al., 2020), SimLex (Hill et al., 2015), WordSim353 (Agirre et al., 2009), Simverb (Agirre et al., 2009) and MTurk (Halawi et al., 2012). The background word model for the experiment is Word2Vec (Herbelot and Baroni, 2017) and the trained contextual corpus is WikiText-103 (Merity et al., 2017) for all models. Furthermore, our model's pre-trained vocabulary comprises words and their embeddings that are present in both the WikiText-103 dataset context corpus and the Word2Vec model vocabulary, with a size of over 80,000.

Apart from our own model, we included two state-of-the-art models, i.e., HiCE (Hu et al., 2019) and LOVE (Chen et al., 2022), as baselines in the experiments. In order to verify the effectiveness of the "similar contexts" introduced in our model, we also included the Ours w/o $\bar{\mathbf{v}}_{\text{sim}}$ model as a comparison, which is formed by removing the $\bar{\mathbf{v}}_{\text{sim}}$ module from our model.

4.2. Results on intrinsic evaluation tasks

Table 1 shows the experimental results of intrinsic evaluation task on Chimeras dataset. It can be seen that, our model with $\bar{\mathbf{v}}_{\text{sim}}$ achieved the highest scores compared to other models with the same contexts (2, 4, or 6 sentences). Furthermore, when the number of context sentences increased from 2 to 4, both LOVE and our models' scores decreased, which may be due to some of the same words have different word frequencies in the test sets with context sentences of 2 and 4. However, as the number of context sentences further increased from 4 to 6, with no change in test vocabulary and an increase in context information, the scores rose, and were higher than those with 2 context sentences, which indicates that sufficient context information is helpful for inferring the meaning of OOV words.

In addition, although Ours w/o $\bar{\mathbf{v}}_{\text{sim}}$ scored higher than HiCE with the same contexts, its scores were slightly lower than LOVE's with 2 and 4 context sentences. It was not until the number of sentences increased from 4 to 6 that the former's score surpassed the latter's, which suggests that introducing OOV word context information alone is insufficient when the OOV word's frequency is low. However, with the introduction of "similar contexts", our model's scores significantly improved and were higher than other models' in all situations. This indicates that the context information was further mined, making it helpful for our model to infer the meaning of OOV words.

Table 2
The results of extrinsic evaluation tasks. The highest score in each column is highlighted in bold, while the second highest score is highlighted with an underline.

Models	Text classification (Acc)			NER (F1-score)		
	MR	SST2	AVG	CoNLL-03	BC2GM	AVG
HiCE	60.49	70.95	65.72	59.38	48.34	53.86
LOVE	**74.67**	79.58	77.13	**71.16**	60.30	65.73
Ours w/o $\bar{\mathbf{v}}_{\text{sim}}$	72.86	79.60	76.23	70.77	61.38	66.08
Ours	74.03	**80.50**	**77.27**	70.98	**63.92**	**67.45**

Table 3
The results of ablation experiments. The highest score in each column is highlighted in bold, while the second highest score is highlighted with an underline.

Models	Word similarity (Spearman's ρ)						
	RareWord	MEN	SimLex	WordSim353	SimVerb	MTurk	AVG
HiCE	6.3	8.0	2.3	7.9	1.8	7.4	5.6
LOVE	41.0	**63.1**	24.7	46.3	25.6	51.7	42.1
Ours w/o $\bar{\mathbf{v}}_{\text{sim}}$	**42.3**	62.9	**27.2**	44.2	**25.9**	**55.5**	43.0
Ours	41.7	63.0	25.6	**47.8**	25.6	55.1	**43.1**

4.3. Results on extrinsic evaluation tasks

Table 2 shows the experimental results of different models on four extrinsic evaluation tasks. In the text classification task, although our models did not achieve the highest score in all given datasets, our model with $\bar{\mathbf{v}}_{\text{sim}}$ achieved a best result and a second-best result respectively. The situation is similar in the named entity recognition task, where although our model with $\bar{\mathbf{v}}_{\text{sim}}$ did not achieve the highest score in all datasets, it still obtained the highest score in one dataset and the second-highest score in another. Moreover, considering the overall performance of our model in both the text classification and named entity recognition tasks, its average scores were higher than other models', indicating that the "similar contexts" we introduced was significantly helpful for inferring the meaning of OOV words.

4.4. Ablation experiments

Table 3 shows the experimental results of our models and baselines over six datasets. Noteworthy, since all the datasets in the word similarity task do not contain contexts, HiCE actually only uses the character-level CNN module in this task. Compared with other degraded models (i.e., the baselines of HiCE and LOVE), our models obtained the highest and the second highest average score in word similarity tasks. Specifically, our models performed best in five of six word similarity task. This indicates that our method can further introduce additional semantic information to help learn embeddings of OOV words while retaining the advantages of the original model.

5. Conclusions

In this paper, we propose an OOV Word embedding learning method based on the human reading comprehension mechanism. We provide additional semantic information for inferring OOV words by imitating the "synonym substitution" strategy used by humans when reading. We compare our model with other baseline models in the tasks of text classification, named-entity recognition, and word similarity, respectively. The results show that our method performs best in most cases, which shows that the introduction of the concept of "similar contexts" really helps us to infer the meaning of OOV words.

In addition, some previous methods based on topic modeling and contrastive learning have shown excellent performance in downstream tasks such as sentiment analysis with OOV words (Huang et al., 2017; Liang et al., 2018b; Chen and Xie, 2020; Xu et al., 2023). This may indicate that these methods have also explored to some extent the information related to the meaning of OOV words. Whether we can discover new information similar to our "similar contexts" information from these methods and further improve the reliability of our inference of the meaning of OOV words is a question that needs to be considered in future work.

CRediT authorship contribution statement

Zhongyu Zhuang: Conceptualization, Methodology, Software, Writing. **Ziran Liang:** Validation, Writing – review & editing. **Yanghui Rao:** Supervision, Writing. **Haoran Xie:** Supervision, Writing. **Fu Lee Wang:** Supervision, Writing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work of Yanghui Rao was supported in part by the National Natural Science Foundation of China (61972426). The work of Haoran Xie was supported in part by Lam Woo Research Fund (LWP20019), the Direct Grant (DR23B2), and the Faculty Research Grants (DB23A3 and DB23B2) of Lingnan University, Hong Kong. The work of Fu Lee Wang was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E01/19).

References

Adams, O., Makarucha, A.J., Neubig, G., Bird, S., Cohn, T., 2017. Cross-lingual word embeddings for low-resource language modeling. In: EACL. pp. 937–947.

Agirre, E., Alfonseca, E., Hall, K.B., Kravalova, J., Pasca, M., Soroa, A., 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In: NAACL-HLT. pp. 19–27.

Anderson, S.R., 1992. A-Morphous Morphology. Cambridge University Press.

Angosto, A., Sánchez, P., Álvarez, M., Cuevas, I., León, J.A., 2013. Evidence for top-down processing in reading comprehension of children. Psicol. Educ. 19 (2), 83–88.

Belinkov, Y., Bisk, Y., 2018. Synthetic and natural noise both break neural machine translation. In: ICLR.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. 5, 135–146.

Chen, L., Varoquaux, G., Suchanek, F., 2022. Imputing out-of-vocabulary embeddings with LOVE makes language models robust with little cost. In: ACL. pp. 3488–3504.

Chen, X., Xie, H., 2020. A structural topic modeling-based bibliometric study of sentiment analysis literature. Cogn. Comput. 12 (6), 1097–1129.

Donig, S., Christoforaki, M., Bermeitinger, B., Handschuh, S., 2020. Multimodal semantic transfer from text to image. Fine-grained image classification by distributional semantics. CoRR abs/2001.02372.

Edizel, B., Piktus, A., Bojanowski, P., Ferreira, R., Grave, E., Silvestri, F., 2019. Misspelling oblivious word embeddings. In: NAACL-HLT. pp. 3226–3234.

Fukuda, N., Yoshinaga, N., Kitsuregawa, M., 2020. Robust backed-off estimation of out-of-vocabulary embeddings. In: Findings of EMNLP. pp. 4827–4838.

Garneau, N., Leboeuf, J., Lamontagne, L., 2019. Predicting and interpreting embeddings for out of vocabulary words in downstream tasks. CoRR abs/1903.00724.

Gülçehre, Ç., Ahn, S., Nallapati, R., Zhou, B., Bengio, Y., 2016. Pointing the unknown words. In: ACL. pp. 140–149.

Halawi, G., Dror, G., Gabrilovich, E., Koren, Y., 2012. Large-scale learning of word relatedness with constraints. In: SIGKDD. pp. 1406–1414.

Harris, Z.S., 1954. Distributional structure. Word 10 (2–3), 146–162.

Herbelot, A., Baroni, M., 2017. High-risk learning: acquiring new word vectors from tiny data. In: EMNLP. pp. 304–309.

Hill, F., Reichart, R., Korhonen, A., 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. Comput. Linguist. 41 (4), 665–695.

Hu, Z., Chen, T., Chang, K., Sun, Y., 2019. Few-shot representation learning for out-of-vocabulary words. In: ACL. pp. 4102–4112.

Huang, X., Rao, Y., Xie, H., Wong, T.-L., Wang, F.L., 2017. Cross-domain sentiment classification via topic-related TrAdaBoost. In: AAAI. pp. 4939–4940.

Jin, D., Jin, Z., Zhou, J.T., Szolovits, P., 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In: AAAI. pp. 8018–8025.

Kim, Y., Kim, K., Lee, J., Lee, S., 2018. Learning to generate word representations using subword information. In: COLING. pp. 2551–2561.

Lazaridou, A., Marelli, M., Baroni, M., 2017. Multimodal word meaning induction from minimal exposure to natural text. Cogn. Sci. 41, 677–705.

Liang, B., Li, H., Su, M., Bian, P., Li, X., Shi, W., 2018a. Deep text classification can be fooled. In: IJCAI. pp. 4208–4215.

Liang, W., Xie, H., Rao, Y., Lau, R.Y.K., Wang, F.L., 2018b. Universal affective model for readers' emotion classification over short texts. Expert Syst. Appl. 114, 322–333.

Luong, T., Socher, R., Manning, C.D., 2013. Better word representations with recursive neural networks for morphology. In: CoNLL. pp. 104–113.

Maluf, M.R., Cardoso-Martins, C., 2013. Alfabetização No Séuulo XXI: Como se Aprende a Ler e a Escrever. Penso Editora.

Merity, S., Xiong, C., Bradbury, J., Socher, R., 2017. Pointer sentinel mixture models. In: ICLR.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: NIPS. pp. 3111–3119.

Pang, B., Lee, L., 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: ACL. pp. 115–124.

Pinter, Y., Guthrie, R., Eisenstein, J., 2017. Mimicking word embeddings using subword RNNs. In: EMNLP. pp. 102–112.

Sang, E.F.T.K., Meulder, F.D., 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: NAACL-HLT. pp. 142–147.

Sasaki, S., Suzuki, J., Inui, K., 2019. Subword-based compact reconstruction of word embeddings. In: NAACL-HLT. pp. 3498–3508.

Schick, T., Schütze, H., 2019. Attentive mimicking: Better word embeddings by attending to informative contexts. In: NAACL-HLT. pp. 489–494.

Smith, L., Tanabe, L.K., Kuo, C.-J., Chung, I., Hsu, C.-N., Lin, Y.-S., Klinger, R., Friedrich, C.M., Ganchev, K., Torii, M., et al., 2008. Overview of BioCreative II gene mention recognition. Genome Biol. 9 (2), 1–19.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In: EMNLP. pp. 1631–1642.

Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P.S., Xiong, C., 2020. Adv-BERT: BERT is not robust on misspellings! generating nature adversarial samples on BERT. CoRR abs/2003.04985.

Taylor, J.M., Raskin, V., Hempelmann, C.F., 2011. Towards computational guessing of unknown word meanings: The ontological semantic approach. In: CogSci. pp. 3581–3586.

Tunmer, W.E., Hoover, W.A., 2019. The cognitive foundations of learning to read: A framework for preventing and remediating reading difficulties. Aust. J. Learn. Diffic. 24 (1), 75–93.

Tunmer, W.E., Nicholson, T., 2011. The development and teaching of word recognition skill. In: Handbook of Reading Research. Routledge, pp. 405–431.

Wang, T., Isola, P., 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: ICML. pp. 9929–9939.

Wieting, J., Bansal, M., Gimpel, K., Livescu, K., 2016. Charagram: Embedding words and sentences via character n-grams. In: EMNLP. pp. 1504–1515.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J., 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR arXiv:1609.08144.

Xu, L., Xie, H., Li, Z., Wang, F.L., Wang, W., Li, Q., 2023. Contrastive learning models for sentence representations. ACM Trans. Intell. Syst. Technol. 14 (4), 67:1–67:34.

Zhang, Y., Chen, Q., Yang, Z., Lin, H., Lu, Z., 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci. Data 6 (1), 52.

Zhao, J., Mudgal, S., Liang, Y., 2018. Generalizing word embeddings using bag of subwords. In: EMNLP. pp. 601–606.