A Word2vec based on Chinese Medical Knowledge

Jiayi Zhu*, Pin Ni*†, Yuming Li*†, Junkun Peng*, Zhenjin Dai*, Gangmin Li*‡ and Xuming Bai§

*Research Lab for Knowledge and Wisdom, Xi'an Jiaotong-Liverpool University, China
†Department of Computer Science, University of Liverpool, UK
§Department of Interventional Radiology, Second Affiliated Hospital of Soochow University, China
Email: ‡Gangmin.Li@xjtlu.edu.cn, §2005baixuming@163.com

***Abstract*—Introducing a large amount of external prior domain knowledge will effectively improve the performance of the word embedded language model in downstream NLP tasks. Based on this assumption, we collect and collate a medical corpus data with about 36M (Million) characters and use the data of CCKS-2019 as the test set to carry out multiple classifications and named entity recognition (NER) tasks with the generated word and character vectors. Compared with the results of BERT, our models obtained the ideal performance and efficiency results.**

***Index Terms*—Word Embedding, Language Model, EMR**

I. Introduction

BERT (Bidirectional Encoder Representation from Transformers) is the current one of the state-of-the-art pre-training language models, however, because of the character of its large amount of calculation, slow training speed and high equipment requirements do not conform to be deployed to the actual medical environment. Therefore, we hope to get a fast and lightweight method to apply NLP to the medical field.

Our contributions are as follows:

- Collected and sorted out a medical corpus of about 36M (Million) Chinese character from 4 mainstream Chinese medical wiki and 13 medical books.
- Word vector and character vector were generated based on Word2vec. As far as we know, these are the first open word and character embedding pre-training language models in the Chinese biomedical field. We respectively based on the word vector and character vector of Word2vec in two main downstream NLP tasks executing a detailed competitive experiment. The experimental results show that in these two tasks, our model achieves a difference of only around 1% F1-score with BERT (both classification and NER tasks) when taking the better one in word vector and character vector. More importantly, our models have more outstanding performance in terms of loading and training efficiency than BERT. They can provide a acceptable balance of performance and efficiency compared to BERT, which provides the possibility for applications in real medical scenarios.

This paper uses word2vec, a pre-training model, to show the word vector trained by our data set, and shows the effect of the word vector obtained by us on the downstream tasks.

II. Related Work

Over the past few years, researchers are continually breakthrough and innovate for better pre-training models. Following the BERT [1] released and refreshed the results of most natural language processing tasks, XLNet [2], GPT-2 [3] models consecutively released and achieved excellent results. All the efforts they strive for is to make the expression of corpus to get better effect, whether in semantics, syntax, and the relationship between the words.

Lee and Yoon et al. [4] proposed the BioBERT, which is a pre-trained biomedical language model for biomedical text mining in English and obtains outstanding effects in the English Medical field. On account of lacking Chinese network resources, the medical corpus in Chinese is more difficult to obtain than in English, which is the problem we want to overcome.

III. Methodology and Experiment

The general flow of the experiment is shown in Fig.1.

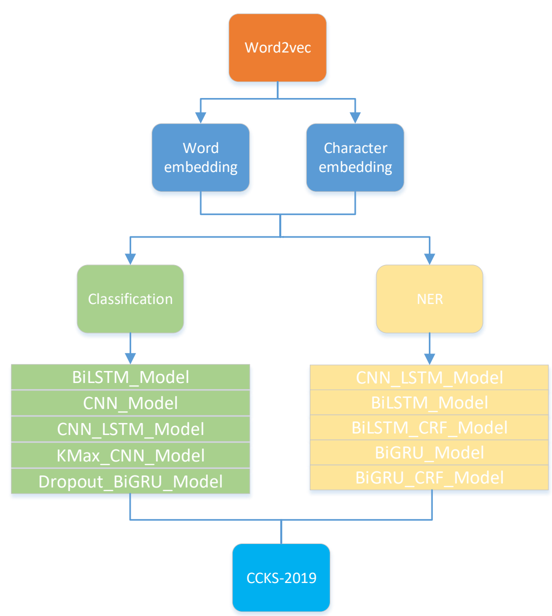

Fig. 1. The general flow of the experiment

A. Pre-train Language Model

The first problem of NLP is to find an appropriate text representation because a precise text representation is a basis for further processing. Previously, the widely used text representation was a one-hot vector, which set the corresponding words as 1 in the position of the corresponding vector after sorting the corpus, and the others are set as 0. This representation method has two disadvantages:

- 1) The input dimension is too large. It increases the number of parameters and calculation of the model, and contains a large number of zeros, making the vector too sparse.
- 2) No semantic information. The inner product of two words is equal to 0, which means these two words have not any relationship.

This paper uses word2vec pre-training model, which uses a method called word embedding, proposed by Bengio, Y. [5], to represent each word and generate word vector. The general process is defining a window size for each word, meaning that the contact range of each word with the words of its context and then according to the distances from the center word gives a weight to each word. In this way, a word co-occurrence matrix and word vector table can be obtained. The whole method can be explained by two models CBOW and Skip-gram, proposed by Mikolov, T. [6].The word vectors obtained by this method will have relevance, and the vector dimension is greatly reduced, which reduces the number of parameters and computation of the model.

Results will be not ideal if word2vec is used for medical downstream task training based on parameters pre-trained from large data sets (e.g. Wikipedia). To obtain a better word vector in the medical field, we have collected medical corpus data with about 36M (Million) characters.

The descriptions of the training corpora for word embedding are listed in Table 1.

TABLE I
CORPORA DESCRIPTION

Corpus Source	Description	Size (word)
Medical Books	The total number of medical books we used is 13, involving vaious medical branches, e.g. diagnostics, immunology,and oncology. The sources can be find from some websites.	4384503
SAHSU	Descriptions of symptoms of patients in The Second Affiliated Hospital of Suzhou University.	2002202
Institutions	All the data were collected from four institutions including 39 Health, Baike Health, Feihua Health, and Wangyi Health, and they were transformed to 'disease and symptoms descriptions' format.	29092216

Here are the steps how we get the word vector:

- 1) Data processing. Symbols, Numbers, words, and other characters that affect the semantics, as well as those that are not medical related, e.g. references, are deleted. Next, transform the corpus format into the input pattern acceptable to word2vec.
- 2) Add dictionaries to word2vec model. Corpus data obtained through tokenizer, e.g. Jieba, often separate medical terms, hence we collates a 5,321 words medical dictionary from CCKS-2019 and let the tokenizer retain the words in it, to obtain a better semantic effect.
- 3) Run the models and save the result.

At the same time, we suppose that embedding the corpus in character will get better effect, then we also generate a character vector.

The results are 50-dimensional word and character vectors that we visualized as 2-D dot graphs in Fig.2 and Fig.3.

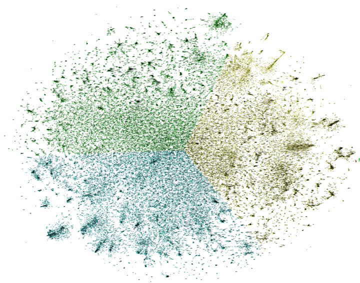

Fig. 2. word vector

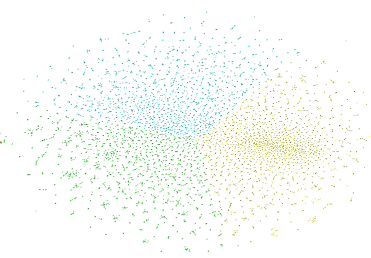

Fig. 3. character vector

After observation, it is found that the drug name and the dose are almost in the yellow area, then the disease name and the anatomical part are generally in the green area, also, the operation name is mostly in the blue area.

B. Downstream task

Downstream tasks are aiming to handle the problems in life. Here we choose two tasks to test our generated word and character vectors:

Named Entity Recognition is one of the most fundamental natural language processing tasks. It aims to recognize specific proper nouns in a medical corpus. For our task is to identify six categories of entities, e.g. disease and diagnosis, operation, and anatomic site.

Classification is a task to classify a text according to its contents. For our task is to classify the text into six categories, including causes, descriptions, diagnosis, prevention, symptom, and treatment.

C. Experiment Environment

Our experimental environment is as follows: CPU Intel Xeon E5-2678 v3, RAM: Dual 2.50GHz, GPU: Dual Nvidia GeForce GTX 1080 Ti.

IV. RESULTS

Although since the release of BERT, has beaten almost all the previous baseline approaches of NLP, the BERT still has a lot of problems under the efficiency and some special conditions. Therefore, we hope to provide a lightweight model in the medical background to replace the huge and high-cost BERT as an alternative to applied for some limited-resource clinical scenarios. Here, we compare the performance of the official BERT Chinese model and our pre-trained Word2Vec

TABLE II
RESULTS OF THE EXPERIMENT (W:WORD VECTOR; C: CHARACTER VECTOR)

Task-Type	Model	W\C	Word2vec				BERT			
			Precision	Recall	F1-score	Time (s/epoch)	Precision	Recall	F1-score	Time (s/epoch)
NER	BiLSTM_CRF	W	0.6631	0.7371	0.6978	29	0.6897	0.7716	0.7282	54
		C	0.6756	0.7600	0.7149	29				
	CNN_LSTM	W	0.5847	0.6632	0.6209	2	0.6033	0.6910	0.6421	24
		C	0.5823	0.6524	0.6146	2				
	BiLSTM	W	0.6075	0.6925	0.6462	3	0.6852	0.7640	0.7214	28
		C	0.6537	0.7326	0.6904	3				
	BiGRU	W	0.6380	0.7172	0.6750	3	0.7047	0.7437	0.7221	27
		C	0.6961	0.7645	0.7284	3				
	BiGRU_CRF	W	0.6979	0.7524	0.7234	28	0.7214	0.7820	0.7499	53
		C	0.7096	0.7884	0.7464	29				
Classification	BiLSTM	W	0.9391	0.9387	0.9385	81	0.9583	0.9579	0.9579	807
		C	0.9431	0.9423	0.9424	80				
	CNN_LSTM	W	0.9433	0.9431	0.9431	12	0.9605	0.9598	0.9600	587
		C	0.9398	0.9383	0.9386	13				
	Dropout_BiGRU	W	0.9580	0.9580	0.9579	50	0.9615	0.9612	0.9612	630
		C	0.9566	0.9561	0.9562	51				
	CNN	W	0.9303	0.9299	0.9299	9	0.7075	0.8064	0.7452	585
		C	0.9119	0.9107	0.9105	9				
	DPCNN	W	0.9019	0.8814	0.8826	93	0.9490	0.9482	0.9476	1532
		C	0.9380	0.9339	0.9342	92				

in various downstream task models based on the CCKS-2019 test set.

The experiment results are illustrated in Table 3.

A. Named Entity Recognition

From the comparison between word vector and character vector, we can know that the character vector performs better than the word vector. Comparing word2vec with BERT shows that BiLSTM and BiGRU based on word2vec have the largest difference from BERT-based (about 6%). While CNN-LSTM based on word2vec, whose effect is the worst, shows a similar result and the least time consumption, the result is still a little worse than BERT-based (about 2%). For the two CRF methods, the results are the best and also absolutely close to the results of BERT-based, despite the longest time consumption, which is about 10 times more than the other models in word2vec.

B. Classification

From the comparison of the results of classification, except CNN (word vector is better) and DPCNN (character vector is better), other models show similar results between word vector and character vector, and we can know that BiLSTM and CNN-LSTM based on word2vec have only 2% difference from BERT-based. Dropout-BiGRU performs the best result with 0.5% distinction from BERT-based among 5 models and the highest values of three indexes. In addition, the time consumption of all the BERT-based models is higher than word2vec in approximately 10 times.

V. CONCLUSIONS

BERT was trained based on the corpus of a huge amount of Wikipedia corpus [1], surprisingly, we get close to BERT in downstream tasks results with a small amount of medical field corpus. Word2vec is faster and more lightweight than BERT in both model loading and training process, since the number of training model parameters in word2vec is only about 2M, and BERT is about 110M [1], which is 55 times higher than ours. In addition, the training of BERT is highly dependent on the computing resource (e.g. GPU and TPU), whereas, in the actual medical equipment deployment scenarios, it is less likely to provide such expensive computing resources to the actual production and operation. Therefore, the Word2vec training through domain knowledge can be deployed in the lower configuration device, and perform the downstream tasks faster and more practical.

VI. ACKNOWLEDGEMENT

This work is partially supported by the AI University Research Centre (AI-URC) through XJTLU Key Programme Special Fund (KSF-P-02) and KSF-A-17. And it is also partially supported by Suzhou Science and Technology Programme Key Industrial Technology Innovation programme with project code SYG201840. We appreciate their support and guidance.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[2] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” 2019.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.

[4] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: a pre-trained biomedical language representation model for biomedical text mining,” 2019.

[5] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, J. Kandola, T. Hofmann, T. Poggio, and J. Shawe-Taylor, “Journal of machine learning research 3 (2003) 1137–1155 submitted 4/02; published 2/03 a neural probabilistic language model,” 03 2003.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013.