Improving Chinese Word Representation with Conceptual Semantics

Tingxin Wei[1, 2], Weiguang Qu[2, 3, *], Junsheng Zhou[3], Yunfei Long[4], Yanhui Gu[3] and Zhentao Xia[3]

Abstract: The meaning of a word includes a conceptual meaning and a distributive meaning. Word embedding based on distribution suffers from insufficient conceptual semantic representation caused by data sparsity, especially for low-frequency words. In knowledge bases, manually annotated semantic knowledge is stable and the essential attributes of words are accurately denoted. In this paper, we propose a Conceptual Semantics Enhanced Word Representation (CEWR) model, computing the synset embedding and hypernym embedding of Chinese words based on the Tongyici Cilin thesaurus, and aggregating it with distributed word representation to have both distributed information and the conceptual meaning encoded in the representation of words. We evaluate the CEWR model on two tasks: word similarity computation and short text classification. The Spearman correlation between model results and human judgement are improved to 64.71%, 81.84%, and 85.16% on Wordsim297, MC30, and RG65, respectively. Moreover, CEWR improves the F1 score by 3% in the short text classification task. The experimental results show that CEWR can represent words in a more informative approach than distributed word embedding. This proves that conceptual semantics, especially hypernymous information, is a good complement to distributed word representation.

Keywords: Word representation, conceptual semantics, hypernymy, similarity computation, short text classification.

1 Introduction

In most of the current natural language processing models, words are represented by vectors known as word embeddings. Word2Vec [Mikolov, Sutskever, Chen et al. (2013)]

[1] International College for Chinese Studies, Nanjing Normal University, Nanjing, 210097, China.

[2] School of Chinese Language and Literature, Nanjing Normal University, Nanjing, 210097, China.

[3] School of Computer Science and Technology, Nanjing Normal University, Nanjing, 210023, China.

[4] School of Computer Science and Electronic Engineering, University of Essex, Essex, CO4 3DQ, UK.

[*] Corresponding Author: Weiguang Qu. Email: wgqu_nj@163.com.

Received: 31 March 2020; Accepted: 08 May 2020.

CMC. doi:10.32604/cmc.2020.010813

and GloVe [Pennington, Socher and Manning (2014)] are popular and effective examples of word embedding methods used to project words into low-dimensional dense vector spaces. Carrying rich information of the word, word embedding has benefited many NLP tasks [Devlin, Zbib, Huang et al. (2014); Liu, Yang, Lv et al. (2019); Qiu, Liu, Chai et al. (2019); Polat, Yaylali and Tanay (2019)]. As word embedding is the fundamental input for all tasks, the quality of how well it represents words determines the final performance of models [Peters, Neumann, Iyyer et al. (2018)]. Learning word representation is based on the distributional hypothesis [Harris (1954)]: words appearing in similar contexts must have similar meanings. However, this hypothesis is established on the condition that the corpus is as large as the whole existing natural language; otherwise, it will suffer from context sparsity. When there are insufficient contexts for training, or contexts are not typically related to the target word, the word embedding cannot represent the word well. Furthermore, the modeling of word embedding is based on the collocation of words in contexts, which results in the embedding carrying more collocation information instead of its own conceptual meaning. Since word embedding heavily relies on the distribution of words in the corpus, word embedding is unstable because a word gets different embeddings when trained on different corpora. To address these issues, we utilize the conceptual semantics of words to enhance and stabilize the word representation from the corpus.

As a critical knowledge base in conceptual semantics, Tongyici Cilin (Cilin) [Mei, Zhu, Gao et al. (1983)] is a Chinese thesaurus in which synonyms are grouped into one synset to denote a distinct concept. It explicitly demonstrates a variety of semantic relations among words, including synonymy, hypernymy, and meronymy. Since it is manually built and revised by linguists, the semantic information and relations of words are accurate, stable, and have been proved useful in many Chinese NLP tasks [Liu, Peng, Qian et al. (2014); Li, Lv, Wang et al. (2016); Peng, Zhu, Chen et al. (2018); Wei, Chen, Shi et al. (2018)]. In this paper, we propose a Conceptual Semantics Enhanced Word Representation (CEWR) model, taking the conceptual meaning of words learned from Cilin into consideration to improve the distributed word representation.

In the experiments, we evaluate our CEWR model in word similarity computation and short text classification. The results show that our model outperforms the baselines which also utilize external knowledge bases to improve word representation. The contributions of our paper are as follows. (1) Our CEWR model integrates conceptual semantics into word representation, which makes the representation of words more accurate and complete, and increases stability. (2) Hypernymous information is used in representing the semantics of words for the first time and has been proved effective. (3) As a post-processing step, the CEWR model can be integrated with any other embedding model with high efficiency and low computational complexity.

2 Related work

Word representation is a fundamental and essential input in many downstream language processing tasks. Distributed word representation projects all words into a continuous low-dimensional semantic space that addresses the issue of data sparsity in conventional one-hot word representation. Due to its effectiveness in semantic representation, it is applied in many tasks, including language modeling [Bengio, Rejean and Pascal (2003)], syntactic parsing [Socher, Lin, Manning et al. (2011)], word sense disambiguation [Chen, Liu and Sun (2014)] and discourse relation classification [Dai and Huang (2018)]. Along with the massive progress and great achievements of deep learning in various fields, learning better word representation has become critical and drawn increasing attention from researchers. Mikolov et al. [Mikolov, Sutskever, Chen et al. (2013)] proposed CBOW and SkipGram models to learn word representation. Both models assume that the meaning of words can be well represented by the contexts in which they appear. GloVe [Pennington, Socher and Manning (2014)] uses matrix factorization on the word affinity matrix to learn word representation. Unsupervised pre-trained language models such as ELMO [Peters, Neumann, Iyyer et al. (2018)] and BERT [Devlin, Chang, Lee et al. (2018)] have contextualized representation and have achieved remarkable results in many NLP tasks. This proves that good representation is crucial for language processing. However, all of these unsupervised corpus-dependent models obtain word representation based on the distribution, and low-frequency words cannot be well represented due to insufficient information in the corpus.

To address this issue, many researchers have proposed utilizing semantic information in existing knowledge bases to improve word representation. Yu et al. [Yu and Dredze (2014)] proposed a relation-constrained model, and tried to incorporate prior knowledge into WordNet and PPDB to improve learned word embeddings. They evaluated their embeddings on the tasks of measuring word similarity and predicting human judgement, and the median reciprocal rank of word pairs with the new model was much better than that of CBOW and SkipGram. Rothe et al. [Rothe and Schűtze (2015)] proposed a model to learn embeddings for synset and lexemes in the lexical resource, and proved its effectiveness on the tasks of word similarity calculation and word sense disambiguation; the lexeme similarity and WSD accuracy were improved to 69.8 and 73.6, respectively. According to Bartusiak et al. [Bartusiak, Augustyniak, Kajdanowicz et al. (2019)], WordNet is a network that embeds relationships between distinct concepts, and it includes rarely used words and their unusual meanings that may not exist or are damped in corpora. Therefore, they created vectors for each word in WordNet, encapsulating its position-role toward all other words, and utilized WordNet2Vec for sentiment analysis on the Amazon product review dataset. The obtained F-1 score was slightly lower than that of Doc2Vec. In Chinese language processing, researchers have also made many attempts in incorporating semantic knowledge in prior knowledge resources into word representation. Chen et al. [Chen, Xu, Liu et al. (2015)] utilized the internal semantic information of words and proposed a word representation model for the joint learning of a word and its composing characters. They used the tasks of word similarity calculation and word analogy to evaluate the new word representation, and that their method outperformed CBOW, SkipGram, and GloVe. HowNet is a widely used Chinese knowledge base used for improving word representation. In HowNet, word sense is defined as a combination of sememes. Niu et al. [Niu, Xie, Liu et al. (2017)] proposed a sememe-encoded word representation learning model with the attention strategy based on HowNet. The correlation of the word similarity between the model result and human judgement was improved to 64.0 on the WordSim-297 dataset and to 61.2 on WordSim-240. As for Cilin, another informative Chinese lexical knowledge base, Yang et al. [Yang and Sun (2015)] took intricate dependencies between composing characters of words learned from Cilin to improve word representation. They demonstrated the effectiveness of the new word representation on tasks of word similarity, word analogy, and document classification. Li et al. [Li, You and Chen (2019)] proposed an algorithm of word semantic similarity computation by representing each word sense of polysems in different contexts with the help of the prior classification information from Cilin. Their experimental results show that the proposed model achieves 64.09 in Spearman correlation on the Chinese word similarity prediction dataset WordSim-297.

In this paper, we assume that the representation of words is composed of its conceptual semantics as well as distributional semantics. The combination of both conceptual semantics and distributional semantics is more stable and accurate in the representation of meaning, especially for low-frequency words that cannot be thoroughly trained and well represented in the corpus. We propose to learn a word's conceptual semantics through its synonyms and hypernyms with the help of Cilin, and integrate these semantics into the distributed word embedding to achieve better word representation.

3 Limitation of distributional word representation

Distributional word representation is based on the distributional hypothesis [Harris (1954)]: words appearing in similar contexts must have similar meanings. However, in the same paper, Harris also pointed out that the distributional structure does not give ideal coverage. Without sufficient exposure in the corpus, the word representation trained from it cannot represent the meaning very well, even for high-frequency words. To illustrate this issue, we use the word representation trained by the SkipGram model as an example. The SkipGram model aims to maximize the predictive probability of context words conditioned on the target word w. The objective function is as follows:

$$L = \sum_{i=k}^{n-k} \log P(W_{i-k}, \dots W_{i+k} \,|W_i). \tag{1}$$

The model assumes that the contexts of a word determine the information it carries. The more contexts the word has, the more information the embedding will get. However, no matter how big the training corpus is, it is impossible for all words to get enough contexts. Let us take a self-built corpus as example, which is crawled from Baidu web pages containing 68 million tokens and has a vocabulary size of around 643 thousand. The word frequency distribution in this corpus is shown in Tab. 1.

Table 1: Word frequency distribution in Baidu corpus

Frequency	>1000	500–999	100–499	20–99	5–19	1–5	total
Tokens	5176	4070	20964	55412	95122	462604	643348
Ratio	0.8	0.63	3.26	8.61	14.78	71.91	100

There are less than 10,000 words that appear over 500 times in the corpus, while over 56,000 words are collected in the "Contemporary Chinese Dictionary" [Chinese Academy of Social Sciences (2016)], the officially issued standard Chinese language dictionary, which means that over 80% of common words do not have sufficient occurrence and contexts in the corpus. Moreover, due to the drawback of unsupervised learning, some contextual terms might not be related to the target word, or there may be insufficient contextual terms to represent the word's meaning even though the frequency of the target word is not very low. Tab. 2 shows some words and their contexts from the corpus.

Table 2: Examples of unrelated contexts

Tokens	Frequency	Contexts *
白霜 White Frost	50	双手 (hands) 结 (bond) 钻心 (core-bit) 满载 (fully loaded) 轻轻 (gently) 上一层 (upper layer) 一双 (a pair) 屋上 (on roof) 一片 (a slice)
工作狂 workaholic	85	世界 (world) 面对 (face) 上司 (boss) 员工(employee) 戏路 (range of acting role types) 一般 (general) 更新 (update) 考古学家 (archaeologist) 功效 (effect) 几乎 (almost) 褒奖 (praise) 荐(recommend) 禁止 (forbid)

*window size is 5, and stop words are excluded.

"白霜 (white frost)" appears 50 times in the corpus, and hence there are only nine tokens in its contexts except for stop words, none of which share the same meaning of "白霜". The same happens to the word "工作狂 (workaholic)". Trained with these contexts, one can imagine that scarce semantic information of the target word is encoded in the final word embedding. Without efficient information being encoded, the word embedding cannot be precise and subtle enough to represent its meaning, which makes it corpus-dependent and thus unstable in application.

Meanwhile, even for high-frequency words, sufficient occurrence does not necessarily lead to good meaning representation. Since the objective function maximizes the predictive probability of context words, the embedding indicates more about which context it appears in than what concept it refers to. Since word embedding is composed of vectors that cannot be read and understood directly, we use words with the closest cosine distances to the target words to better understand the word representation. Tab. 3 shows the words most similar to "咖啡 (coffee)" and "消费者 (consumer)" in the Baidu corpus.

Table 3: The most similar words with closest cosine distances to target words

Tokens	Frequency	Most similar words
咖啡 coffee	936	甜品 (dessert) 奶茶 (milk tea) 日式 (Japanese style) 三文治 (sandwich) 甜点 (sweet dessert) 轩尼 (a name) 港式 (Hong Kong style) 餐 (meal) 下午茶 (afternoon tea) 餐厅 (restaurant)
消费者 consumer	3341	消费 (consume) 维权 (safeguard rights) 生产者(producer) 顾客 (customer) 权益 (legal right) 保护法 (protection law) 客户 (client) 购买 (buy) 决策 (make policy) 造假者 (counterfeiters)

As shown in Tab. 3, the words similar to "咖啡 (coffee)" are mostly things frequently appearing in the same scene with coffee, such as "甜品 (dessert)", "三明治 (sandwich)", and “餐厅 (restaurant)”. However, people cannot obtain a clear image of what coffee is from its representation. For “消费者 (consumer)”, since it tends to appear in the same contexts as word “消费 (consume)”, the cosine distance between it and “消费 (consume)” is much closer than that between “消费 (consume)” and “顾客 (customer)”, despite “顾客 (customer)” being semantically closer to “消费者 (consumer)”. From this perspective, distributed word representation represents more contexts than the meaning.

In studies of lexical semantics, it is generally agreed that the meaning of a word consists of its conceptual meaning and distributive meaning. The conceptual meaning refers to the general or essential attributes of things that are reflected in the human mind. It is the basis and core component of a word’s semantics, and usually lies in dictionaries and other semantic resources. Inspired by this, we utilize the conceptual meaning of words extracted from semantic resources to complement distributional word representation. Since semantic knowledge in dictionaries and thesauri is manually summarized and annotated, and therefore the concept has very good stability and universality to represent the word’s meaning, we believe that incorporating conceptual semantics with distributed word embedding will improve the stability and accuracy of word representation.

4 Conceptual semantics enhanced word representation

In this section, we present our CEWR model, which considers the synonym concept and hypernym concept in word representation. First, we introduce the Tongyici Cilin thesaurus and its semantic network. Then, we propose two models: the Synset Concept Encoded Embedding Model (SEEM) and the Hypernym Concept Encoded Embedding Model (HEEM). They extract, respectively, the synset semantics and hypernym semantics of words and integrate them with the distributional representation. The structure of CEWR is shown in Fig. 1.

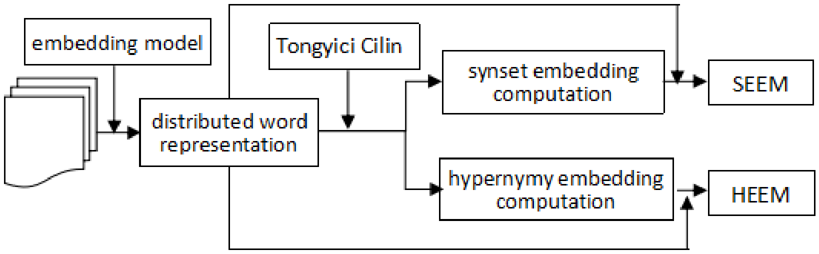
embedding model
Tongyici Cilin
synset embedding computation
SEEM
distributed word representation
hypernymy embedding computation
HEEM

Figure 1: Model structure of CEWR

4.1 Concepts and semantic hierarchy in Cilin

Cilin is a Chinese semantic thesaurus published in 1983 and revised in 2005 by enlarging the vocabulary and making it machine-readable [HIT-IR (2005)]. It is one of the most widely used semantic resources in Chinese language processing, and we use it as the semantic resource in this paper. In Cilin, synonyms are grouped in one line as a synset. Each synset expresses a distinct concept and is represented by a unique code. For example, the synset “Bo25C01=车轮 (wheel) 轮子 (wheel) 轮 (wheel) 轱辘 (wheel)” represents the concept of “wheel”. The form of the code is shown in Tab. 4.

Table 4: Cilin's code format

Position	1	2	3	4	5	6	7	8
Symbol	B	o	2	5	C	0	1	=\#\@
Property	class	category	subcategory		hypersynset	synset		
Layer	one	two	three		four	five		

There are over 70,000 words categorized in 17,809 synsets in Cilin, each of which denotes one distinct concept. Synsets are interlinked by semantic relations in the semantic hierarchy. With the semantic meaning and hierarchical information encoded, the synset code can be used to retrieve and compute the semantics of words. We combine the hierarchical semantic information in Cilin with the distributed word embedding to improve word representation.

4.2 Synset concept encoded embedding model (SEEM)

In Cilin, even though each synset denotes a distant concept, it is invisible and cannot be computed directly. SEEM is based on a straightforward idea that the conceptual semantics can be represented with the common meaning of the words in one synset. Therefore, the concept's embedding can be computed with the average embedding of synonyms in the synset. For w in synset $S_j=\{w_1, w_2, w_3..., w_n\}$, the embedding of the concept of w is defined as

$$E_{syn}(w) = \frac{\sum_{i=1}^{n} E(w_i))}{|S_J(w)|}. \tag{2}$$

Here, E(w) stands for the word embedding of w in the synset. $|S_j(w)|$ is the overall number of words in the synset. When a concept's embedding is obtained, it can be incorporated into word representation. For target word w_i, the embedding of SEEM is defined as follows:

$$E_{SEEM}(w_i) = E_{distribution}(w_i) \oplus E_{syn}(w_i). \tag{3}$$

The concept's embedding is concatenated to the distributed embedding; hence, the word representation contains the information of both the context and the conceptual meaning of the word itself.

4.3 Hypernym concept encoded embedding model (HEEM)

A word's meaning is not only defined by itself, but also by its semantic relations with other words in the semantic network, especially its hypernyms. In dictionaries, a word is usually defined by its hypernym and some modifiers and qualifiers. Hypernyms contain the core meaning of their hyponyms, regardless of similar contexts and grammatical structures, and thus they can be used as a complement to the word's semantic representation, particularly for low-frequency words. Therefore, we integrate the hypernym concept into word representation.

In Cilin, there are five semantic layers in its hierarchy denoted by its code. Usually, the first synset in each layer is the hypernym synset. For example, Synset "Bi06D01" denotes the concept of "羊" (sheep) and represents the hypernym concept of the hypersynset

“Bh03B”; meanwhile, the concept “政策” (policy) in “Di09A01” is the hypernym concept of subcategory “Di09”. Based on knowledge of Cilin, for target word *w*, the embedding of its j-th layer’s hypernym synset is defined:

$$E_{hyperlayer_j}(w) = \frac{\sum_{i=1}^{n} E(w_i)}{|Hypersyn_j(w)|}. \tag{4}$$

Here, HyperSyn$_j$(w) stands for the j-th layer hypernym synset of target word *w*. Theoretically, for each word, five layers of hypernym concepts can be obtained, including that of synset, hypersynset, subcategory, category, and class. We then take them into consideration with weights. The hypernym concept encoded embedding is defined as follows:

$$E_{concept}(w_i) = \alpha E_{syn}(w_i) + \beta \sum_{j=1}^{n} v_j E_{hyperlayer_j}(w_i); \tag{5}$$

$$\sum_{j=1}^{n} v_j = 1. \tag{6}$$

Here, α is the weight of the synset’s concept embedding, and β is that of the hypernym embedding. We use v_j to adjust the weight of each layer’s semantics that are taken into the final embedding. In this way, the conceptual semantics of the word itself and the hypernym semantics of different layers are combined to represent the word. Finally, the embedding of HEEM is defined as follows:

$$E_{HEEM}(w_i) = E_{distribution}(w_i) \oplus E_{concept}(w_i). \tag{7}$$

HEEM is encoded with both the information of the distribution in the corpus and the conceptual meaning of the target word itself and its hypernyms. As shown in the experiments, this helps improve the word representation.

5 Experiments and analysis

In this section, we evaluate the effectiveness of our CEWR models on two tasks: word similarity computation and short text classification. Before presenting our experimental results, we first describe the datasets used in our experiment and other experimental settings.

5.1 Experimental settings

We select Sogou-News[5] as our primary text corpus for distributed word embedding learning. This corpus contains 649 million words, and the vocabulary size is 1.22 million. We use the SkipGram model as an example to train the distributed word embedding on this corpus. For the parameter settings, we set the vector dimension as 300, the context window size as 5, and the number of negative samples as 5. When training our CEWR models, we use the same settings.

5.2 Word similarity

Word similarity computation is often used to evaluate the quality of word representation

[5] http://www.sogou.com/labs/resource/cs.php.

by comparing similarity scores of word pairs predicted by word representation models and human judgement.

5.2.1 Experiment results

We initially selected Wordsim-297 and Wordsim-240[6] for evaluation, both of which are commonly used in Chinese word similarity computation. The Spearman correlations between human judgement and the results of the models are shown in Tab. 5.

Table 5: Spearman correlations between human judgement and results of models

Model	Wordsim-240	Wordsim-297
SkipGram	59.71	60.90
CWE [Chen et al. (2015)]	59.64	63.58
SEWRL [Niu et al. (2017)]	61.20	64.00
WSME [Li et al. (2019)]	63.54	64.09
CEWR-SEEM	60.26	63.94
CEWR-HEEM	60.56	64.71

From the observation of the evaluation results on Wordsim-297, we can find that:

(1) Both SEEM and HEEM outperform all of the selected baseline models, especially the conventional SkipGram model, which models on word distribution. This indicates that the conceptual semantic embedding is a good complement to distributed word representation.

(2) Compared with other models like SEWRL and WSME, which also have semantic meanings encoded within, HEEM achieves the best results. This indicates that hypernymous information is an important part of the composition in word meaning, and it enriches the semantic information when it is integrated into word representation.

Meanwhile, the CEWR model does not significantly improve on Wordsim-240. The reason for this is that Wordsim-240 tends to score the contextual relevance between words rather than the similarity. Tab. 6 presents some high-ranking word pairs in the two datasets.

Similar words refer to those that share some sememes in common and can be interchanged by each other in some context; meanwhile, related words refer to those that usually appear in the same semantic frame with a high frequency of collocation. In Wordsim-240, highly related word pairs are given a high rank, such as 李白 (name of a famous Chinese poet) and 诗 (poem); this rank is even higher than that for similar word pairs like "白天 (daytime) – 晚上 (night)". Moreover, the pairing "喝水 (drinking water) – 嘴 (mouth)" is ranked higher than "春节 (spring festival)-正月 (January, the month of the spring festival)". From this perspective, Wordsim-240 is applicable for relevance computation rather than similarity computation. Regardless, both SEEM and HEEM

[6] https://github.com/thunlp/SE-WRLaster/datasets.

outperform SkipGram.

Table 6: High-ranking word pairs in Word-sim240 and Word-sim297

Wordsim-240			Wordsim-297		
Word 1	Word 2	Human ranking (1-10)	Word 1	Word2	Human ranking (1-5)
李白 name of a poet	诗 poem	9.2	入场券 entrance ticket	门票 ticket	4.59
医生 doctor	责任 responsibility	8.85	钱 money	现金 cash	4.58
白天 daytime	晚上 night	8.8	类型 type	种类 category	4.24
喝水 drink water	嘴 mouth	8.45	新闻 news	报道 report	4.10
春节 spring festival	正月 January	8.3	医生 doctor	护士 nurse	3.85

Furthermore, to verify the performance of CEWR in similarity computation, we utilize two other datasets. The RG65 dataset contains 65 pairs of nouns, and the MC30 dataset contains 30 noun pairs. Since these are English datasets, we use their Chinese versions [Chen, Li, Zhu et al. (2016)]. Moreover, since these two datasets are smaller, there are more high-similarity word pairs in them than in Wordsim-297, as is shown in Tab. 7.

Table 7: Comparison of high-similarity word pairs in three datasets

	Wordsim-297	MC30	RG65
High-similarity Pairs	30	10	24
Total Pairs	296	30	65
Ratio	0.076	0.333	0.369

Due to this reason, the performances of the investigated models on these datasets vary greatly. The Spearman correlations between the cosine similarity scores with the three models and human judgement on these datasets are shown in Tab. 8.

Table 8: Experimental results on three datasets

Model	Wordsim-297	MC30	RG65
SkipGram	60.90	75.47	79.52
CEWR-SEEM	63.94	75.27	83.25
CEWR-HEEM	64.71	81.84	85.16

Through combining the findings in Tabs. 7 and 8, we observe that:

(1) There are more high-similarity word pairs in MC30 and RG65, which means that the word pairs in these two datasets are more relatively similar than those in Wordsim-297.

Accordingly, the CEWR model performs better on them. This indicates that the CEWR can capture the similarity relations in words very well.

(2) SEEM performs better on Wordsim-297 and RG65, but does not yield much improvement on MC30. The reason for this is that the words in MC30 are high-frequency words such as “汽车 (car), 男孩 (boy), 鸟 (bird)” which results in them being well trained in the corpus and capturing sufficient semantic information. In this case, the conceptual semantic complement does not help that much. Meanwhile, in the other two datasets, there are low-frequency words such as “庇护所 (shelter)” and “坟堆 (cemetery)” in RG65 and “质子 (proton)” and “繁殖力 (fecundity)” in Wordsim-297. For these words, conceptual semantics are very important and crucial to improve their word representation.

(3) With the hypernym concept encoded, HEEM performs much better on all three datasets, especially MC30. It exhibits a statistically significant improvement of over 6% against the SkipGram model and SEEM. This indicates that hypernymous information conveys essential information of words, and this is very useful in revealing semantic relations between words. Hence, the semantic relations cannot be well captured if only learning from the word’s distribution. From this point of view, hypernymous information should be taken as an essential and irreplaceable component in word representation.

5.2.2 Case study

To demonstrate the validity of CEWR in capturing the meaning of words, we show some examples from a case study in Tab. 9.

Table 9: Examples of similarity computation with different models

Word Pairs		SkipGram	SEEM	HEEM	Human Ranking (Normalized)
Word 1	Word 2				
钱 money	财产 property	0.22	0.33	0.36	0.68
街道 street	大街 avenue	0.43	0.51	0.56	0.81
老虎 tiger	美洲虎 jaguar	0.36	0.46	0.56	0.70

As shown, all three pairs of words are very similar in meaning. However, due to various reasons, such as low frequency (美洲虎 jaguar) and differentiation in collocation (街道 street-大街 avenue), their similarity computed by SkipGram based on distribution is much lower than that made through human judgement. Meanwhile, with CEWR, the similarity increases and becomes much closer to human judgement. For example, for the pair “钱 (money)-财产(property)”, even though both words have a high frequency in the corpus, the similarity by SkipGram is only 0.22 since the words are usually used in different contexts and with different collocations. With SEEM, the similarity is increased to 0.33, and HEEM increases the similarity to 0.36, which is much closer to human judgement. Another example is “老虎 (tiger)-美洲虎 (jaguar)”. “美洲虎 (jaguar)” is a large animal of the cat family that lives in Central and South America. However, it is not a common word with high frequency in the corpus. Insufficient contexts lead to low-quality embedding, and thus we cannot precisely reveal the semantic relation with the word “tiger”. With the semantics of the hypernym concept integrated, the representation of jaguar is improved, and the similarity between this word pair is much closer to what it should be. This proves that concept semantics is a very informative complement to word representation.

5.2.3 Influence of corpus size

To explore how much the CEWR model improves word representation, we conduct our experiments on corpora of various sizes. We select the self-built Baidu corpus (0.4 G in size), Wikipedia_zh corpus (1.3 G), and Sogou-News corpus (3.7 G). The results are shown in Tab. 10.

Table 10: Influence of corpus sizes

corpus	size	tokens	model	Wordsim297	MC30	RG65
Baidu	0.4 G	68 M	SkipGram	55.81	55.69	68.96
			SEEM	60.09	59.11	75.03
			HEEM	60.39	71.61	81.47
Wiki_zh	1.3 G	223 M	SkipGram	58.09	63.15	63.87
			SEEM	60.70	66.71	71.10
			HEEM	61.57	71.69	74.51
Sogou	3.7 G	649 M	SkipGram	60.90	75.47	79.52
			SEEM	63.94	75.27	83.25
			HEEM	64.71	81.84	85.16

From evaluation of the table, we can observe that:

(1) The size of the corpus is very important for obtaining good word representation. We find a positive correlation between corpus size and model performance for both baseline models and our proposed models. Across all size settings, CEWR effectively improves word representation. This shows that conceptual semantics is very useful information for word representation.

(2) With synset concept semantics encoded, the word representation improves for corpora of all sizes. With the hypernym concept encoded, the word representation is even better. This indicates that the hypernym contains considerable semantics of words and that the meaning of words is better represented with the hypernym concept included.

(3) For a small corpus, CEWR helps in boosting the performance in a more significant way since the conceptual meaning is very necessary for distributed word embedding when training contexts are insufficient. It also brings considerable improvement on a large corpus, which shows that CEWR represents words in some aspects that distributed word embedding cannot capture.

5.3 Short text classification

Short text classification is one of the fundamental and intensively studied NLP tasks. Due to its shortness and sparsity, short text classification is more challenging than document classification. Since there are fewer words in a short text and the context cannot provide sufficient contextual information for the target word, we believe that the word embedding combined with conceptual semantics can better represent the meaning of words and give some semantic clues for classification. To verify our hypothesis, we implement experiments on short text classification to evaluate the quality of CEWR.

5.3.1 Experimental settings and dataset

Since CEWR can improve word representation, especially when training data is insufficient (as shown in the above experiments), we create a small dataset and conduct experiments on it to verify CEWR's quality. Our dataset is created by extracting 24,000 headlines from six categories in THUCNews[7], including entertainment, education, politics, society, science & technology, and economy. The average length of headlines in the new dataset is 10.2 tokens.

In the experiments, we train a convolutional neural network (CNN) composed of a single hidden layer, and the word embedding trained by SkipGram and those by CEWR are used as inputs for comparison. We use the same hyperparameters so that the evaluation score solely depends on the input embedding.

5.3.2 Evaluation

Tab. 11 reports the precision, recall, and F-score of each category with SkipGram and CEWR as embedding models.

From the Table we can see that:

(1) With additional conceptual semantics information, SEEM slightly improves the F-score of classification from 0.76 to 0.77. Furthermore, with the hypernym concept encoded, HEEM improves it to 0.79. This proves that conceptual semantics, especially hypernym information, helps improve word representation.

(2) Specifically, HEEM performs well in the classification of the categories entertainment, science & technology, and economy. The reason for this is that there are relatively more sub-topics and diversified vocabulary in these three categories, and when the training set is not large enough, it is harder to catch common characteristics in these categories for classification as there are only about ten tokens in each headline. SEEM and HEEM enrich word representation by introducing extra semantic knowledge, and aid the classification by presenting common semantics in the semantic network. The result proves that conceptual semantics are useful and helpful in word representation, particularly when the training set is small.

(3) In science & technology, there are comparatively more professional terms and new words that have very low frequency in the corpus. For those words, distributional embedding cannot represent the meaning very well, and semantic relations with other words can hardly be revealed, which results in a relatively low recall. With conceptual semantics encoded, both precision and recall are significantly improved. This shows that CEWR significantly improves word representation for low-frequency words.

[7] http://thuctc.thunlp.org/.

Table 11: Classification results with SkipGram and CEWR

Embedding model	Category	P	R	F
SkipGram	Entertainment	0.70	0.98	0.82
	Education	0.90	0.87	0.88
	Politics	0.90	0.87	0.88
	Society	0.92	0.82	0.87
	Science & technology	0.49	0.14	0.22
	Economy	0.58	0.94	0.72
	Macro Avg.	0.75	0.77	0.76
SEEM	Entertainment	0.82	0.95	0.88
	Education	0.83	0.89	0.86
	Politics	0.83	0.89	0.86
	Society	0.82	0.87	0.84
	Science & technology	0.60	0.22	0.32
	Economy	0.60	0.91	0.72
	Macro Avg.	0.75	0.79	0.77
HEEM	Entertainment	0.92	0.93	0.93
	Education	0.74	0.92	0.82
	Politics	0.74	0.92	0.82
	Society	0.86	0.86	0.86
	Science & technology	0.62	0.33	0.43
	Economy	0.68	0.92	0.78
	Macro Avg.	0.76	0.81	0.79

6 Conclusion and future work

6.1 Conclusions

In this paper, we have presented a CEWR model that incorporates conceptual semantics, particularly hypernym information learned from the Cilin knowledge base, with a distributed word embedding to improve word representation. As a post-processing model, it can be easily integrated into other word embedding models. Its performance on the word similarity task shows that using conceptual semantics can significantly improve word representation, especially for low-frequency words. The experiment on short text classification also verifies its effectiveness in improving word representation. Compared with other word representation models, CEWR is very efficient in capturing semantic relations between words and enriching semantic information for word representation, which distributed word embedding can hardly capture.

6.2 Future work

For future work, we plan to validate the effectiveness of integrating conceptual semantics into word representation in other languages. We believe that semantic knowledge is a good complement to distributed word representation, and thus we will explore more useful semantic information demonstrated in other semantic knowledge bases to improve word representation.

Acknowledgment: The authors would like to thank all anonymous reviewers for their suggestions and feedback.

Funding Statement: This research is supported by the National Science Foundation of China (grant 61772278, author: Qu, W.; grant number: 61472191, author: Zhou, J. http://www.nsfc.gov.cn/), the National Social Science Foundation of China (grant number: 18BYY127, author: Li B. http://www.cssn.cn), the Philosophy and Social Science Foundation of Jiangsu Higher Institution (grant number: 2019SJA0220, author: Wei, T. https://jyt.jiangsu.gov.cn) and Jiangsu Higher Institutions' Excellent Innovative Team for Philosophy and Social Science (grant number: 2017STD006, author: Gu, W. https://jyt.jiangsu.gov.cn)

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Bartusiak, R.; Augustyniak, Ł.; Kajdanowicz, T.; Kazienko, P.; Piasecki, M. (2019): WordNet2Vec: Corpora agnostic word vectorization method. *Neurocomputing*, vol. 326, pp. 141-150.

Bengio, Y.; Rejean, D.; Pascal, V. (2003): A neural probabilistic language model. *Journal of Machine Learning Research*, no. 3, pp. 1137-1155.

Chen, H.; Li, F.; Zhu, X.; Ma, R. (2016): A path and depth-based approach to word semantic similarity calculation in Cilin. *Journal of Chinese Information Processing*, vol. 30, no. 5, pp. 80-88.

Chen, X.; Liu, Z.; Sun, M. (2014): A unified model for word sense representation and disambiguation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1025-1035.

Chen, X.; Xu, L.; Liu, Z.; Sun, M.; Luan, H. (2015): Joint learning of character and word embeddings. *Proceedings of Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1236-1242.

Chinese Academy of Social Sciences. (2016): *Contemporary Chinese Dictionary* (7th version). The Commercial Press, Beijing.

Dai, Z.; Huang, R. (2018): Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 141-151.

Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. (2019): BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the North American Chapter of the Association for computational Linguistics*, pp. 4171-4186.

Devlin, J.; Zbib, R.; Huang, Z.; Lamar, T.; Schwartz, R. M. et al. (2014): Fast and robust neural network joint models for statistical machine translation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 1370-1380.

Harris, Z. (1954): Distributional structure. *Word*, vol. 23, no. 10, pp. 146-162.

HIT-SCIR. (2005): Tongyici Cilin (Extended Version). https://www.ltp-cloud.com/download /#down_cilin.

Li, G.; Lv, L.; Wang, R.; Li, J.; Li, R. (2016): Semantic role labeling based on Tongyici Cilin derived features. *Journal of Chinese Information Processing*, vol. 30, no. 1, pp. 101-107.

Li, X.; You, S.; Chen, W. (2019): An algorithm of similarity between words based on word single-meaning embedding model. *Acta Automatica Sinica*, pre-print, doi: 10.16383/j.aas.c180312.

Liu, D.; Peng, C.; Qian, L.; Zhou, G. (2014): The effect of Tongyici Cilin in Chinese entity relation extraction. *Journal of Chinese Information Processing*, vol. 28, no. 2, pp. 91-99.

Liu, J.; Yang, Y. H.; Lv, S. Q.; Wang, J.; Chen, H. (2019): Attention-based BiGRU-CNN for Chinese question classification. *Journal of Ambient Intelligence and Humanized Computing*, doi.org/10.1007/s12652-019-01344-9.

Mei, J.; Zhu, Y.; Gao, Y.; Yin, H. (1983): *Tongyici Cilin*. Shanghai Lexicographical Publishing House, Shanghai.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J. (2013): Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, pp. 3111-3119.

Niu, Y.; Xie, R.; Liu, Z.; Sun, M. (2017): Improved word representation learning with sememes. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 2049-2058.

Peng, Q.; Zhu, X.; Chen, Y.; Sun, L.; Li, F. (2018): IC-based approach for calculating word semantic similarity in Cilin. *Application Research of Computers*, vol. 35, no. 2, pp. 400-404.

Pennington, J.; Socher, R.; Manning, C. (2014): GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532-1543.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C. et al. (2018): Deep contextualized word representation. *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 2227-2237.

Polat, N. C.; Yaylali, G.; Tanay, B. (2019): A method for decision making problems by using graph representation of soft set relations. *Intelligent Automation and Soft Computing*, vol. 25, no. 2, pp. 305-311.

Qiu, J.; Liu, Y.; Chai, Y.; Si, Y.; Su, S. et al. (2019): Dependency-based local attention approach to neural machine translation. *Computers, Materials & Continua*, vol. 59, no. 2, pp. 547-562.

Rothe, S.; Schütze, H. (2015): AutoExtend: extending word embeddings to embeddings for synsets and lexemes. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol. 1, pp. 1793-1803.

Socher, R.; Lin, C.; Manning, C.; Ng, A. Y. (2011): Parsing natural scenes and natural language with recursive neural networks. *Proceedings of the 28th International Conference on Machine Learning*, pp. 129-136.

Wei, T.; Chen, X.; Shi, S.; Zhou, J.; Gu, Y. et al. (2018): Optimizing the taxonomy and hierarchy of a Chinese lexical database-Cilin. *Proceedings of 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing*, pp. 99-102.

Yang, L.; Sun, M. (2015): Improved learning of Chinese word embeddings with semantic knowledge. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, Springer International Publishing, pp. 15-25.

Yu, M.; Dredze, M. (2014): Improving lexical embeddings with semantic knowledge. *Proceedings of the Association for Computational Linguistics*, pp. 545-550.