Improved Learning of Chinese Word Embeddings with Semantic Knowledge

Liner Yang[1] and Maosong Sun[1,2]
[1] Department of Computer Science and Technology,
State Key Lab on Intelligent Technology and Systems,
National Lab for Information Science and Technology,
Tsinghua University, Beijing 100084, China
[2] Jiangsu Collaborative Innovation Center for Language Ability,
Jiangsu Normal University, Xuzhou 221009, China
lineryang@gmail.com, sms@tsinghua.edu.cn

Abstract. While previous studies show that modeling the minimum meaning-bearing units (characters or morphemes) benefits learning vector representations of words, they ignore the semantic dependencies across these units when deriving word vectors. In this work, we propose to improve the learning of Chinese word embeddings by exploiting semantic knowledge. The basic idea is to take the semantic knowledge about words and their component characters into account when designing composition functions. Experiments show that our approach outperforms two strong baselines on word similarity, word analogy, and document classification tasks.

Keywords: Word embeddings, CBOW, semantic knowledge

1 Introduction

Distributed word representations, also known as word embeddings, have proven to be effective in capturing both semantic and syntactic regularities in language [1,14,13,16,6]. These word embeddings have benefited a range of natural language processing (NLP) tasks, including named-entity recognition [5], word sense disambiguation [3], syntactic parsing [17] and sentiment analysis [18].

While early approaches treat words as the basic unit for learning distributed representations from unlabeled data (e.g., [13,19]), a number of researchers have demonstrated the usefulness of exploiting the internal structure of words and modeling the minimum meaning-bearing units, such as morphemes in English or characters in Chinese [10,2,4]. Luong et al. [10] propose a recursive neural network (RNN) model to encode morphological structure of words. Botha and Blunsom [2] introduce a log-bilinear model which uses addition as composition function to derive word vectors from morpheme vectors. Chen et al. [4] extend their idea and present a character-enhanced word embedding (CWE) model. These morpheme- and character-based models significantly outperform the original word-based models in a variety of tasks.

We believe there is still a room to improve word embeddings by considering the intricate dependencies between the minimum meaning-bearing units rather than simply taking addition as composition function when deriving word vectors. For example, the ways how characters interact to determine the meaning of a word are significantly different between two words “远眺 (overlook)” and “村落 (villages)”. Instead of simply adding the vectors of two characters, our intuition is that the semantic relations between characters should be modeled to better learn distributed representations of Chinese words. In this work, we propose to exploit semantic knowledge to improve the learning of distributed representations of Chinese words. Based on semantic categories and relations derived from Tongyi Cilin, a Chinese semantic thesaurus, we design new composition functions to compute word vectors from character vectors. Experiments on word similarity, word analogy, and documentation classification tasks show that our approach significantly outperforms the state-of-the-art baseline methods.

2 Background

2.1 The CBOW Model

The continuous bag-of-words (CBOW) model [12] is a recently proposed framework for learning continuous word representations based on the distributional hypothesis. In the model, each word $w \in W$ is associated with vector $v_w \in \mathbb{R}^d$, where W is the word vocabulary and d is the vector dimension. The entries in the vectors are treated as parameters to be learned. Specifically, we learn these parameters values so as to maximize the log likelihood of each token given its context:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \log p(w_i | w_{i-k}^{i+k}) \tag{1}$$

where N is the size of corpus and w_{i-k}^{i+k} is the set of words in the window of size k centered at w_i (w_i excluded). The CBOW model formulates the probability $p(w_i|w_{i-k}^{i+k})$ using a softmax function as follows:

$$p(w_i|w_{i-k}^{i+k}) = \frac{\exp\left(v'_{w_i} \cdot \sum_{-k \le j \le k, j \ne 0} v_{w_{i+j}}\right)}{\sum_{w \in W} \exp\left(v'_w \cdot \sum_{-k \le j \le k, j \ne 0} v_{w_{i+j}}\right)} \tag{2}$$

where v_w and v'_w represent the input and output vectors of the word w respectively. In order to learn model efficiently, the techniques of hierarchical softmax and negative sampling are used [13]. One key limitation of the CBOW model is that it treats each word as the basic unit and fails to capture the internal structure of words. Therefore, some morphological-based methods have been proposed, for example Chen et al. [4].

2.2 The CWE model

To the best of our knowledge, the closest work to ours for learning Chinese word embeddings is character-enhanced word embedding (CWE) model [4], which learns character

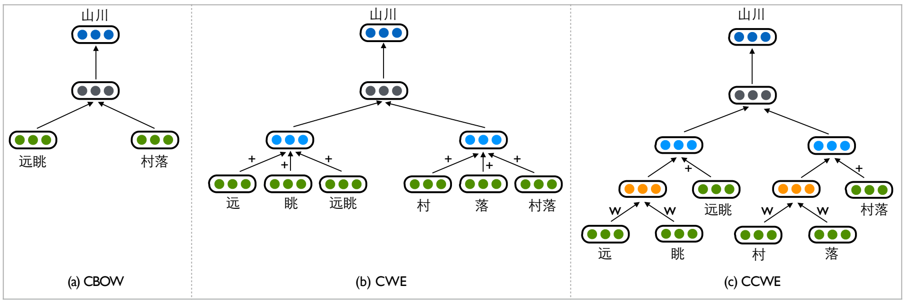

Fig. 1. The architectures of (a) continuous bag-of-words (CBOW) model, (b) character-enhanced word embeddings (CWE) model and (c) compositional Chinese word embeddings (CCWE) model. Here "远眺 (overlook) 山川 (mountains) 村落 (villages)" is a word sequence. The word "远眺 (overlook)" is composed of characters "远 (far)" and "眺 (view)", and the word "村落 (villages)" is composed of characters "村 (village)" and "落 (place)".

and word embeddings jointly. The key idea of the model is to represent the word with its surface form itself and its component characters as follows:

$$v_{w_i} = v^f_{w_i} + \frac{1}{n_i}\sum_{j=1}^{n_i} v^f_{c^i_j} \tag{3}$$

where $v^f_{w_i}$ is the surface form word vectors, $v^f_{c^i_j}$ is the character vector, n_i is the number of characters in word w_i and c^i_j is the j-th character in word w_i. They use the addition operation for simplicity. This is a principled way of handling new words, we can get the vector of a new word by adding vectors of its component characters. Our work does not simply add vectors of characters, but rather combines them using more linguistically-motivated composition functions.

3 Our Models

The way how characters are composed to form the meaning of a word is far more intricate than addition. Thus, we propose to learn different compositional functions for different semantic relations of words and their characters. In this paper, we use a semantic formation corpus to identify the semantic relation of words and propose two novel models for learning compositional functions as well as word representations.

In this section, we first describe the semantic formation corpus as they serve as the basis of our model. We then introduce a compositional Chinese word embeddings (CCWE) using semantic category and semantic relations to learn word embeddings and compositional functions. Figure 1(c) shows the overview of our proposed model. Finally, we provide complexity analysis about our model and some baseline models.

3.1 Semantic Formation Lexicon

The Tongyici Cilin [11], a Chinese thesaurus, is adopted in this paper that contains 12 main categories labeled “A-L”, 96 middle categories labeled with lower case letters and 1,506 subcategories labeled with numbers. Each small category consists of a group of synonyms that have the same or similar meaning. For example, under the major category “B”, the middle category “Bh” groups all words that refer to “plant”. Under the middle category “Bh”, the subcategory “Bh02” groups all words that refer to flower, e.g., “兰花 (orchid)”.

Compound		First character		Second character	
村落(villages),	Cb25	村(village),	Cb25	落(place),	Cb08
远眺(overlook),	Fc04	远(far),	Eb21	眺(view),	Fc04
木瓜(pawpaw),	Bh07	木(wood),	Bm03	瓜(cucurbitaceae),	Bh07
同窗(classmate),	Aj04	同(together),	Ka23	窗(window),	Bn04

Table 1. Examples of annotated disyllabic compounds.

Chinese characters are usually meaningful in words. Therefore we annotate 52,362 disyllabic compounds with semantic information, in which compound and component characters are appended with semantic categories, as shown in Table 1. In Table 1, “Cb25” is the semantic category of compound “村落 (villages)” which is composed of character “村 (village)” with semantic category “Cb25” and character “落 (place)” with semantic category “Cb08”. Although Chinese characters are highly ambiguous, i.e. having more than one semantic category, the semantic category of the character in a word is determined.

3.2 The CCWE Model

As illustrated in Figure 1(c), the word vectors are derived from their component character vectors. The word vector v_{w_i} is constructed by character vector $v^f_{c^i_j}$ and the surface form word vector $v^f_{w_i}$ as follows:

$$v_{w_i} = v^f_{w_i} + \frac{1}{n_i}\sum_{j=1}^{n_i} h^t_j \odot v^f_{c^i_j} \tag{4}$$

where $h^t_j \in \mathbb{R}^{d\times 1}$ are tag-specific weight vectors [1] and $\odot$ denotes element-wise multiplication. This forms the basis of our CCWE model with $\theta = \{h^t_j, v^f_{w_i}, v^f_{c^i_j}\}$ being parameters to be learned.

It is very important to define tag c which we will use to incorporate semantic knowledge. In this paper, we propose two different tags, i.e. categorical tags and relational tags.

[1] We use tag-specific weight vectors rather than weight matrices, as the vLBL model [14] does, for significantly faster training. This has been discussed by Mnih and Teh [15].

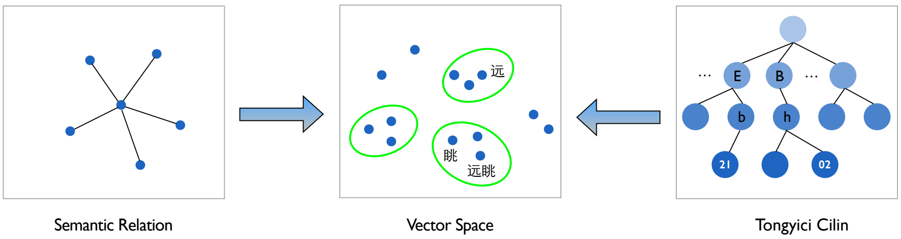

Fig. 2. Overview of our models which leverage semantic category and semantic relation information to improve the quality of word representations.

Category-based Model According to the right part of Figure 2, semantic category knowledge encodes the semantic properties of words, from which we can group similar words according to their attributes. Then we may require the representations of words that belong to the same semantic category to be close to each other. Therefore, we give the definition of tag $t \triangleq (s, p)$, where $s \in S$, S is the semantic category sets and $p \in \{B, E\}$, B, E corresponding to the position of character in a word, i.e. Begin and End.

Therefore, we replace tag t in Eq. (4) with (s, p). This gives the new representation of the word w:

$$e_{w_i} = e^f_{w_i} + \frac{1}{n_i} \sum_{j=1}^{n_i} h_j^{(s,p)} \odot e^f_{c^i_j} \tag{5}$$

We call this model category-based model and denote this category-based model as C-CCWE for ease of reference.

Relation-based Model We first give the definition of semantic relation knowledge.

Definition 1 (Semantic Relation). *Semantic relation r indicates which character's meanings is closer to word meanings, where $r \in R$, R is the set of semantic relationships.*

Since we only take disyllabic compounds into consideration, we divide r into three main types:

1. Beginning character biased, which means that the meaning of first character is closer to the meaning of word.
2. Ending character biased, which means that the meaning of last character is closer to the meaning of word.
3. Unbiased, which means that either of two characters semantic distance to the word is approximate or the word is non-compositional.

For example, the meaning of “远眺 (overlook)” should be closer to character “眺 (view)” than “远 (far)”, we label the (r, p) of “远眺 (overlook)” as “ending character biased”, where p is also position of character in a word.

Similarly, we replace tag t in Eq. (4) with (r, p):

$$e_{w_i} = e^f_{w_i} + \frac{1}{n_i} \sum_{j=1}^{n_i} h_j^{(r,p)} \odot e^f_{c^i_j} \tag{6}$$

We call this model relation-based model and denote this relation-based model as R-CCWE for ease of reference.

3.3 Optimization

In this paper, the proposed compositional Chinese word embeddings (CCWE) are learned using stochastic gradient descent (SGD) algorithm.

3.4 Complexity Analysis

We now analyze model complexities of the CBOW, CWE, C-CCWE and R-CCWE models.

Table 2 shows the complexity of model parameters of various models. In the table, the dimension of vector is d, the word vocabulary size is $|W|$, the character vocabulary size is $|C|$, the semantic category set size is $|S|$ and the semantic relation size is $|R|$. The CBOW window size is $2k$, the corpus size is N, the average number of characters of each word is n, and the computational complexity of negative sampling and hierarchical softmax for each target word is f.

From the complexity analysis, we can observe that, compared with CWE, the computational complexity of CCWE does not increase and the CCWE only requires a little more parameters for saving weight vectors (note that $|S| \ll |W|$). In our experiment, we set $|S| = 1000$ and $|R| = 3$.

Model	Model Parameters	Computational Complexity
CBOW	$\|W\|d$	$2kNf$
CWE	$(\|W\| + \|C\|)d$	$2kN(f + n)$
C-CCWE	$(\|W\| + \|C\| + \|S\|)d$	$2kN(f + n)$
R-CCWE	$(\|W\| + \|C\| + \|R\|)d$	$2kN(f + n)$

Table 2. Model complexities.

4 Experiments

In this section, we first describe our experimental settings, including the datasets and baseline methods. Then we compare our models with baseline methods on three tasks, *i.e.*, word similarity, word analogy, and document classification.

4.1 Experimental Settings

We use a text corpus with news articles from *The People's Daily* for learning word embeddings, which is also used by Chen et al. [4]. The corpus in total has about 31 million words. The word vocabulary size is 105 thousand and the character vocabulary size is 6 thousand.

Following the parameter settings in Chen et al. [4], the context window size is 5 and the dimension of word vector is 200. For training model we use hierarchical softmax and also adopt the same linear learning rate strategy described in [13], where the initial learning rate is 0.05.

4.2 Word Similarity

In this task, each model is required to compute semantic relatedness of given word pairs. The correlations between results of models and human judgements are reported as the model performance.

In this paper, we evaluate the word vectors with semantic similarity dataset provided by organizers of SemEval-2012 Task 4 [7]. This dataset contains 296 Chinese word pairs with similarity scores estimated by humans and the words in 60 word pairs have appeared less than 100 times. We compute the Spearman correlation between relatedness scores from a model and the human judgements for comparison. The relatedness score of two words are computed via cosine similarity of word vectors.

Method	wordsim-296	
	60 pairs	296 pairs
CBOW	55.24	60.89
CWE	60.07	62.13
C-CCWE	62.30	63.46
R-CCWE	**63.03**	**65.17**

Table 3. Evaluation results on wordsim-296 ($\rho \times 100$).

The evaluation results of CCWE and baseline methods on wordsim-296 are shown in Table 3. From the evaluation results, we observe that: CCWE and its extensions all significantly outperform baseline methods on both 60 word pairs and 296 word pairs.

4.3 Word Analogy

The word analogy task is introduced by [12] to quantitatively evaluate the linguistic regularities between pairs of word representations. The task consists of question like "男人 (man) is to 女人 (woman) as 父亲 (father) to ___", where ___ is missing and must be predicted from the entire vocabulary. To answer such question, we need to find a word x such that its vector x is close to vec(女人) - vec(男人) + vec(父亲) according to the cosine similarity. The question is judged as correctly answered only if x is exactly the answer in the evaluation set. The evaluation metric for this task is the percentage of questions answered correctly.

We use Chinese analogy dataset from [4]. The dataset contains 1,124 analogies and 3 analogy types: (1) capitals of countries (687 groups); (2) states/provinces of cities (175 groups); and (3) family words (240 groups).

Table 4 shows the results of word analogy. The R-CCWE method outperforms C-CCWE methods and performs significantly better than all baseline methods.

Model	Total	Capital	State	Family
CBOW	45.15	36.34	55.43	62.50
CWE	56.04	52.58	69.71	55.83
C-CCWE	58.88	53.47	77.14	60.83
R-CCWE	61.63	55.24	69.71	73.75

Table 4. Evaluation accuracies (%) on word analogy.

4.4 Document Classification

Another way to evaluating the quality of the word embeddings is using the word vectors to compute document representation, which can be evaluated with document classification tasks. To obtain document vectors, we choose a very simple approach that takes the average of the word vector representations in that document. This is because we aim to compare the word embeddings with different approaches instead of finding the best method for document embeddings.

In this paper, we run experiments on the dataset Chinese Encyclopedia, which is from the electronic version of the Chinese Encyclopedia. This dataset was also used by Li and Sun [8]. This dataset contains 55 categories and about 70,000 documents and is split into training set and test set with 9:1. Each document belongs to only one category. All document vectors are used to train classifier using the LibLinear package [2]. We report the classification metrics Micro-F1 and Macro-F1. The results are averaged over 10 different runs.

Method	Micro-F1	Macro-F1
CBOW	75.93	74.23
CWE	77.01	75.76
C-CCWE	77.65^{*}	76.00^{*}
R-CCWE	77.97^{*+}	76.45^{*+}

Table 5. Results on document classification. "*": significantly better than CBOW ($p < 0.05$). "+": significantly better than CWE ($p < 0.05$).

[2] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Table 5 shows the results of document classification. Similar conclusion can be made as in the word analogy task. The R-CCWE method outperforms C-CCWE methods and performs significantly better than all baseline methods.

4.5 Qualitative analysis

In order to demonstrate the characteristics of CCWE model, we select two example words and use R-CCWE model to find the most similar words of these words. For comparison, we also used CBOW model and CWE model to find similar words of these example words. In Table 6, we can observe that, the most similar words returned by the CBOW model are syntactically related words. The most similar words returned by the CWE model tend to share at least one character with the given word, for example: “他途 (other way)”, “牛蒡 (great burdock, which is a species of plants)”. The most similar words found by the R-CCWE model are a mixture of syntactically and semantically related words.

Words	CBOW	CWE	CCWE
仕途 (official career)	失意 (frustrated)	仕宦 (be an official)	仕宦 (be an official)
	功名 (scholarly honour)	仕 (be an official)	功名 (scholarly honour)
	官场 (official circle)	宦途 (official career)	中举 (pass imperial exams)
	穷愁 (depressed)	失意 (frustrated)	宦途 (official career)
	闻达 (illustrious)	他途 (other way)	书生 (intellectual)
种牛 (stud bull)	蛋鸡 (laying hen)	种羊 (stud sheep)	肉牛 (beef cattle)
	肉牛 (beef cattle)	种畜 (breeding stock)	奶牛 (milking cow)
	出栏 (of livestock)	肉牛 (beef cattle)	黄牛 (ox)
	存栏 (livestock)	种羊场 (stud sheep farm)	养牛 (cowboying)
	鸡场 (chicken farm)	牛蒡 (great burdock)	种羊 (stud sheep)

Table 6. Target words and their most similar words under different word representations.

5 Related Work

This work is inspired by two lines of research: (1) compositional semantic models and (2) exploiting word internal structure.

Compositional semantic models More recently, a number of authors have paid some efforts to learn compositional semantic models. Luong et al. [10] proposed a neural language model to learn morphologically-aware word representations by combining recursive neural network and neural language model. Botha and Blunsom [2] introduced the additive log-bilinear model (LBL++) which learns separated vectors for each component morpheme of a word and derves word vector from these vectors.

Finally, most similar to our model, Chen et al., [4] presented a general framework to integrate the character knowledge and context knowledge to learn word embeddings and also provides an efficient solution to character ambiguity. We solve this issue through annotating the sense of each component character in words.

Exploiting word internal structure Exploiting word internal structure to improve Chinese word segmentation and parsing has gained increasing popularity recently. Zhao [21] investigate character-level dependencies for Chinese word segmentation task in a dependency parsing framework. Their results show that annotated word dependencies can be useful for Chinese word segmentation. Li [9] annotate morphological-level word structures and proposed a unified generative model to parse the Chinese morphological and phrase structures. Zhang et al., [20] annotate character-level word structures which cover entire words in CTB and present a unified framework for segmentation, POS tagging and phrase structure parsing. Compared to their work, we annotate internal word structures from semantic view and use the knowledge to improve word embeddings. To the best of our knowledge, it is the first work in this direction.

6 Conclusion

We have presented a compositional neural language models that incorporates semantic category and semantic relation knowledge in resources to improve word embeddings. Compared to existing word representation models, CCWE is very efficient and can capture semantic relation between words and their component characters, which are crucial for semantic similarity tasks. We have demonstrated improvements on word similarity, word analogy, and document classification tasks. In summary, our contributions include:

1. We annotated the internal semantic structures of Chinese words, which are potentially useful to character-based studies of Chinese NLP.
2. We proposed a novel compositional Chinese word embeddings and investigated the effectiveness of our model in three tasks.

For future work, we plan to extend our models to learn word embeddings and semantic category of words jointly.

Acknowledgments

The authors thank Yang Liu, Xinxiong Chen, Lei Xu, Yu Zhao and Zhiyuan Liu for helpful discussions and three anonymous reviewers for the valuable comments. This research is supported by the Key Project of National Social Science Foundation of China under Grant No. 13&ZD190 and the Project of National Natural Science Foundation of China under Grant No. 61170196.

References

1. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A Neural Probabilistic Language Model. JMLR 3, 1137–1155 (2003)
2. Botha, J.A., Blunsom, P.: Compositional Morphology for Word Representations and Language Modelling. In: Proceedings of ICML (2014)
3. Chen, X., Liu, Z., Sun, M.: A unified model for word sense representation and disambiguation. In: Proceedings of EMNLP (2014)
4. Chen, X., Xu, L., Liu, Z., Sun, M., Luan, H.: Joint learning of character and word embeddings. In: Proceedings of IJCAI (2015)
5. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural Language Processing (Almost) from Scratch. JMLR 12 (2011)
6. Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A.: Retrofitting Word Vectors to Semantic Lexicons. In: Proceedings of NAACL (2015)
7. Jin, P., Wu, Y.: Semeval-2012 task 4: Evaluating chinese word similarity. In: Proceedings of SemEval (2012)
8. Li, J., Sun, M.: Scalable term selection for text categorization. In: Proceedings of EMNLP (2007)
9. Li, Z.: Parsing the internal structure of words: A new paradigm for chinese word segmentation. In: Proceedings of ACL (2011)
10. Luong, T., Socher, R., Manning, C.D.: Better Word Representations with Recursive Neural Networks for Morphology. In: Proceedings of CoNLL (2013)
11. Mei, J., Zhu, Y., Gao, Y., Yin, H.: TongYiCi CiLin. Shanghai Cishu Publisher, Shanghai, China (1983)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR (2013)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS (2013)
14. Mnih, A., Kavukcuoglu, K.: Learning word embeddings efficiently with noise-contrastive estimation. In: Proceedings of NIPS (2013)
15. Mnih, A., Teh, Y.W.: A fast and simple algorithm for training neural probabilistic language models. In: Proceedings of ICML (2012)
16. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of EMNLP (2014)
17. Socher, R., Lin, C.C., Ng, A.Y., Manning, C.D.: Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In: Proceedings of ICML (2011)
18. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of EMNLP (2013)
19. Yu, M., Dredze, M.: Improving Lexical Embeddings with Semantic Knowledge. In: Proceedings of ACL (2014)
20. Zhang, M., Zhang, Y., Che, W., Liu, T.: Chinese parsing exploiting characters. In: Proceedings of ACL (2013)
21. Zhao, H.: Character-level dependencies in chinese: Usefulness and learning. In: Proceedings of EACL (2009)