Complex & Intelligent Systems (2021) 7:2977–2990
https://doi.org/10.1007/s40747-021-00482-y

ORIGINAL ARTICLE

Modeling multi-prototype Chinese word representation learning for word similarity

Fulian Yin[1] · Yanyan Wang[1] · Jianbo Liu[1] · Marco Tosato[2]

Received: 22 December 2020 / Accepted: 19 July 2021 / Published online: 4 August 2021

Abstract
The word similarity task is used to calculate the similarity of any pair of words, and is a basic technology of natural language processing (NLP). The existing method is based on word embedding, which fails to capture polysemy and is greatly influenced by the quality of the corpus. In this paper, we propose a multi-prototype Chinese word representation model (MP-CWR) for word similarity based on synonym knowledge base, including knowledge representation module and word similarity module. For the first module, we propose a dual attention to combine semantic information for jointly learning word knowledge representation. The MP-CWR model utilizes the synonyms as prior knowledge to supplement the relationship between words, which is helpful to solve the challenge of semantic expression due to insufficient data. As for the word similarity module, we propose a multi-prototype representation for each word. Then we calculate and fuse the conceptual similarity of two words to obtain the final result. Finally, we verify the effectiveness of our model on three public data sets with other baseline models. In addition, the experiments also prove the stability and scalability of our MP-CWR model under different corpora.

Keywords Chinese word representation · Multi-prototype · Synonym knowledge base · Word semantic disambiguation

Introduction

Word similarity (WS) is a critical task in NLP, and its intended scope is to measure the relatedness or similarity degree between word pairs [1–4]. Currently, the most popular method to solve this task is word embedding [5–7], which yields low-dimensional word vectors from corpora to calculate word similarity. It also has become a relevant topic in recent years and plays a really important role in NLP downstream tasks, such as Word sense disambiguation [8,9], machine translation [10,11], text summarization [12,13], context identification system [14].

Most research methods, such as Word2vec [15] and GloVe [16], obtain a single vector of each word by training special corpora. They can not capture polysemy, because every word vector is associated to a certain meaning. This has led to research on languages that contain rich semantic information, such as Chinese, becoming a hot topic. To overcome this problem, multi-prototype word embedding models have been built [17–20], which mainly utilize context clustering to represent the different meanings of words. However, there are still some problems that should be addressed. These include the fact that it is difficult to determine the number of clusters and the clustering results may not correspond to each meaning of a word. In addition, these methods are difficult to use, since they require a really significant amount of data with high quality for training model.

Some recent studies attempted to incorporate additional external knowledge base, which can not only clarify each of the different meanings of a word, but also include additional relationships between words in case of insufficient data. WordNet [21] is one of the most commonly used external knowledge bases in English. Chen et al. [22] applied WordNet to train vectors for different meanings of polysemous words, and proposed a character-enhanced word embedding model (CWE). After that, more fine-grained Chinese radicals and other information have been added to represent words [23,25,39]. The most popular Chinese knowledge bases include HowNet [26] and Tongyici Cilin [27]. Dong

✉ Jianbo Liu
ljbcuc@163.com

[1] Institute of Information and Communication, Communication University of China, Beijing 100024, China

[2] Laboratory for Industrial and Applied Mathematics, York University, Toronto M3J 1P3, Canada

and Dong [26] employed one or more sememes which represent word concepts in HowNet (a Chinese concept database), while Tongyici Cilin utilizes synonyms or related words to aggregate each word sense. Niu et al. [20] proposed a sememe attention over target model (SAT) used HowNet to learn representations of sememes, senses and words, and added an attention scheme for sense detection. The above methods all calculated the word similarity using the word embedding. However, a common limitation of these models is that they ignore whether the word embedding can represent accurate semantic information or not due to the quality problems of the training corpus.

In this paper, to study Chinese polysemous expressions on limited data for improving the word similarity, we first introduce synonym knowledge base to learn semantic information of words, and propose a multi-prototype Chinese word representation model, which regards the frame of Skip-gram in word2vec [15,28] as the foundation. In our model, we look up synonyms for words and construct two similarity matrices for each word: the former is based on the pre-trained word vectors, the latter is based on synonyms or related words from synonym knowledge base. Then we use an attention mechanism to identify and modify the incorrect vectors obtained in the pre-trained vectors. Finally, each word is represented by multi-prototype vectors.

To evaluate the model proposed in this paper, we conduct experiments on several different corpora. In the word similarity task, the results show that our model has an optimal performance and is adaptable to various data set choices compared to other methods. In the nearest neighbor detection task, our model outperforms the existing baseline models.

The rest of the paper is organized as follows: we introduce existing approaches of word embedding for word similarity in second section, and describe our model in third section. Then, we present the results and performance comparisons in fourth section followed by the conclusions and next research plan in the final section.

Related work

Since word representation plays an important role in word similarity [2,3], it has become a relevant topic both for industry and academia. This paper is mainly based on word embedding method to improve the word similarity task, hence we mainly explore the research of word embedding.

Word embedding models

In the early stages, it was common to use one-hot vectors to represent words, which leaded to dimensional disaster. To address this problem, there are two methods: matrix factorization and distributed representation. The former is a typical method that involves low-dimensional vectors from sparse matrices [29], such as LSA and PPMI [30]. Currently, with the continuous development of deep learning, the later distributed representation is becoming more and more popular, various models are shown in Table 1. Neural Network Language Model (NNLM) [31] trained a fully connected network to reduce dimensions of words and at the same time the vector representation of similar words is similar. However, the input of the model is a fixed number of context words, which cannot obtain the information of further words. Recurrent Neural Network Language Model (RNNLM) [32] used RNN instead of the fully connected neural network to model language model. Later, Word2Vec [15,28] and GloVe [16] are successively proposed. Word2vec [15,28] consists of two models: CBOW (Continuous Bag-of-Word Model) and Skip-gram. CBOW model predicts the vector representation of the current word by context words, and Skip-gram model is based on the vector representation of the current word to predict the vector representations of context words. GloVe [16] (Global vectors for word representation) model takes into account both the global statistical features and the local context features of the corpus. And the experimental effect is similar to that of word2vec, but it takes up more memory.

Since then, most of the methods are just based on the above models. Lu et al. [33] proposed a position-sensitive Skip-gram model to the task of similarity calculation and entity recognition. Bojanowski et al. [34] added sub-word information to construct the Fasttext model, which took the Skip-gram model with negative sampling and performed fast in model training with a large corpus. At the same time, N-grams co-occurrence statistics information was applied to four word representation models: SGNS, GloVe, PPMI matrix and its SVD factorization [35], and it was proved to be useful to find antonyms and relieve computing burden. Then, Song et al. [36] decided to distinguish the left and right context in Word2vec model, by proposing a Directional Skip-gram (DSG) model. The experiment attested its effectiveness in both semantic and syntactic information representation. Heo et al. [37] proposed a model to evaluate semantic similarity using both global and local information. Meng et al. [38] proposed an unsupervised word embedding learning model combining local and global context. By simply extending the Word2Vec structures, their objective function included the loss corresponding to global context. In addition, some fine-tuned models are proposed, they have integrated some other information, such as sentimental information [39], character information [22,39], document labels [40,41], syntactic information [42], on the basis of their original pre-trained vectors. In recent years, some new models have also been proposed, such as ELMo (Embeddings from Language Models) [43] and BERT (Bidirectional Encoder Representations from Transformer) [44]. Peters et al. [43]

Table 1 A table of the related work for distributed word representation

Models	Years	Techniques	Contributions
NNLM [31]	2003	A fully connected neural network is used	A pioneering work
RNNLM [32]	2010	RNN	Complete historical information is captured
Word2Vec [15,28]	2013	Word context information is considered	Word embedding is proposed
Glove [16]	2014	The global statistical and local context features are added	The effect is similar to word2vec
CWE [22]	2015	Character information is introduced	Semantic ambiguity is alleviated
SCWE [23]	2016	Internal structure of words is considered	Semantic ambiguity is alleviated
Fasttext [34]	2017	Sub-word information is added	The OOV problem is alleviated
SAT [20]	2017	Concept information of HowNet is incorporated	Semantic ambiguity is alleviated
DSG [36]	2018	The orientation of context words is considered	The directions of words can be pointed out
WE model [45]	2019	Distributional and ontology-based information are combined	Deep analysis of different methods are explored
BERT [44]	2019	The context of all layers are considered	Deep bidirectional representation is pre-trained
Meng et al. [38]	2020	Both global and local information is considered	Complementary word contexts are captured

proposed ELMo model, which used linear combination of layers to represent word vectors based on bidirectional language model. Devlin et al. [44] proposed BERT, which aimed to pre-train deep bidirectional representation according to the context of all layers. However, these two methods are not suitable for context-free tasks, and require high computer performance.

Although these above methods yield good performances, they present several limitations in Chinese word representation. First, they generally use a single vector to represent words, ignoring polysemy. This is a major limitation for languages which have rich semantics, such as Chinese, since word sense disambiguation is essential for word representation. Second, a lager amount of corpora are needed for training model. This is not always possible, since there are not enough corpora in specific fields. At the same time, large-scale corpora can cause a great hardware burden. Thirdly, these methods based on pre-trained results ignore quality problems of the corpus. Finally, most of the above methods are based on the English language. Chinese is completely different in both historically and culturally, thus it is unreasonable to use such models directly.

Multi-prototype word embedding models

To address the above problems, the general method is to construct multi-prototype word embedding models by clustering contexts. First, Reisinger and Mooney [17] clustered local contexts to provide a context-dependent vector representation of different senses for words. The effectiveness of their model has been demonstrated by a semantic similarity experiment. After that, a new neural network architecture has been proposed by Huang et al. [18], which combined both local and global document context. Tian et al. [19] then designed an expectation–maximization algorithm to learn the multi-prototype vectors of words. Chen et al. [22] proposed a CWE model to learn representation by embedding multi prototype characters. Xu et al. [23] proposed SCEW model to represent Chinese words by exploiting internal structure of words, which could realize semantic disambiguation to some extent.

In addition, external expert resources have been introduced which can not only represent polysemous words, but also complement certain relationships between words, by adding sematic information. In English system, Juan et al. [45] evaluated and compared the similarity calculation methods based on word embedding and knowledge base on multiple data sets, and found that the combination of the two methods could get better results. And in the field of Chinese, knowledge base HowNet has been introduced as a prior knowledge, in which sememe is provided to construct attention mechanism for word representation [20,46]. Niu et al. [20] introduced HowNet knowledge base to express knowledge information, and then synthesized word semantic representation. However, most methods ignore the inaccuracy of pre-trained word embedding and the large amount demand of corpora. Compared with the existing methods, in this paper, we introduce Chinese synonym knowledge base into word representation with small data for the first time to build a multi-prototype Chinese word representation model. Our method can revise the representations of pre-trained words through similar words or related words in synonym base to achieve a much more accurate word representation for each concept of each word. The final experimental results show that our method outperforms the methods listed previously in word similarity evaluation.

A multi-prototype Chinese word representation Model

Overview

Word semantic representation has become an increasingly important problem in the field of natural language processing. Most methods often fail to express the ambiguity of words, because they produce an unique vector representation for each word. To solve this problem, we propose a Multi-prototype Chinese Word representation (MP-CWR), the specific research architecture is shown in Fig. 1. The main idea is that each word, which may reveal more than one meaning, is related to not only the context words, but also prior knowledge of words.

The model integrates synonym knowledge for lexical semantics, the specific steps include knowledge representation and word similarity evaluation. For knowledge representation, first query the synonym word set $sense_i$ of word w_t, through a prior knowledge base, namely, Cilin. It should be noted that if w_t has multiple different meanings $\{\text{sense}_1, ..., \text{sense}_i, ..., \text{sense}_N\}$, where N represents the number of meanings for w_t, there will be different synonym sets, and each one consists of multiple words, containing different amounts of information, the ith concept set sense_i for w_t is $\left\{w_i^1, w_i^2, ..., w_i^M\right\}$, M means the number of the words in the ith concept of w_t. And then these words are assigned to different word embedding vectors based on the Skip-gram model. After that, a word dual weight mechanism is proposed to calculate the importance of words in each meaning set to distinguish the contribution of different prior knowledge, and then to assign a corresponding knowledge vector representation to the different meanings of each word. In the word similarity evaluation module, for word pairs w_1 and w_2, their respective synonyms and related word sets are obtained through knowledge representation, and the contribution of synonyms and related words to the similarity calculation is coordinated based on the coordination factor β to realize the evaluation of the semantic representation model. The model uses the Cilin as a prior knowledge to modify the semantics of words, and can obtain rich semantic representation under small corpus, reducing the dependence on corpus. At the same time, the MP-CWR model can not only capture the ambiguity of words, but also recognize inaccurate pre-trained embedding representation.

Knowledge base resources

Synonyms knowledge information can indicate intuitively the relationships between words compared to HowNet which mainly includes contextual conceptual information. In this paper, we choose the most common synonyms knowledge base in Chinese, Tongyici Cilin, for looking up synonyms of words.

Tongyici Cilin is an important tool which is able to identify the synonyms of polysemous words. We use the latest version HIT IR-Lab Tongyici Cilin (Extended) that is provided by Harbin Institute of Technology, simplified as Cilin in the following section. Each line in Cilin represents a synonym or related set coded by an 8-bit encoding, where the 8th bit is "=", "#" or "@". The state of each word is represented by each line, where "=" denotes that all words in a concept are synonyms, "#" indicates that the words are related, "@" indicates that the word exists independently without any synonym or related word. Figure 2 gives an example of the distribution for the word "Chinese medicine (中医)" in Cilin. The first layer represents the word "Chinese medicine", the second layer denotes the two concepts of the word. One concept is associated to the people who work in the Chinese medicine field, and the second is associated to the profession of Chinese medicine. The last layer indicates the synonyms of the two concepts.

Each word can match different concepts including synonym sets or related set from Cilin. An example of the results is shown in Fig. 3. The word "骄傲" (pride) matches two synonym sets in Cilin, indicating that it has two concepts, one means pride with positive sentiment, the other means arrogant with negative sentiment. The word "角度" (angle) has a synonym and related word set, respectively. Through looking up the expert knowledge resource Cilin, most of words can match the synonym or related word sets, so that we can represent the polysemous words to realize the semantic disambiguation. At the same time, it supplements the relationship among different words, thus the representation of

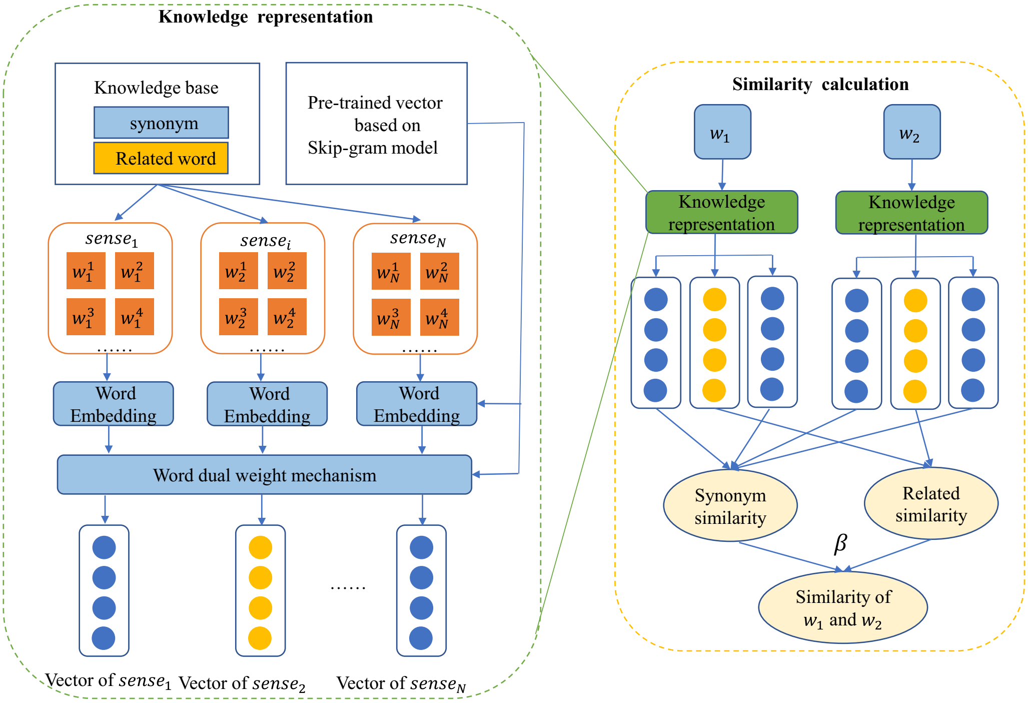


Fig. 1 Overall research framework of MP-CWR model

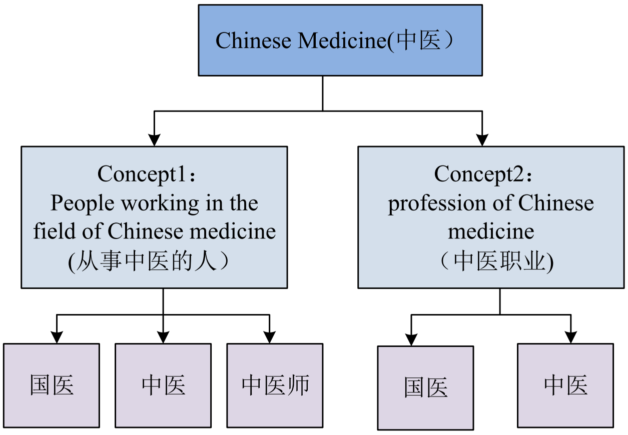
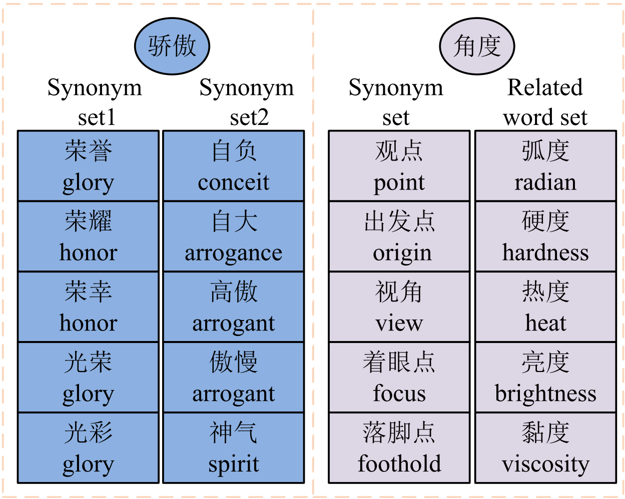

Chinese Medicine(中医）
Concept1：People working in the field of Chinese medicine (从事中医的人）
Concept2：profession of Chinese medicine （中医职业)
国医
中医
中医师
国医
中医

Fig. 2 Taking the word “Chinese Medicine” (中医) as an example to show how the Cilin structure works. Since the last layer is synonymous, the English expressions of the words are the same as “Chinese medicine”, there are no additional English annotation



骄傲

Synonym set1	Synonym set2
荣誉 glory	自负 conceit
荣耀 honor	自大 arrogance
荣幸 honor	高傲 arrogant
光荣 glory	傲慢 arrogant
光彩 glory	神气 spirit

角度

Synonym set	Related word set
观点 point	弧度 radian
出发点 origin	硬度 hardness
视角 view	热度 heat
着眼点 focus	亮度 brightness
落脚点 foothold	黏度 viscosity

Fig. 3 Two examples for looking up Cilin

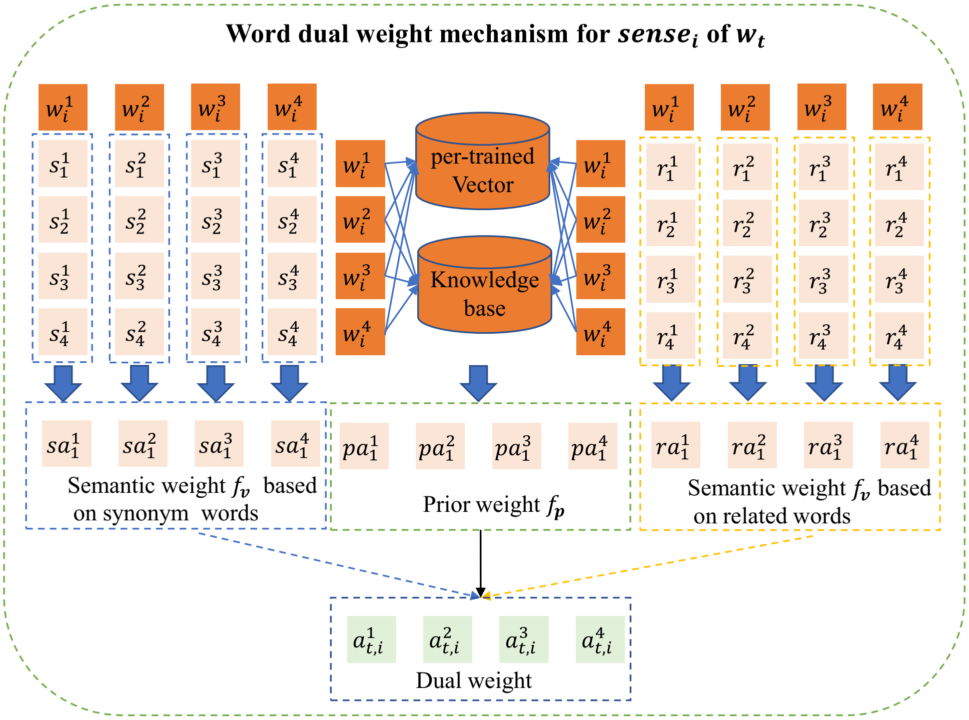
words can be fine-tuned to avoid the phenomenon that the word representation is poor due to insufficient data.

Knowledge representation

First, we define that the word w_t has N concepts, the ith concept sense$_i$ of w_t consists of a word set $\left[w_i^1, w_i^2, \ldots, w_i^M\right]$, which may be a set of synonyms or related words. They are both obtained from the Cilin, and w_i^m indicates mth word in the word concept set sense$_i$. Compared with traditional methods which only consider the context knowledge from the corpus, we incorporate semantic information from prior knowledge base Cilin to jointly learn word semantic representation.

To realize this idea, the key is to integrate prior knowledge and corpus-based knowledge. We propose a dual weight mechanism to organically combine the two parts, as shown in Fig. 4. The weight of prior knowledge is calculated by the path relationship between words in the prior knowledge base, and the weight of corpus-based knowledge is obtained

Fig. 4 Dual weight mechanism


Word dual weight mechanism for $sense_i$ of w_t
w_i^1 w_i^2 w_i^3 w_i^4
s_1^1 s_2^1 s_3^1 s_4^1
s_1^2 s_2^2 s_3^2 s_4^2
s_1^3 s_2^3 s_3^3 s_4^3
s_1^4 s_2^4 s_3^4 s_4^4
per-trained Vector
Knowledge base
r_1^1 r_2^1 r_3^1 r_4^1
r_1^2 r_2^2 r_3^2 r_4^2
r_1^3 r_2^3 r_3^3 r_4^3
r_1^4 r_2^4 r_3^4 r_4^4
sa_1^1 sa_1^2 sa_1^3 sa_1^4
Semantic weight f_v based on synonym words
pa_1^1 pa_1^2 pa_1^3 pa_1^4
Prior weight f_p
ra_1^1 ra_1^2 ra_1^3 ra_1^4
Semantic weight f_v based on related words
$a_{t,i}^1$ $a_{t,i}^2$ $a_{t,i}^3$ $a_{t,i}^4$
Dual weight

by the cosine similarity between the words in each concept. Finally, the integration results of the two is regarded as the final importance of each word in each concept.

The process of a dual weight mechanism is as follow, for ith concept with respect to w_t, there are many words in its synonym or related concept set. Assume that w_i^m and w_i^k are any two words in ith concept set, the dual weight is specifically defined as:

$$f_{\mathrm{p}}\left(w_i^m, w_i^k\right) = \mathrm{sim}\left(w_i^m, w_i^k\right), \tag{1}$$

$$f_{\mathrm{v}}\left(w_i^m, w_i^k\right) = \frac{\left\langle \mathbf{v}_o\left(w_i^m\right) \cdot \mathbf{v}_o\left(w_i^k\right)\right\rangle}{\|\mathbf{v}_o(w_i^m)\| \|\mathbf{v}_o\left(w_i^k\right)\|}, \tag{2}$$

where $f_{\mathrm{p}}\left(w_i^m, w_i^k\right)$ indicates the importance of prior knowledge, $sim(w_i^m, w_i^k)$ represents the similarity obtained by the hypernyms and hyponyms relationship between words, can be calculated thanks to the method provided by Zhu et al. [47], $f_{\mathrm{v}}\left(w_i^m, w_i^k\right)$ indicates the semantic context importance, which is obtained by pre-trained vectors through point multiplication with Skip-gram model. $\mathbf{v}_{\mathrm{o}}\left(w_i^m\right)$ and $\mathbf{v}_{\mathrm{o}}\left(w_i^k\right)$ denote the initial embedding obtained by the pre-trained model.

The important weight $a_{t,i}^m$ of mth word in the ith concept of w_t is defined as:

$$a_{t,i}^m = \sum_{k=1}^{M} f_{\mathrm{p}}\left(w_i^m, w_i^k\right) \cdot f_{\mathrm{v}}\left(w_i^m, w_i^k\right), \tag{3}$$

Furthermore, the dual weight-based representation $\hat{v}_{t,i}^m$ of the mth word in ith concept of the word w_t can be obtained, which is defined as follows:

$$\hat{v}_{t,i}^m = a_{t,i}^m \cdot v_{t,i}^m, \tag{4}$$

where $v_{t,i}^m$ means the original pre-trained vector representation of mth word in the ith concept of w_t. Finally, the word embedding representations of all words in the ith concept of the word w_t are combined to obtain the representation of the ith concept of the word:

$$\hat{v}_{t,i} = \frac{1}{M} \sum_{m=1}^{M} \hat{v}_{t,i}^m, \tag{5}$$

where M represents the number of words in the ith concept set with respect to word w_t.

Word similarity

By landing the word representation module to the word similarity task, the semantic representation effect is evaluated. As for any two words w_1 and w_2, we first obtain the meaning of each word through querying knowledge base resources Cilin, and some of which are represented by synonymous word sets, and some are composed of related word sets. Then through the "Knowledge base resources" section, we can obtain different concept vector representations for each two word w_1

and w_2, using S and R, respectively, to distinguish synonymous and related word set of ith concept of the two words.

$w_1 : \{\hat{v}^S_{w_1,i_1}, \hat{v}^S_{w_1,i_2}, \ldots, \hat{v}^S_{w_1,i_{n_1}}\}\{\hat{v}^R_{w_1,i_1}, \hat{v}^R_{w_1,i_2}, \ldots, \hat{v}^R_{w_1,i_{n_2}}\}$

$w_2 : \{\hat{v}^S_{w_2,i_1}, \hat{v}^S_{w_2,i_2}, \ldots, \hat{v}^S_{w_2,i_{m_1}}\}\{\hat{v}^R_{w_2,i_1}, \hat{v}^R_{w_2,i_2}, \ldots, \hat{v}^R_{w_2,i_{m_2}}\}$

Among them, $n_1 + n_2 = N_1$, $m_1 + m_2 = M_1$, N_1 and M_1 are the concept numbers of words w_1 and w_2, respectively. For the multiple concepts of the words w_1 and w_2, we calculate the vector cosine similarity based on the synonyms and related words, respectively, and obtain the synonym similarity $s(i, j)$ and related word similarity $r(i, j)$ of any two concept sets i and j.

$$s(i, j) = \frac{\langle \hat{v}^S_{w_1,i} \cdot \hat{v}^S_{w_2,j} \rangle}{\|\hat{v}^S_{w_1,i}\| \|\hat{v}^S_{w_2,j}\|}, \tag{6}$$

$$r(i, j) = \frac{\langle \hat{v}^R_{w_1,i} \cdot \hat{v}^R_{w_2,j} \rangle}{\|\hat{v}^R_{w_1,i}\| \|\hat{v}^R_{w_2,j}\|}, \tag{7}$$

Then, we use the combination method to calculate the word similarity $sim(w_1, w_2)$, the definition is as follows:

$$sim(w_1, w_2) = \frac{\sum_i^{n_1} \sum_j^{m_1} s(i, j) + \beta(\sum_i^{n_2} \sum_j^{m_2} r(i, j)}{n_1 m_1 + n_2 m_2}, \tag{8}$$

where $sim(w_1, w_2)$ indicates the similarity of words w_1 and w_2, β represents the coordination factor, using to adjust the importance of synonyms and related word sets, it will be explored in the following experiments and the optimal value is determined to be 1.

Experiments

Data and experimental settings

For model training, we use the corpora from Sogou Labs, namely, SogouCA[1], which is provided by [48]. To extract the noise of the corpus, we use the existing Chinese stopword dictionary, and treat the categories, such as "it" and "of" as meaningless words, and finally remove them. To verify the broad applicability of our model, we also use some pre-trained word vectors[2] provided by [50], including Baidu Encyclopedia, Wikipedia, People's Daily News, Sogou News, Zhihu QA, Weibo and Literature.

For model evaluation, we choose the WordSim-240 [22], WordSim-297 [49][3] and RG35 [51] to obtain the performance of word similarity computation. The evaluation data WordSim-240 has 240 word pairs which are mainly related words, WordSim-297 has 297 words pairs which are mainly similar word pairs, and RG35 has 35 words pairs. Specially, WordSim-240 and WordSim-297 are both ordered in a decreasing manner according to their relevance or similarity.

To evaluate the effectiveness of our proposed method MP-CWR, we use Spearman and Pearson correlation coefficient, which are widely applied in word similarity task.

Assume that $D = \left\{\left(w_1^1, w_1^2, X_1\right), \ldots, \left(w_n^1, w_n^2, X_n\right), \ldots, \left(w_N 1^1, w_N^2, X_N\right)\right\}$ is used to represent each evaluation data set, N represents the total number of word pairs, $\left(w_n^1, w_n^2, X_n\right)$ is the nth word pair, w_n^1 and w_n^2 indicate the two words in nth word pair, X_n is the nth gold-standard similarity score. Through our MP-CWR model, we can predict the similarity Y_n of the nth word pair, and then get two sequences $X = X_1, \ldots, X_n, \ldots, X_N$ and $Y = Y_1, \ldots, Y_n, \ldots, Y_N$. The key to the evaluation of the similarity task is to find the correlation between the two sequences. The Pearson (r) is defined as

$$r = \frac{\sum_n \left(X_n - \bar{X}\right)\left(Y_n - \bar{Y}\right)}{\sqrt{\sum_n \left(X_n - \bar{X}\right)^2}\sqrt{\sum_n \left(Y_n - \bar{Y}\right)^2}}, \tag{9}$$

where $\bar{X}$ and $\bar{Y}$ are the average value of two sequences X and Y.

The Spearman correlation coefficient (ρ) is defined as

$$\rho = 1 - \frac{6\sum_{n=1}^{N} \left(R_{X_n} - R_{Y_n}\right)^2}{N(N^2 - 1)}, \tag{10}$$

where R_{X_n} and R_{Y_n} are the rank of X_n in X and the rank of $Y_n$in Y, respectively.

In the experiment, Skip-gram model is chosen as the basic pre-trained method with 100 dimensions. The parameters of our model are similar to Word2Vec model with a window of 5, a minimum word frequency of 20, negative adoption and a sampling rate of 0.001. Other comparative experimental models are trained based on the same corpus and use the same parameters. Because of the excessive training loss of BERT [44], we use pre-trained Chinese words embedding BERT-Base-Chinese[4], and vector dimension of each word is 768. In the experiment, we train each evaluation data set for five times to get the average results, so as to ensure the reliability of models.

[1] http://www.sogou.com/labs/resource/ca.php.

[2] https://github.com/Embedding/Chinese-Word-Vectors.

[3] https://github.com/Leonard-Xu/CWE/tree/master/data.

[4] https://www.worldlink.com.cn/osdir/bert-as-service.html.

Word similarity

Experiments of algorithmic advantage

We utilize about 1 million of SogouCA corpora, to conduct evaluation experiments based on Pearson(r) and Spearman(ρ). We take the similarity task on different models.

1. CiLin-based method [47]: This method calculates the similarity of words through the path distance between words in CiLin. Cilin sets an 8-bit encoding for each word, and the distance between words is the difference between their encoding.

2. HowNet-based method [47]: This method calculates the similarity by the concept sets of different words in HowNet.

3. The Skip-gram method [15,28]: It is based on the vector representation of the current word to predict the vector representations of context words.

4. The CBOW method [15,28]: It predicts the vector representation of the current word by context words.

5. The SAT method [20]: It uses the attention mechanism for knowledge representation, and then synthesizes the semantic representation of words.

6. The CWE method [22]: It is proposed to obtain multiple-prototype character embedding for word similarity task.

7. The SCWE model [23]: It considers the Chinese word and internal structure character to learn the word embedding.

8. The BERT [44]: It pre-trains deep bidirectional representation according to the context of all layers.

9. Our MP-CWR model: In this model, we choose the average between the synonym set and the related set from Cilin as prior knowledge, that is $\beta = 1$.

Table 2 shows the performances of MP-CWR and other existing models on the word similarity task. In general, the MP-CWE model is superior to other existing models. In addition, we can observe that:

1. The performance of the lexicon methods based on HowNet and Cilin is very unstable. This is because if a word pair exists in the lexicons, the similarity between words can be calculated. However, if it is not included, the similarity cannot be calculated, which also reflects the disadvantage of the lexicon-based method, that is, the semantic knowledge of a word that is not in the lexicon or knowledge base resources cannot be obtained. On the whole, the CiLin-based method is better than HowNet.

2. The Skip-gram model is the best method based on multiple word embedding methods, and the overall performance of different word embedding methods is less volatile. The performance of small evaluation data set (RG35) is better than that of the larger data sets (WordSim-240 and WordSim-297). Meanwhile, though vector dimension of BERT method is higher than that of the existing methods, the effect is worse. It is found that most words are highly similar whether they are similar or not in real world. This may be because BERT considers the context information of all layers, so that any two words are highly related. This also shows that BERT model is not suitable for word similarity task.

3. The proposed model MP-CWR surpasses other existing methods, including lexion-based method, word embedding method. Compared with the HowNet and Cilin methods, when there are many word pairs (such as WordSim-240 and WordSim-297), the performance of MP-CWR model is significantly improved, and the result is more stable. Compared with the existing word embedding methods, the effect of MP-CWR method is also significantly improved by adding synonym knowledge, especially for the RG35 data set. In general, MP-CWR model has achieved excellent results in the three data sets, which shows that the method of integrating synonym knowledge into word embedding can effectively represent words and enrich the semantic representation of words.

Applicability experiments

To study the performance of the MP-CWR model, we use various Chinese word pre-trained vectors from different corpora by Skip-gram with negative-sampling (SGNS)[50] to calculate the word similarity using WordSim-297 data set. Because the similarity of words in WordSim-297 data set is in descending order, to further explore the significance of synonym information for model, we divided WordSim-297 into three classes (top 100, top 200, all) and analyzed the various cases, respectively.

From Table 3, we can see that the performances of our model are better than word2vec, especially for top 100 and top 200, which demonstrates that the MP-CWR model is more suitable to analyze similar words. In addition, the effect is improved significantly for People's Daily News, Baidu Encyclopedia and Chinese Wikipedia data sets. The top 100 word-pairs outperforms the traditional Skip-gram model by improving more than 10%, and the top 200 data increased by more than 5%. Compared with other data sets, the original corpora of these three data sets are written by professionals and the qualities are high. The performance indicates that our method has good applicability and scalability to different corpora.

Qualitative analysis for detecting nearest neighbors

We select some polysemous and monosemous words to verify the validity of our MP-CWR model for detecting nearest neighbors, and compare the results with existing methods, such as Word2vec [15], CWE [22] and SAT [20].

Table 2 Similarity evaluation experiments

Methods	WordSim-240		WordSim-297		RG35	
	ρ	r	ρ	r	ρ	r
HowNet	0.0089	– 0.0537	0.2573	0.1908	0.1394	– 0.0101
CiLin	0.0427	0.162	0.3013	0.4037	0.7939	0.8606
Skip-Gram	**0.5514**	0.5390	0.5820	0.5866	0.5945	0.6601
CBOW	0.51345	0.5155	0.51345	0.5155	0.5804	0.638
CWE	0.5217	0.5304	0.5577	0.5616	0.6432	0.6497
SCWE	0.5292	0.5354	0.5513	0.5596	0.5851	0.591
BERT	0.3250	0.3687	0.4141	0.4407	0.7040	0.7364
SAT	0.5127	0.5224	0.5534	0.5532	0.4313	0.4634
MR-CEW	0.5451	**0.5517**	**0.5919**	**0.5951**	**0.9143**	**0.8784**

Table 3 Similarity evaluation experiment on different corpora

Corpora	Size(Bytes)	Model	WordSim-297		
			top 100	top 200	all
Literature	511M	Word2Vec	0.4232	0.5094	0.6094
		MP-CWR	**0.5124**	**0.5204**	0.5850
Weibo	531M	Word2Vec	0.4108	0.4277	0.5554
		MP-CWR	**0.4770**	**0.4355**	0.5144
Zhihu	707M	Word2Vec	0.5393	0.4991	0.6363
QA		MP-CWR	**0.5961**	**0.5174**	0.6125
Chinese	960M	Word2Vec	0.3696	0.4234	0.5768
Wikipedia		MP-CWR	**0.4836**	**0.4870**	0.5761
People’s	972M	Word2Vec	0.4834	0.4398	0.5851
Daily News		MP-CWR	**0.5655**	**0.4894**	**0.5968**
Sogou	994M	Word2Vec	0.5288	0.4897	0.6131
		MP-CWR	**0.5967**	**0.5076**	0.5813
Baidu	1.69G	Word2Vec	0.4171	0.4580	0.5946
Encyclopedia		MP-CWR	**0.5522**	**0.5156**	**0.6151**

The values in bold represent the best-performing results in this column of indicators

Nearest neighbors for polysemous words

We take polysemous words “骄傲” (pride) and “队伍”(team) as examples to find their nearest neighbors.

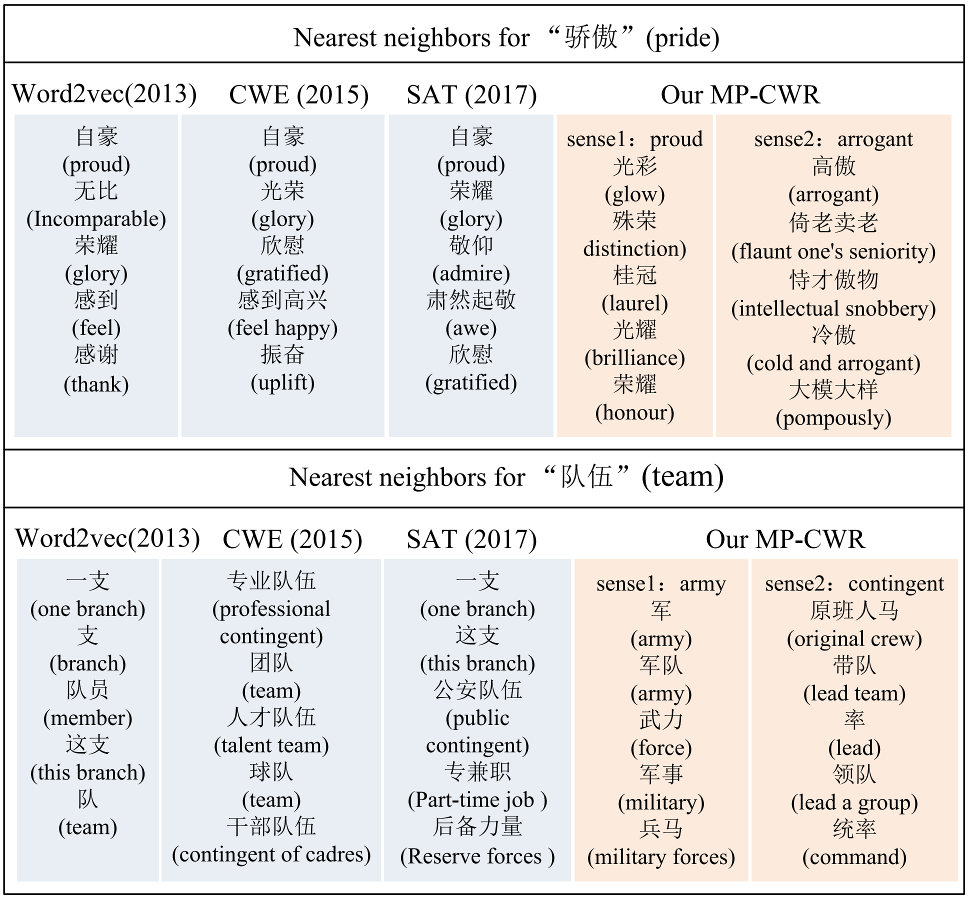
From Fig. 5, the word vectors obtained by word2vec, CWE and SAT, as we would expect, contain only a single meaning. Even though CWE and SAT both construct a multi-prototype model, in the final stage, the results are combined in a unique word vector. Meaning uniqueness appears to be a great drawback, since these models could not be able to express polysemous words. Through the MP-CWR model proposed in this paper, we can see that the words “骄傲”(pride) and “队伍”(team) have two meanings after Cilin’s refining. The two meanings of the first word are proud and arrogant. Important to note that a distinctive characteristic of our model is the fact that it is able to represent words in a more complete and reasonable manner.

Nearest neighbors for monosemous words

In addition to the nearest neighbor experiments of polysemous words, we also analyze the performance of our model in dealing with monosemous words by choosing Chinese words “青蛙”(frog) and “怀孕”(pregnancy) as examples.

Figure 6 shows that when analyzing the nearest neighbors of the word “青蛙”(frog), the results obtained by the other existing models are mostly related words, but with very different meanings. On the other hand, the method proposed in this paper detects words that have also similar meaning, which shows its better ability to find nearest neighbors for monosemous words with respect to the other models. The qualitative analyses show that the MP-CWR model yields a better performance than the other methods, since it is not only applicable to polysemous words, but also to monosemous words.

Fig. 5 Nearest neighbors for polysemous words(highest cosine similarity



Nearest neighbors for “骄傲”(pride)

Word2vec(2013)	CWE (2015)	SAT (2017)	Our MP-CWR	
			sense1：proud	sense2：arrogant
自豪 (proud)	自豪 (proud)	自豪 (proud)	光彩 (glow)	高傲 (arrogant)
无比 (Incomparable)	光荣 (glory)	荣耀 (glory)	殊荣 distinction)	倚老卖老 (flaunt one's seniority)
荣耀 (glory)	欣慰 (gratified)	敬仰 (admire)	桂冠 (laurel)	恃才傲物 (intellectual snobbery)
感到 (feel)	感到高兴 (feel happy)	肃然起敬 (awe)	光耀 (brilliance)	冷傲 (cold and arrogant)
感谢 (thank)	振奋 (uplift)	欣慰 (gratified)	荣耀 (honour)	大模大样 (pompously)

Nearest neighbors for “队伍”(team)

Word2vec(2013)	CWE (2015)	SAT (2017)	Our MP-CWR	
			sense1：army	sense2：contingent
一支 (one branch)	专业队伍 (professional contingent)	一支 (one branch)	军 (army)	原班人马 (original crew)
支 (branch)	团队 (team)	这支 (this branch)	军队 (army)	带队 (lead team)
队员 (member)	人才队伍 (talent team)	公安队伍 (public contingent)	武力 (force)	率 (lead)
这支 (this branch)	球队 (team)	专兼职 (Part-time job)	军事 (military)	领队 (lead a group)
队 (team)	干部队伍 (contingent of cadres)	后备力量 (Reserve forces)	兵马 (military forces)	统率 (command)

Model parameters analysis

To verify the stability of our model under different parameters, we explored three parameters in the model including the coordination factor β, corpora size, and vector dimension.

Coordination factor β

We first train our MP-CWR model using about half of SogouCA corpora, and then explore the performance of the model on word similarity tasks under different factor β.

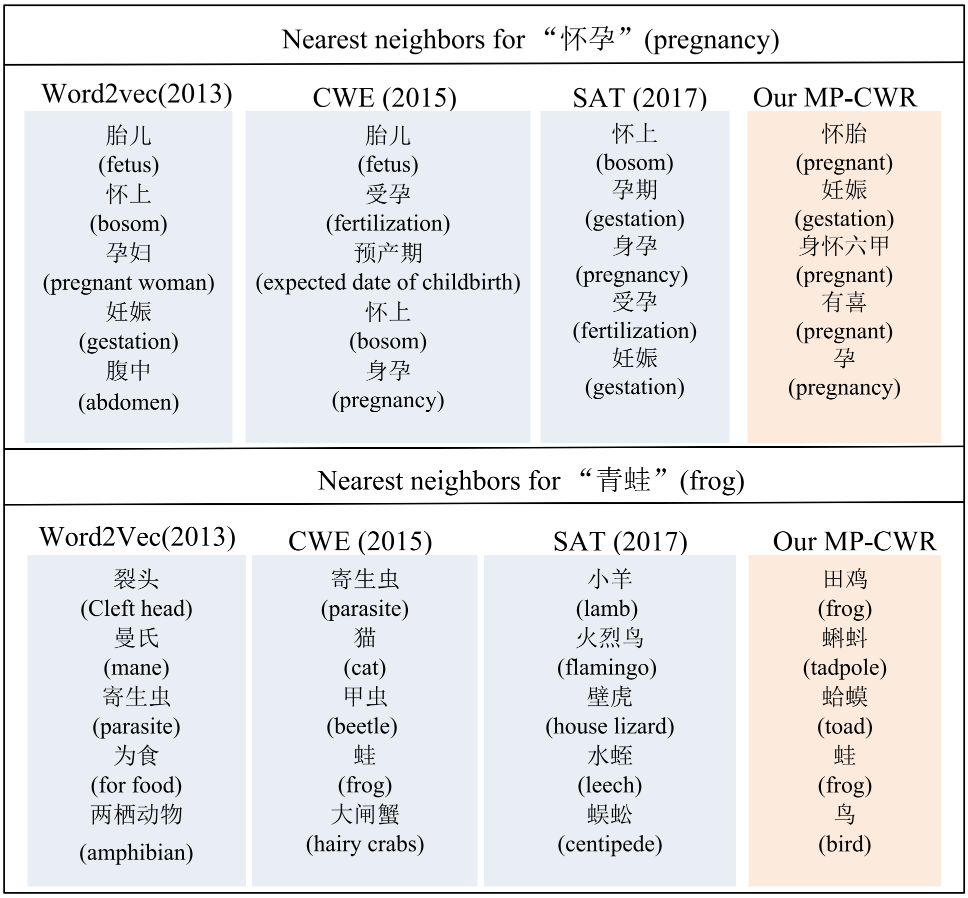
As shown in Fig. 7, the results have similar rules with the increase of β for different evaluation data sets, first, they increase gradually, and then keep flat. Especially when $\beta = 1$, the best effect is obtained, and it indicates that synonyms and related words in the Cilin have the same importance. It also proves that it can complement the information of word pair by integrating synonyms and context knowledge of words.

Influence of different corpus size

To explore the influence of different training corpora of our model, we extract 0.1 million, 0.3 million, 0.5 million, 0.7 million and 1.1 million from all data which contains 1.1 million pieces for experiments. Then we have carried out the experiments on WordSim-240 and WordSim-297 data sets, respectively, by fixing the coordination factor and vector dimension. The performance of our model using different corpora on word similarity task was explored by analyzing the evaluation index Pearson correlation coefficient. The experimental results are shown in Fig. 8.

As can be seen from Fig. 8, on the WordSim-240 evaluation data set, with the increase of corpus, the Pearson correlation coefficient of similarity task fluctuates greatly, which is similar to the trend of traditional Skip-gram method. On the WordSim-297 data set, with the increase of training corpus, our model and Skip-gram model have certain regularity, the Pearson correlation coefficient is rising. But compared with Skip-gram method, the MP-CWR method always has better performance. In a word, the effect of the proposed model is better than the existing method on both kinds of data, especially on WordSim-297 data set. At the same time, the corpus size is not stable for the change of the results, which also shows that our model integrating prior knowledge is more stable, not easily affected by the change of data size, and more suitable for lexical semantic representation under small data.

Fig. 6 Nearest neighbors for monosemous words(highest cosine similarity)



Nearest neighbors for “怀孕”(pregnancy)

Word2vec(2013)	CWE (2015)	SAT (2017)	Our MP-CWR
胎儿 (fetus)	胎儿 (fetus)	怀上 (bosom)	怀胎 (pregnant)
怀上 (bosom)	受孕 (fertilization)	孕期 (gestation)	妊娠 (gestation)
孕妇 (pregnant woman)	预产期 (expected date of childbirth)	身孕 (pregnancy)	身怀六甲 (pregnant)
妊娠 (gestation)	怀上 (bosom)	受孕 (fertilization)	有喜 (pregnant)
腹中 (abdomen)	身孕 (pregnancy)	妊娠 (gestation)	孕 (pregnancy)

Nearest neighbors for “青蛙”(frog)

Word2Vec(2013)	CWE (2015)	SAT (2017)	Our MP-CWR
裂头 (Cleft head)	寄生虫 (parasite)	小羊 (lamb)	田鸡 (frog)
曼氏 (mane)	猫 (cat)	火烈鸟 (flamingo)	蝌蚪 (tadpole)
寄生虫 (parasite)	甲虫 (beetle)	壁虎 (house lizard)	蛤蟆 (toad)
为食 (for food)	蛙 (frog)	水蛭 (leech)	蛙 (frog)
两栖动物 (amphibian)	大闸蟹 (hairy crabs)	蜈蚣 (centipede)	鸟 (bird)

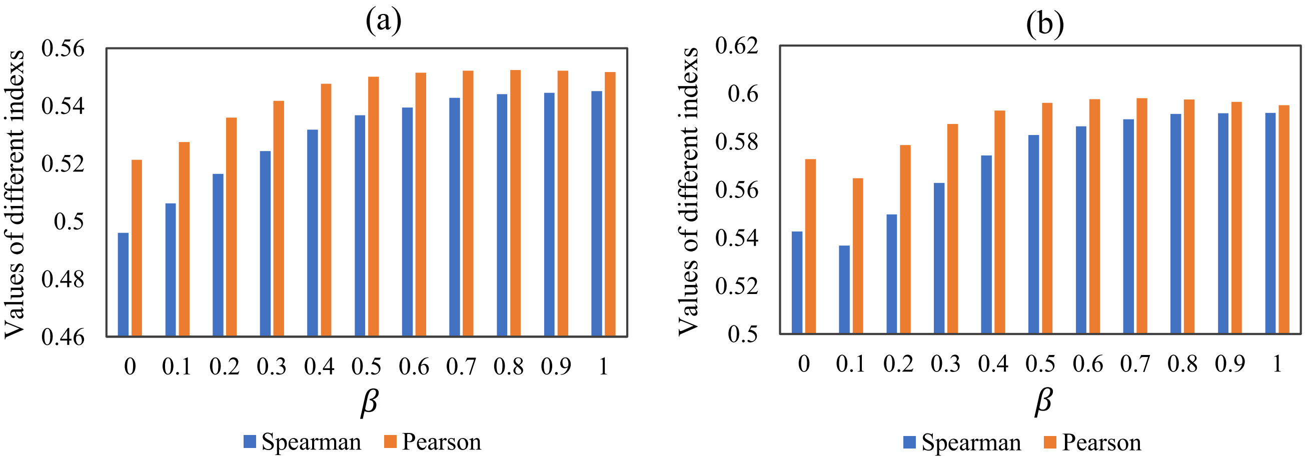


Fig. 7 Results of word similarity for different β on **a** WordSim-240 and **b** WordSim-297 using Spearman and Pearson correlation coefficient

Size of vector dimension

To explore the performance of our model in different vector dimensions, the experiments are conducted on WordSim-240 and WordSim-297 data sets by fixing the coordination factor and the size of the training corpus. The performance of the MP-CWR model on word similarity task in different embedded dimensions is explored by analyzing the evaluation index Pearson correlation coefficient. The experimental results are shown in Fig. 9.

The results shown in Fig. 9 show that the evaluation data sets of WordSim-240 and WordSim-297 have similar rules. With the increase of dimensions, the Pearson correlation coefficient of our model first increases and then decreases, the best effect is when the dimension is 200 or 300. Second, it can be clearly seen that MP-CWR method has better

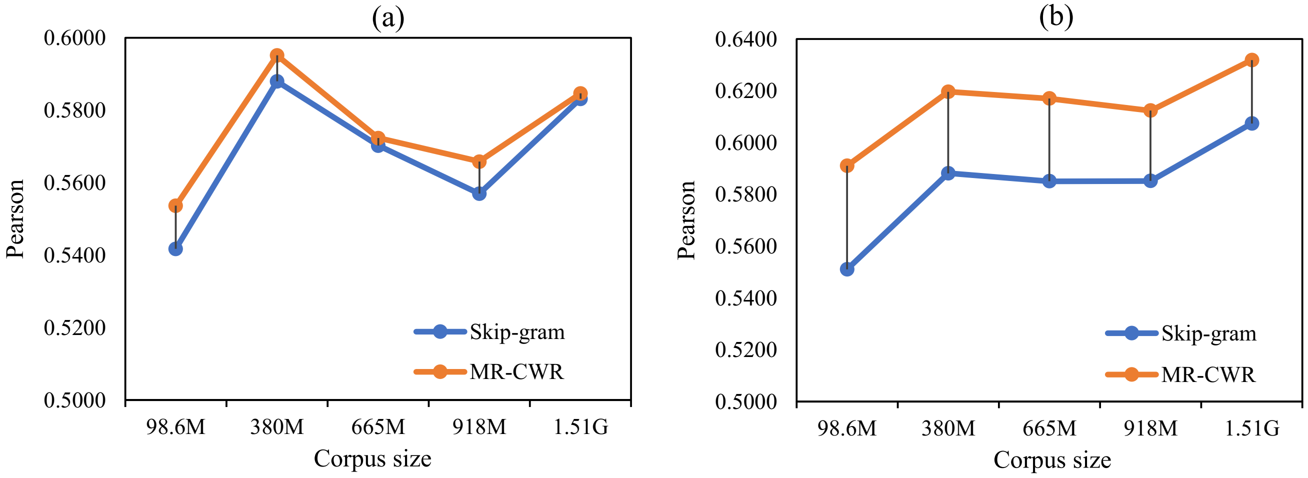


Fig. 8 Results of word similarity for different corpus sizes on **a** WordSim-240 and **b** WordSim-297

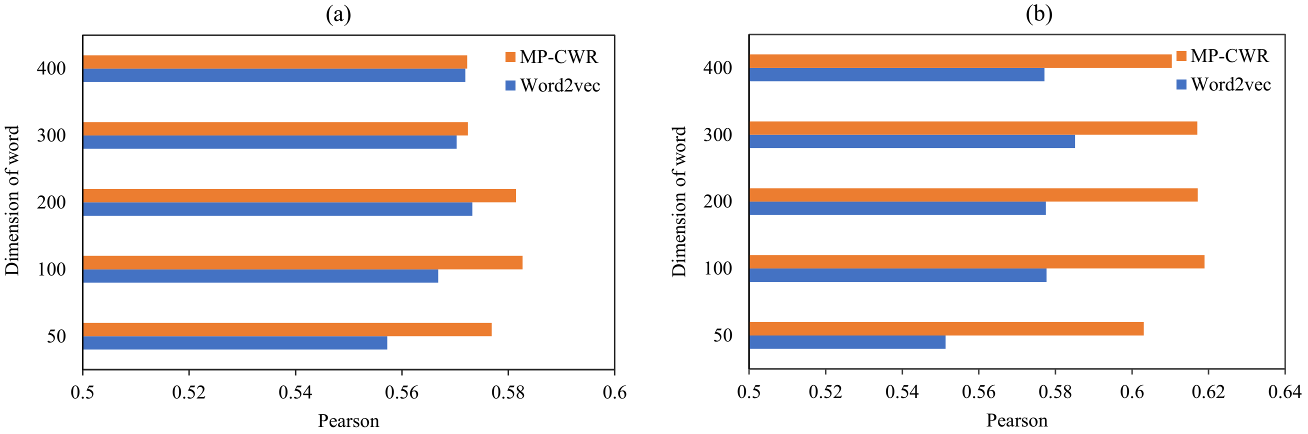


Fig. 9 Results of word similarity for different vector dimensions on **a** WordSim-240 and **b** WordSim-297

performance than traditional pre-trained method in semantic capture of model under low dimensional vector, which may be because that MP-CWR model can integrate rich expert knowledge and play a complementary role in semantic information in low dimension. The overall results show that our model is better than the existing Skip-gram in the analysis of small data, and the model can reach a stable level faster. The results indicate that our proposed model integrates knowledge into the pre-trained word embedding, and the model can enrich the semantic information of words. In addition, what needs to be explained is that higher dimensions do not guarantee better performance for two reasons. One is because the increase in dimensions may lead to confusion of information, resulting in inaccurate results. Another reason is that there are many noises in the training data itself, and the increase in dimensions will increase the noise, which may eventually lead to inaccurate results. Hence the vector dimension is not proportional to the performance of our model.

Conclusion

We propose a multi-prototype Chinese word representation model based on expert knowledge base for Chinese word similarity. Compared with the existing methods, not only considering the context of corpus, our model incorporates prior knowledge from synonym knowledge base. We also fine-tune the word embedding by searching for the less effective pre-trained vectors, avoiding the difficulty in obtaining high-quality corpus. This is our revised approach to the pre-trained model. We search for synonyms of each word through the Tongyici Cilin, and calculate the cosine similarity between them through the vector of each word, and select words with low similarity as less effective words. Then we modify the vector of the word by the vector of the synonyms of the word. Finally, the experiments demonstrate that our MP-CWR model is useful for word similarity and nearest neighbor word detection tasks, and the performances has an overall improvement compared with the existing methods. In addition, our MP-CWR model also has widely applicability for words, and its performances, which obtains under small corpus, are comparable to that of large data volume.

Nowadays, the research on word embedding is almost focused on big data and stronger GPUs to improve the accuracy of word representation. This will inevitably lead to a "computational burden" which requires a great need of computational power. It is necessary to improve the accuracy in small data to obtain similar results when using large data. In addition, due to the special structure of Chinese language, there are several important and related arguments that could be of great relevance, such as stroke order, radical of Chinese characters and the representation of words which are missing in the corpus vocabulary. In future studies, we would like to concentrate on these above issues for word representation.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No. 61801440), the High-quality and Cutting-edge Disciplines Construction Project for Universities in Beijing (Internet Information, Communication University of China), State Key Laboratory of Media Convergence and Communication (Communication University of China), and the Fundamental Research Funds for the Central Universities.

Declaration

Conflict of interest The authors declare no competing interests.

References

1. Smarandache F, Colhon M, Vladutescu S et al (2019) Word-level neutrosophic sentiment similarity. Appl Soft Comput 80:167–176
2. Khalid U, Hussain A, Arshad MU et al (2021) Co-occurrences using Fasttext embeddings for word similarity tasks in Urdu. arXiv:2102.10957
3. Janai S et al (2021) Speech-to-speech conversion: an approach to enhance the speech intelligibility of dysarthric speaker. IJACI 12(1):184–206
4. Lastra-Diaz JJ, Goikoetxea J, Taieb MAH et al (2021) A large reproducible benchmark of ontology-based methods and word embeddings for word similarity. Inf Syst 96:101636
5. Huang D, Pei J, Zhang C et al (2018) Incorporating prior knowledge into word embedding for Chinese word similarity measurement. ACM Trans Asian Low-Resourc Lang Inf Process (TALLIP) 17(3):1–21
6. Lee YY, Ke H, Yen TY et al (2020) Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement. J Am Soc Inf Sci 71(6):657–670
7. Sivakumar S, Rajalakshmi R (2021) Analysis of sentiment on movie reviews using word embedding self-attentive LSTM. IJACI 12(2):33–52
8. Kwon S, Oh D, Ko Y (2021) Word sense disambiguation based on context selection using knowledge-based word similarity. Inf Process Manag 58(4):102551
9. Jia L, Tang J, Li M et al (2021) TWE-WSD: An effective topical word embedding based word sense disambiguation. CAAI Trans Intell Technol 2021:5
10. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, pp 3104–3112
11. Jitao X U, Crego J M, Senellart J (2020) Boosting neural machine translation with similar translations. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 1580-1590
12. Rahman N, Borah B (2019) Improvement of query-based text summarization using word sense disambiguation. Compl Intell Syst 2019:1–11
13. Yoon W, Yeo YS, Jeong M et al (2002) Learning by semantic similarity makes abstractive summarization better. arXiv:2002.07767
14. Mundotiya RK, Yadav N (2021) Forward context-aware clickbait tweet identification system. IJACI 12(2):21–32
15. Mikolov T, Sutskever I, Chen K et al (2013a) Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 26:3111–3119
16. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
17. Reisinger J, Mooney R (2010) Multi-prototype vector-space models of word meaning. In: Human Language Technologies: the 2010 annual conference of the north american chapter of the association for computational linguistics, pp 109–117
18. Huang EH, Socher R, Manning CD et al (2012) Improving word representations via global context and multiple word prototypes. In: Proceedings of the 50th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 873–882
19. Tian F, Dai H, Bian J et al (2014) A probabilistic model for learning multi-prototype word embeddings. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, pp 151–160
20. Niu Y, Xie R, Liu Z et al (2017) Improved word representation learning with sememes. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 2049–2058
21. Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38(11):39–41
22. Chen X, Xu L, Liu Z et al (2015) Joint learning of character and word embeddings. In: Twenty-fourth international joint conference on artificial intelligence, pp 1236–1242
23. Xu J, Liu J, Zhang L et al (2016) Improve Chinese word embeddings by exploiting internal structure. In: Proceedings of the 2016 Conference of the North American Chapter of the association for computational linguistics: human language technologies, pp 1041–1050
24. Yu J, Jian X, Xin H et al (2017) Joint embeddings of chinese words, characters, and fine-grained subcharacter components. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 286–291
25. Sun C, Qiu X, Huang XJ (2019) VCWE: visual character-enhanced word embeddings. In: Proceedings of the 2019 conference of the north american chapter of the association for computational lin-

guistics: human language technologies, volume 1 (Long and Short Papers), pp 2710–2719
26. Dong ZD, Dong Q (2003) Hownet-a hybrid language and knowledge resource. In: Proceedings of NLP-KE. IEEE, pp 820–824
27. Mei JJ (1983) Tongyici cilin. Shanghai Lexicographical Publishing House, Shanghai
28. Mikolov T, Chen K, Corrado G et al (2013b) Efficient estimation of word representations in vector space. arXiv:1301.3781
29. Deerwester S, Dumais ST, Furnas GW et al (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407
30. Levy O, Goldberg Y (2014) Linguistic regularities in sparse and explicit word representations. In: Proceedings of the eighteenth conference on computational natural language learning, pp 171–180
31. Bengio Y, Ducharme R, Vincent P et al (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155
32. Kombrink S, Mikolov T, Karafiat M et al (2011) Recurrent neural network based language modeling in meeting recognition. In: Twelfth annual conference of the international speech communication association, pp 1045–1048
33. Lu Y, Zhang Y, Ji D (2016) Multi-prototype Chinese character embedding. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), pp 855–859
34. Bojanowski P, Grave E, Joulin A et al (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguis 5:135–146
35. Zhao Z, Liu T, Li S et al (2017) Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 244–253
36. Song Y, Shi S, Li J et al (2018) Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In: Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 2 (Short Papers), pp 175–180
37. Heo TS, Kim JD, Park CY et al (2021) Global and local information adjustment for semantic similarity evaluation. Appl Sci 11(5):2161
38. Meng Y, Huang J, Wang G et al (2020) Unsupervised word embedding learning by incorporating local and global contexts. Front Big Data 3:9
39. Yu LC, Wang J, Lai KR et al (2018) Refining word embeddings using intensity scores for sentiment analysis. IEEE/ACM Trans Audio Speech Lang Process 26(3):671–681
40. Yang L, Chen X, Liu Z et al (2017) Improving word representations with document labels. IEEE/ACM Trans Audio Speech Lang Process 25(4):863–870
41. Yang X, Mao K (2016) Task independent fine tuning for word embeddings. IEEE/ACM Trans Audio Speech Lang Process 25(4):885–894
42. Yang J, Li Y, Gao C et al (2021) Measuring the short text similarity based on semantic and syntactic information. Futur Gener Comput Syst 114:169–180
43. Peters M, Neumann M, Iyyer M, et al (2018) Deep contextualized word representations. In: Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (Long Papers), pp 2227–2237
44. Devlin J, Chang M W, Lee K et al (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (Long and Short Papers), pp 4171–4186
45. Lastra-Diaz JJ, Goikoetxea J, Taieb MAH, Garcia-Serrano A, Aouicha MB, Agirre E (2019) A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art. Eng Appl Artif Intell 85:645–665
46. Zeng X, Yang C, Tu C et al (2018) Chinese LIWC lexicon expansion via hierarchical classification of word embeddings with sememe attention. In: Proceedings of the AAAI conference on artificial intelligence, pp 5650–5657
47. Zhu X, Ma R, Sun L et al (2016) Word semantic similarity computation based on hownet and cilin. J Chin Inf Process 30(4):29–36
48. Wang C, Zhang M, Ma S et al (2008) Automatic online news issue construction in web environment. In: Proceedings of the 17th international conference on World Wide Web, pp 457–466
49. Jin P, Wu Y (2012) Semeval-2012 task 4: evaluating chinese word similarity. SEM 2012: the first joint conference on lexical and computational semantics-volume 1: proceedings of the main conference and the shared task, and volume 2: proceedings of the sixth international workshop on semantic evaluation (SemEval 2012), pp 374–377
50. Li S, Zhao Z, Hu R et al (2018) Analogical reasoning on chinese morphological and semantic relations. In: Proceedings of the 56th annual meeting of the association for computational linguistics, pp 138–143
51. Rubenstein H, Goodenough JB (1965) Contextual correlates of synonymy. Commun ACM 8(10):627–633