Radical Enhanced Chinese Word Embedding

Zheng Chen and Keqi Hu

University of Electronic Science and Technology of China,
Section 2-4, Jianshe north road. 610054 Chengdu, China
zchen@uestc.edu.cn

Abstract. The conventional Chinese word embedding model is similar to the English word embedding model in modeling text, simply uses the Chinese word or character as the minimum processing unit of the text, without using the semantic information about Chinese characters and the radicals in Chinese words. To this end, we proposed a radical enhanced Chinese word embedding in this paper. The model uses conversion and radical escaping mechanisms to extract the intrinsic information in Chinese corpus. Through the improved parallel dual-channel network model on a CBOW-like model, the word information context is used together with the Chinese character radical information context to predict the target word. Therefore, the word vector generated by the model can fully reflect the semantic information contained in the radicals. Compared with other similar models by word analogy and similarity experiments, the results showed that our model has effectively improved the accuracy of word vector expression and the direct relevance of similar words.

Keywords: Word Embedding, Radical Enhanced Chinese Word Embedding.

1 Introduction

Vectorized representation of text is one of the core research areas of natural language processing. One-hot embedding, as the traditional solution, has the advantage of simplifying and efficiency. However, it ignores too much semantic information, failing to meet people's expectations in practical use. Distributed word representation represents a word as a vector in a continuous vector space, makes similar words close to each other. It allows the machine learning model to directly obtain relevant semantic information through text vectors and is conducive to follow-up works. Take the advantage of distributed word representation, also known as word embedding, scholars achieve many excellent results in different natural language processing tasks, such as named entity recognition [1], text classification [2], semantic analysis [3], and question answering system [4]. Among many word-embedding methods [5][6], Continuous Bag of Words (CBOW) model and Skip-Gram model are most popular ones. They could gain excellent word embedding from large-scale corpus [7].

Although the word embedding method has excellent performance in English, this method that uses the word as the smallest unit of the language does not have a good

effect on all languages, especially on Chinese, which is a structurally complex pictographic language. The Chinese characters have semantics itself. In many cases, a single Chinese character can be a word on its own. And, furthermore, even the sub-word items of Chinese characters, including radicals and components, also contain rich semantic information. For example, the word "智能(intellect)", on the one hand, its semantic information we can learn from the relevant context in the corpus, and on the other hand we can also infer from the individual Chinese characters "智(wisdom)" and "能(ability)". And the Chinese characters "江(large river)", "河(river)", "湖(lake)" and "海(sea)" can be inferred that they are all related to water according to their common radical "氵".

However, as one of the oldest writing system in the human history, Chinese is complicated. At the very beginning, Chinese only have single characters. Subsequently, because of the need of complex expressions, people create a lot of compound characters, which compounded by two, three, or even more characters. Therefore, each part of compound characters need to be simplified, from character to radical. For example, the radical "氵" is actually the reduced form of character "水(water)". If these correspondences cannot be used, the model is difficult to find the semantic relationship of the Chinese character that has common sidelines.

In this paper, we propose a model to jointly learn the embeddings of Chinese words, characters, and sub-character components. By using simplified transformation and radical escaping techniques, the learned Chinese word embeddings can leverage the external context co-occurrence information and incorporate rich internal sub-word semantic information. Both the word similarity experiment and the word analogy experiment prove that this method is real and effective and has better effect than other similar methods.

2 Model

The Radical Enhanced Chinese Word Embedding (RECWE) proposed in this paper is based on the CBOW model. It can effectively synthesizes the Chinese characters and the radicals. The overall structure is shown in figure 1.

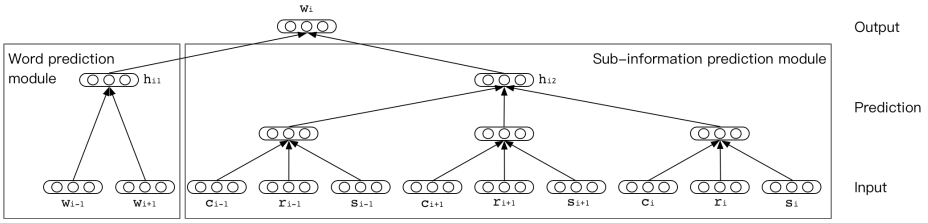

Fig. 1. Radical Enhanced Chinese Word Embedding.

RECWE has dual-prediction modules: word prediction module and sub-information prediction module. The first model is roughly the same as the CBOW model, where w_i

is the target word, w_{i-1} and w_{i+1} is the left and right words of the target word in the context, and the h_i^1 is word context vector.

The sub-information prediction module is juxtaposed with the word prediction module. Let c_{i-1}, r_{i-1}, s_{i-1} and c_{i+1}, r_{i+1}, s_{i+1}, corresponding to the context words w_{i-1} and w_{i+1} in the word prediction module, be the "input" Chinese character, radical, and component.h_i^2 is the sub-information context vector. However, because the semantic information contained in Chinese characters and radicals is not as rich as words, using the context vector directly to predict the target words will have a large error. To increase the semantic information, the sub-information prediction module uses the information of target word (c_i, r_i and s_i in the figure 1).

To prediction target words, sub-information prediction module mainly relies on the semantic information contained in the Chinese characters, but not all the words can obtain semantic information through the Chinese characters and the radicals. For example, word "东西(thing)" that represents the physical object, the internal Chinese characters "东(east)" and "西(west)" are only indicative of the position alone and are far from the semantics of the word "东西". In addition, for most of the transliterated words that appear in modern Chinese, such as "苏打(soda)" and "沙发(sofa)", this approach is not suitable. So, for such words, the sub-information prediction module will directly use their word vector to construct h_i^2, instead of disassembling into the Chinese character and the radical.

To make full use of the semantic information of Chinese characters and radicals, before training RECWE, the input text need to perform simplified conversion and radical convert(As shown in table 1).

Table 1. Radical convert table

Radical	Converted character	Radical	Converted character
艹	艸 (grass)	亻	人 (people)
刂	刀 (knife)	犭	犬 (dog)
灬	火 (fire)	釒	金 (gold)
麥	麥 (wheat)	飠	食 (eat)
礻	示 (show)	月	肉 (meat)
攵	攴 (knock)	罒	网 (net)
扌	手 (hand)	氵	水 (water)
糸	糸 (silk)	耂	老 (old)
牜	牛 (cow)	忄	心 (heart)
衤	衣 (cloth)	王	玉 (jade)
辶	走 (walk)	疒	病 (illness)

Similar to the CBOW model, the objective function of the RECWE model is the log-likelihood function of the conditional probabilities of the two context vectors for the target word w_i, as shown in equation 1.

$$L(w_i) = \sum_{k}^{2} \log P(w_i \mid h_i^k) \tag{1}$$

Where h_i^1 and h_i^2 are the word context vector and the child information context vectors.

The conditional probability $P\left(w_i \middle| h_i^k\right)$ for each context vector for the target word w_i can be calculated using the SoftMax function, as shown in Eq. 2.

$$P\left(w_i \mid h_i^k\right) = \frac{\exp\left(h_i^{kT} \hat{v}_w^i\right)}{\sum_{j=1}^{N} \exp\left(h_i^{kT} \hat{v}_w^j\right)}, k = 1, 2 \tag{2}$$

Where $\hat{v}_w^i$ is the "output" vector of the target word w_i, $\hat{v}_w^j$ is the "output" vector of each word in the input corpora, and N is the length of the input corpora.

The context vector h_i^1 is the average of the "input" vector of each word in the context, obtained by equation 3:

$$h_i^1 = \frac{1}{2T} \sum_{-T \le j \le T, j \ne 0} v_w^{i+j} \tag{3}$$

Where T is the size of the context window and v_w^{i+j} is the "input" vector of words in the context window.

Similarly, the sub-information context vector h_i^2 is the average of the "input" vectors of the radical and components. The calculation formula is shown in equation 4.

$$h_i^2 = \frac{1}{X} \sum_{-T \le j \le T, j \ne 0} v_c^{i+j} + v_r^{i+j} + v_s^{i+j} \tag{4}$$

Where v_c^{i+j}, v_r^{i+j} and v_s^{i+j} are the word, radical, and components vector, and X is the number of v_c^{i+j}, v_r^{i+j} and v_s^{i+j}.

Thus, for corpus D, the overall log-likelihood function of the RECWE model is shown in equation 5.

$$L(D) = \sum_{w_i \in D} L\left(w_i\right) \tag{5}$$

3 Experiment

To verify the feasibility and effectiveness of the word embedding trained by our model, in this section, we conduct experiments based on headline data from the news network named "今日头条". We compare our method with word2vec [7], CWE[11], SCWE[8], and JWE[13].

3.1 Parameter settings

The training corpus is processed through the Ansj segment tool[1], then filter the stop words in the news through HIT and Baidu's stop word vocabulary. All word embedding models use same training parameters, as shown in table 2. The detail of our implementation can be found at https://github.com/UESTC1010/RECWE.

Table 2. Training parameters

Parameter	Explanation	Value
size	Word embedding dimension.	200
alpha	Learning rate.	0.025
mincount	The lower bound of low frequency words in corpus.	5
sample	Threshold for down sampling of high frequency words.	1e-4
workers	Number of threads.	4
iter	Number of iterations.	5
window	Windows size.	5

To verify the influence of the sub-information of the target word on the word embedding, this experiment adopts three different sub-information superposition modes: The pattern 1 uses the sub-information corresponding to the word context, labeled "p1"; Pattern 2 uses only the sub-information of the target word, labeled "p2"; The pattern three uses the sub-information corresponding to the word context and the target word at the same time, labeled "p3".

3.2 Word similarity

This task is mainly used to evaluate the ability of word embedding to determine semantically similar word pairs [13]. In this task, two different similar word databases, wordsim-240 and wrodsim-296, which were provided in [11], were selected. They contain 240 Chinese word pairs and 296 Chinese word pairs. Each word pair contains a manually labeled similarity. Wordsim-240 is mainly for semantically related words, while wordsim-296 is for synonyms.

For each word embedding model, the similarity of a word pair is expressed using the cosine distance of the word embedding corresponding to the two words. At the end of the task, we compute the Spearman correlation [14] between the manually labeled similarity and similarity computed by embeddings, which computed as the cosine similarity of word pair's embeddings generated by the model. The evaluation results are shown in Table 3.

[1] https://github.com/NLPchina/ansj_seg

Table 3. Word similarity result

Model	Wordsim-240	Wordsim-296
Word2vec	0.4221	0.4479
CWE	0.4363	0.4750
SCWE	0.4311	0.4648
JWE	0.4467	0.5831
RECWE-p1	0.4962	0.5849
RECWE-p2	0.5011	0.5554
RECWE-p3	0.5290	0.5765

From the results, we can see that RECWE outperforms other models in two similar databases. Compared with the SCWE, the CWE has improved word2vec results. However, since only the Chinese character information in the word is used, the result of the JWE with the addition of radical information is worse than that of the RECWE. It can also be seen that, in the wordsim-240 database, the RECWE is greatly improved compared to the JWE, especially after the addition of the radical word of the target word. This is mainly because the database contains mostly word pairs that are semantically affiliated, and the RECWE with a radical transformation mechanism can find such relationships well. For example, for the words "淋浴(shower)" and "水(water)", the RECWE will convert the radical "氵" of character "淋" and "浴" to the character "水" in the pre-processing phase. The intrinsic link between water and water. And then the model can find the inherent connection between "淋浴" and "水". Moreover, comparing the results of the three pattern, it can be found that by adding the sub-information of the target word and word context, the word embedding can be further optimized and the model can achieve better results. However, this feature has little effect on the word sim-296 and may even result in deterioration of the results. Therefore, the specific use depends on the characteristics of the corpus.

3.3 Word analogy

This task is another technique for measuring the quality of word embedding. It mainly judges whether word embedding can reflect the linguistic regularity between word pairs [14]. For example, given the word relationship group "Beijing-China: Tokyo-Japan", a good word embedding should have the word embedding "$vec(Japan)$" of the word "Japan" close to the embedding produced by the expression "$vec(China) - vec(Beijin) + vec(Tokyo)$". That is, given an analogy relationship "$a - b : c - d$", if the word embedding can find the relationship "$a - b : c - x$" by looking for the embedding x in the vocabulary through formula 6, then it is determined that the word embedding contains this analogy relationship.

$$\arg \max_{x \neq a, x \neq b, x \neq c} \cos\left(\vec{b} - \vec{a} + \vec{c}, \vec{x}\right) \tag{6}$$

This task uses the Chinese word analogy database provided in the literature [11], which contains 1124 group word analogy relationships, each group of analogical relations

contains 4 words, all analogy relations are divided into three categories: “Capital” (group 677), “The provincial capitals” (175 groups) and “people” (272 groups).

The accuracy rate is used as a result indicator. The higher the accuracy rate is, the more the word analogy relationship the word vector contains. The experimental results of the three types of word analogy are shown in table 4.

Table 4. All kinds of analogy experimental results

Model	Capital	Provincial capitals	people
word2vec	0.11816	0.11428	0.34558
CWE	0.20787	0.14285	0.32720
SCWE	0.21381	0.14022	0.33021
JWE	0.22101	0.16	0.32352
RECWE-p1	0.23632	0.18285	0.38603
RECWE-p2	0.33479	0.2	0.32079
RECWE-p3	0.39606	0.18857	0.44852

From the result, we can see that compared with other word vector models, the RECWE has achieved optimal results in the three types of analogy relationship, indicating that the radical information can enhance the performance of the model in language law. Particularly in the category of “people”, the radical transformation mechanism allows the RECWE to find the relationship between word pairs relatively easily based on radicals. For example, RECWE can contact the word “妈妈(mother)” and “姐姐(sister)” through radical “女”. Therefore, the accuracy of RECWE compared with the JWE model has greatly improved. For categories such as “Capital” that contain a large number of transliterated words, their internal radicals cannot provide additional semantics, and the promotion is smaller. However, it can be seen from the results of the three patterns of “p1”, “p2”, and “p3” that using the target information and the sub-information corresponding to the word context can alleviate this shortcoming to a certain extent, and the model can be better result.

4 Relate work

With the continuous development of neural networks and deep learning, word embedding has achieved many achievements in English. However, there has been no breakthrough in how to construct word embedding of Chinese. At the beginning of the study, the researchers tried to directly use the English word embedding model (such as the CBOW and the Skip-gram) to train Chinese word embedding on the Chinese corpora after word segmentation. However, there is obviously a problem with this approach: Most English word embedding model use words as the smallest unit of operation when training, while ignoring the internal morphological in-formation of words. Different from English and other alphabetic characters, Chinese characters are still have a lot of semantic information.

To address this issue and make full use of the semantic information of Chinese words, Chen et al. [11] added Chinese character information to the ordinary CBOW and proposed the CWE. Subsequently, Xu et al. [12] based on the CWE, assigning each

character weight in the word to optimize the whole model. However, these models did not use the radicals in Chinese characters and ignored the internal information of all characters.

On the other hand, Sun et al. [8] added radical information to train Chinese character embedding based on the CBOW; Yu et al. [13] designed a multi-granularity word vector model by combining the information of "radical-character" and "character-word" in the JWE. However, these methods simply add the radical information to the model and do not take into account the evolution of the radicals. That is, they do not associate the radicals with the corresponding Chinese characters (such as "氵" to "水"), which makes the information obtained from radicals is very limited, affecting the quality of the word embedding.

5 Conclusion

In this paper, we proposed a Chinese word embedding training model RECWE by effectively using the semantic information Chinese characters and their radicals. Through simplified conversion and radical escaping techniques, RECWE can directly determine the internal semantic relations between Chinese characters based on radicals, allowing words with the same radical to be close to each other in the vector space. Experiments show that our method is more effective than other similar methods in word similarity and word analogy.

Due to the rapid developing of deep learning technology, there're a lot of potential work. About the predict architecture, we will try to employ attention technology in the prediction layer of the model. But more work should to be done on the sematic side, which is to understand Chinese character, and try to employ more features to our model.

Acknowledgements. Financial support for this study was provided by the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2016J198) and Science and Technology Planning Project of Sichuan Province, China (Grant No. 2017JY0080).

References

1. Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(Aug): 2493-2537.
2. Grave E, Mikolov T, Joulin A, et al. Bag of tricks for efficient text classification[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017: 427-431.
3. Tang D, Wei F, Yang N, et al. Learning sentiment-specific word embedding for twitter sentiment classification[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014, 1: 1555-1565.
4. Zhou G, He T, Zhao J, et al. Learning continuous word embedding with metadata for question retrieval in community question answering[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015, 1: 250-259.

5. Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb): 1137-1155.
6. Mnih A, Hinton G E. A scalable hierarchical distributed language model[C]//Advances in neural information processing systems. 2009: 1081-1088.
7. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
8. Sun Y, Lin L, Yang N, et al. Radical-enhanced chinese character embedding[C]//International Conference on Neural Information Processing. Springer, Cham, 2014: 279-286.
9. Li Y, Li W, Sun F, et al. Component-enhanced chinese character embeddings[J]. arXiv preprint arXiv:1508.06669, 2015.
10. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 160-167.
11. Chen X, Xu L, Liu Z, et al. Joint Learning of Character and Word Embeddings[C]//IJCAI. 2015: 1236-1242.
12. Xu J, Liu J, Zhang L, et al. Improve chinese word embeddings by exploiting internal structure[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 1041-1050.
13. Yu J, Jian X, Xin H, et al. Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 286-291.
14. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
15. Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013: 746-751.