Build Chinese Language Model with Recurrent Neural Network

Li Lin, Jin Liu(✉), Zhenkai Gu, Zelun Zhang, and Haoliang Ren

College of Information Engineering, Shanghai Maritime University,
Shanghai, China
{lilin, jinliu, zhenkaigu, zelunzhang,
haoliangren}@shmtu.edu.cn

Abstract. In recent years, the introduction of Deep Learning based machine learning methods have greatly enhanced the performance of Natural Language Processing (NLP). However, most Deep Learning based NLP studies in the literature are aimed at the Latin family languages. There is seldom research which takes Chinese language model as the objective. In this paper, we use Deep Learning method to build language model for Chinese. In our model, the Fully-connected Neural Network which is a popular structure used in NLP is replaced by the Recurrent Neural Network to build a better language model. In the experiments, we compare and summarize the differences between the results obtained by using the original Deep Learning method and our model. And the results prove the effectiveness of our proposed model.

Keywords: Natural language processing · Recurrent neural networks
Deep learning · Language model

1 Introduction

With the arrival of "Big Data" era and the development of Deep Learning based machine learning methods, it is becoming possible for both academia and industry to make use of large amount of data. The advent of word embedding has utilized Unsupervised Learning to extract large amounts of text from the Internet effectively. However, original word embedding methods are designed for Latin languages. There are many similarities and differences between Latin and Chinese languages, so it is necessary for us to make some specific improvements to the word embedding method.

A word is the smallest unit that contains the semantic information. The original segmentation methods work better in word segmentation with general corpus, but they cannot handle unknown words. At the same time, for a professional vocabulary, the segmentation process also depends on the corresponding professional dictionary. The traditional word segmentation method can no longer satisfy Chinese word segmentation now.

The remaining of this paper are as follows. Section 2 introduces the current research situation of Chinese word segmentation. In Sect. 3, we modify the Deep Learning method according to the characteristics of the Chinese language family, and

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*, Lecture Notes in Electrical Engineering 474,
https://doi.org/10.1007/978-981-10-7605-3_146

use the Recurrent Neural Network instead of the Fully-connected Neural Network. Finally, we evaluate the experimental results and draw a conclusion in Sect. 4.

2 Related Work

In 1954, Harris proposed the Distribution Hypothesis, that is, “If the words are similar in context, their semantics are similar, too [1]”. It provided a theoretical basis for the distribution representation of words. After that, researchers have proposed a variety of word representation models based on the Distribution Hypothesis, such as matrix-based LSA model [2], clustering-based Brown clustering model [3], sense-aware semantic analysis (SaSA) [4] and so on.

In 1992, Brown et al. [3] designed a context clustering model to construct word embedding, which created a precedent for the distributed word representation model based on clustering. Xu and others [5] proposed an embedded representation method for fixed length words based on neural networks. Its idea was the case that, a given vector dimension determines the size of the vector space. The neural network model can be used to model the context in a composite way. Dual Embedding Space Model (DESM) [6] was proposed by Nalisnick et al. in 2016, which provided evidence that a document is about a query term, and the performance of the word2vec model is application dependent.

Lai et al. [7] analyzed three critical components in training word embedding: model, corpus, and training parameters. They discovered that corpus domain is more important than corpus size. So we choose a corpus in a suitable domain for the desired task, after that, using a larger corpus yields better results. A framework for the Chinese Lexical Similarity Computation (CLSC) [8] task was proposed by Pei et al. It measured the Chinese word similarity by combining word embedding and Tongyici Cilin and utilize retrieval techniques to extend the contexts of word pairs and calculate the similarity scores to weakly supervise the selection of a better result, which rank No. 2 with the result of 0.457/0.455 of Spearman/Pearson rank correlation coefficient.

In recent years, the word embedding model has performed well in Latin language. As for Chinese language, it has to do with word segmentation first, so the extraction of morphemes becomes difficult and there may be the case of the error.

3 Chinese Language Model Based on Recurrent Neural Networks

In order to generate character embedding effectively and use the character sequence information to segment the word better, this system uses the Recurrent Neural Network as the hidden layer of word embedding generation model and statistical segmentation model.

Recurrent Neural Networks (RNNs) is an input neural network model designed to handle inputs that contains sequence information. The traditional neural network model is a fully-connected model, where the input layer and the hidden layer are linked while there is no link between the hidden layers. This design makes it difficult for the hidden

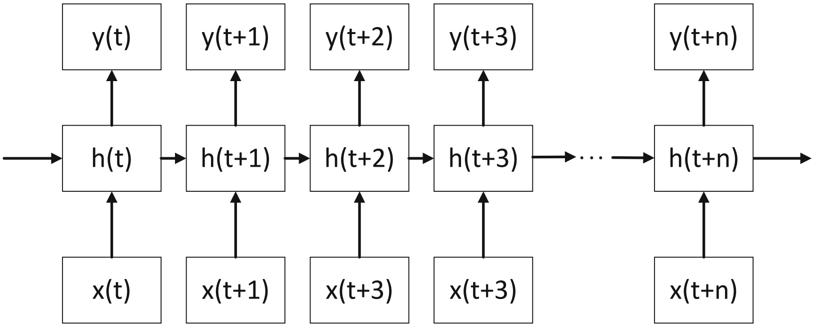

Fig. 1. Structure chart of recurrent neural networks.

layer to acquire the input sequence relationship. The structure chart of Recurrent Neural Networks is illustrated in Fig. 1.

In the Recurrent Neural Network, the input is no longer solely linked to the hidden layer. Here the input is split into multiple timing. In each time, the input layer and the hidden layer are fully-connected, and the output is not just as the current timing output, but also as the input *h(t−1)* for the next timing. In conjunction with the next timing input *x(t−1)*, we can obtain the next timing output y(t).

We use *W* to represent the weight matrix of the hidden layer. In order to simplify the expression, the bias *B* is set to 0, and the *tanh* function is used as the activation function, then the Fully-connected Neural Network can be expressed as:

$$Y = tanh(W * X) \tag{1}$$

And the Recurrent Neural Networks can be expressed as:

$$Y_t = tanh(W * (X_t + Y_{t-1})) \tag{2}$$

By comparing Formulas 1 and 2, it can be observed that, in the recurrent neural networks, the output of the previous sequence will affect the output of the current sequence. In other words, the output of the current sequence is not only based on the input of the current sequence, but also implicitly on the input of the previous sequence, so that the model can extract the sequence relations between inputs.

4 Experiment Results and Analysis

In order to compare the effect of the Fully-connected and the Recurrent Neural Networks for Chinese word segmentation, this paper uses the same character embedded in the comparative experiments, extracts 320 million words from Wikipedia, and use the recurrent neural network model as generation method.

The training corpus used in Chinese segmentation is a set of manually annotated Chinese word segmentation corpus provided by Peking University, and the test set is also the corresponding test set provided by Peking University. Due to the generation of a character vector does not depend on the training set and test set, but use the no-related Wikipedia data, the test set and the training set is not coincide, so the test results can be considered as open training.

After comparing experiments, the results are as follows:

Table 1 and Fig. 2 show the results of the comparison between the three groups of reference and five other models. Among them, the first two groups are the best results in Chinese word segmentation evaluation competition held in 2005. The third group is the most commonly used Chinese word segmentation result based on manual selection feature.

Table 1. Experimental results of network structure contrast

Number	Experiment	Accuracy rate	Recall rate	F value	Unknown words recall rate
	Best results (closed)	0.946	0.953	0.950	0.636
	Best results (open)	0.968	0.969	0.969	0.838
#1	Maximum entropy	0.953	0.945	0.949	0.825
#2	Fully connected network	0.942	0.937	0.939	0.787
#3	CRF	0.948	0.942	0.935	0.784
#4	HMM	0.932	0.928	0.931	0.777
#5	SVM	0.929	0.921	0.920	0.767
#6	RNN	0.947	0.935	0.940	0.791

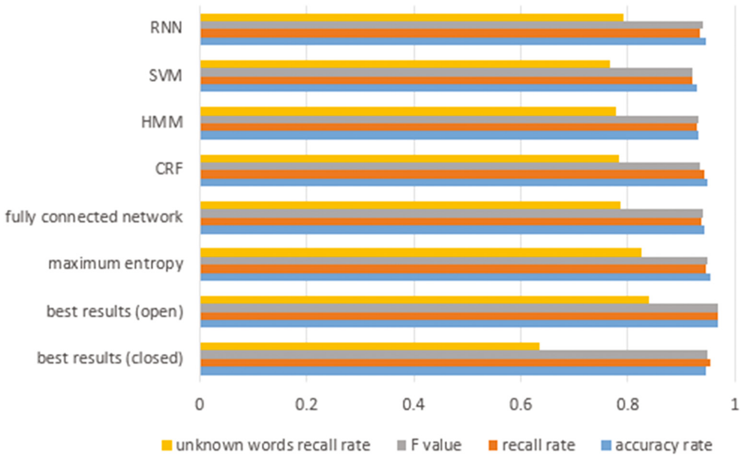

Fig. 2. Experimental results of network structure contrast

In contrast experiment #1, #2, #3, #4, #5 and #6, we found that in the neural network model without manual selection, it is difficult to achieve the performance of artificial selection if it only relies on the character embedding generated by large corpus. This paper speculated that this may result from the following aspects:

(a) The theme of generating character embedding does not match the subject of Chinese word segmentation training set.
(b) The generation of character embedding is not perfect.
(c) The size of the training set does not reach the threshold applicable to the neural network.
(d) The training process of the neural network needs to be improved.

5 Conclusion

The traditional methods for Chinese word segmentation usually use dictionary which are built by human. However, the new Chinese words are created fast with the development of Internet; and this requires an automatic method to do the word segmentation job.

In this paper, we proposed a novel Chinese word segmentation model based on recurrent neural network. Compared to other network structure like fully-connected network, or some traditional models like HMM, CRF, SVM, our model achieves better performance in unknown words segmentation. In the future, we will focus our research on the word embedding technology, change it to be more suitable for Chinese language.

References

1. Harris, Z.S.: Distributional structure. World **10**(2–3), 146–162 (1981). Springer, Netherlands
2. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Process. **25**(2–3), 259–284 (1998)
3. Brown, P.F., Desouza, P.F., Mercer, R.L., et al.: Class-based n-gram models of natural language. Comput. Linguist. **18**(4), 467–479 (1992)
4. Wu, Z., Giles, C.L.: Sense-aware semantic analysis: a multi-prototype word representation model using Wikipedia. In: Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 2188–2194 (2015)
5. Xu, W., Rudnicky, A.: Can artificial neural networks learn language models? In: Sixth International Conference on Spoken Language Processing (2000)
6. Nalisnick, E., Mitra, B., Craswell, N., et al.: Improving document ranking with dual word embedding. In: International Conference Companion on World Wide Web, pp. 83–84 (2016)
7. Lai, S., Liu, K., He, S., et al.: How to generate a good word embedding. IEEE Intell. Syst. **31** (6), 5–14 (2016)
8. Pei, J., Zhang, C., Huang, D., et al.: Combining word embedding and semantic lexicon for Chinese word similarity computation. In: International Conference on Computer Processing of Oriental Languages, pp. 766–777 (2016)

9. Bengio, Y., Ducharme, R., Vincent, P.: A neural probabilistic language model. In: Advances in Neural Information Processing Systems, pp. 932–938 (2001)
10. Turian, J., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 384–394 (2010)
11. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., Khudanpur, S.: Recurrent neural network based language model. In: Conference of the International Speech Communication Association, INTERSPEECH 2010, Makuhari, Chiba, Japan, September, DBLP, pp. 1045–1048 (2010)
12. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations Workshop Track (2013)
13. Lai, W.: Research on semantic vector representation of words and documents based on neural networks. Institute of automation, Chinese Academy of Sciences (2016)