On Learning Better Word Embeddings from Chinese Clinical Records: Study on Combining In-Domain and Out-Domain Data

Yaqiang Wang[1*], Yunhui Chen[2], Hongping Shu[1], Yongguang Jiang[2]
[1] Department of Software Engineering, Chengdu University of Information Technology, Chengdu, Sichuan 610225, China
[2] School of Fundamental Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan 610075, China
*Corresponding author: yaqwang@cuit.edu.cn

Abstract

High quality word embeddings are of great significance to advance applications of biomedical natural language processing. In recent years, a surge of interest on how to learn good embeddings and evaluate embedding quality based on English medical text has become increasing evident, however a limited number of studies based on Chinese medical text, particularly Chinese clinical records, were performed. Herein, we proposed a novel approach of improving the quality of learned embeddings using out-domain data as a supplementary in the case of limited Chinese clinical records. Moreover, the embedding quality evaluation method was conducted based on Medical Conceptual Similarity Property. The experimental results revealed that selecting good training samples was necessary, and collecting right amount of out-domain data and trading off between the quality of embeddings and the training time consumption were essential factors for better embeddings.

1 Introduction

Word embeddings, or embeddings for short, have been widely used in various natural language processing tasks, such as language modeling (Bengio et al., 2003; Sundermeyer, et al. 2012; Adams et al., 2017), syntactic parsing (Grefenstette et al., 2014; Tu et al., 2017) and part-of-speech tagging (Yang and Eisenstein, 2016). Owing to the advantage of embeddings in boosting performance, a surge of interest in applying embeddings has become increasingly evident with numerous encouraging results in the field of biomedical applications, e.g. disease prediction (Miotto et al., 2016), clinical events prediction (Choi et al., 2016a), medical concept disambiguation (Tulkens et al., 2016), and biomedical information retrieval (Mohan et al., 2017).

Learning embeddings from English medical texts, as a hot topic in recent years, has been extensively studied due to the efforts of open datasets, such as UMLS of NLM (Bodenreider, 2004), medical journal abstracts from PubMed (Choi et al., 2016a), and some released clinical data (Finlayson, et al., 2014; Stubbs and Uzuner, 2015). These datasets have been widely used as gold standards by the biomedical natural language processing domain for learning embeddings (De Vine et al., 2014; Choi et al., 2016b).

However, the development of learning embeddings from Chinese medical texts has fallen far behind, especially from Chinese clinical records. Due to the privacy concerns, Chinese clinical records that can be used are generally limited. Learning better embeddings based on neural network architectures, for instance the widely used skip-gram model (Mikolov et al., 2013a), usually needs a large number of training data. As a result, the learned embeddings from Chinese clinical records are not good enough.

Moreover, to the best of our knowledge, there is a limited number of studies focusing on learning embeddings from Chinese clinical records, not to mention the embedding evaluation. Many methods have been developed to learn embeddings from English medical texts, however, Chinese medical texts, especially clinical records, have their particular language features. Therefore, adaptions to the approaches of learning embeddings from English medical texts are urgently needed for learning embeddings from Chinese clinical records.

In this paper, we focused on learning embeddings from Chinese clinical records, and our major contributions were as follows:

- We proposed an in-domain and out-domain data combination method for learning better

Proceedings of the BioNLP 2018 workshop, pages 177–182
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

embeddings from Chinese clinical records by the skip-gram model under the situation that we only have limited Chinese clinical records.

- Referring to the evaluation method for medical concept embeddings proposed in (Choi et al., 2016b) which is based on medical conceptual similarity property, we proposed a method for distantly evaluating the learned embeddings from Chinese clinical records using an additional standard medical terminology dataset.

- We found that selecting good training samples is necessary. Collecting right amount of out-domain data, trading off between the quality of embeddings and the training time consumption are essential factors for better embeddings.

2 Skip-Gram Model for Learning Embeddings

The skip-gram model is one of the most popular methods for learning embeddings from texts. The training objective of the skip-gram model is to find an embedding that is useful for predicting context words of one target word in a sequence. The sequence usually refers to a sentence in a specific task. In the skip-gram model, if two different target words w_k and $w_{k'}$ have (very) similar context words, then learned embeddings of w_k and $w_{k'}$ by the model would be (very) similar, because a common output weight matrix is used (Mikolov et al., 2013b). In other words, if we want to clearly distinguish two target words' embeddings, we can provide more informative context words that differentiate the target words.

The skip-gram model has been used in various domain to learn embeddings from different types of texts, and there have been also various relevant attempts to learn embeddings from medical texts by the skip-gram model. Most works directly applied the model on various medical corpora to complete this domain-specific task (Giménez et al., 2013; Liu, et al., 2016). In this paper, we continued the previous work using the skip-gram model to learn embeddings from Chinese clinical records to further explore a data combination method for improving the quality of the learned domain-specific embeddings.

3 Skip-Gram Model for Learning Embeddings from Chinese Clinical Records

3.1 Observation

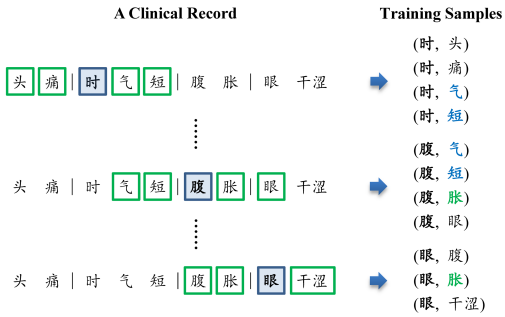

Figure 1: An example of training sample generating process of the skip-gram model.

Content of Chinese clinical records are usually brief, the occurrence of symptoms and diseases has certain correlation, and doctors have a certain habit in inquiring procedures and making records. These domain-specific characteristics challenge learning embeddings from Chinese clinical records, because it gives general domain words a high probability of having similar or even identical context words to those medical words. For example, in Figure 1, general domain word “时” (sometimes) and medical term “眼” (eye, the body part) have similar context words with medical word “腹” (abdomen, the body part), and “时” (sometimes) has more common context words with “腹" (abdomen) than “眼" (eye). Moreover, it would like to be a fixed pattern to describe certain medical problems. As a result, learned embeddings of “时” (sometimes) and “腹” (abdomen) would be more similar than embeddings of “眼" (eye) and “腹” (abdomen), although “腹” (abdomen) and “眼” (eye) belong to the same type of medical concept (i.e. the body part).

In summary, the main challenge of learning better embeddings from Chinese clinical records is to let the skip-gram model make a clearer distinction between medical words and general domain words.

3.2 Usage of Out-Domain Data

As mentioned earlier, making a clearer distinction between learned embeddings of two target words by skip-gram model requires more evidences, i.e. adding diverse context words to illustrate the difference between the two target words. Therefore,

we proposed a hypothesis that adding general domain Chinese texts, i.e. the out-domain data, to Chinese clinical records, i.e. the in-domain data, would facilitate the learning of embeddings from Chinese clinical records. The intuition is that the medical words in Chinese clinical records have domain-specific usage but are not widely used in the out-domain data. However, the general domain words have a wide range of usage in the out-domain data, which is the exact opposite of using medical words. Combining out-domain data with Chinese clinical records can improve the diversity of context words of the general domain words, but without the side-effect of impairing the contexts of the medical words. Better embeddings, in turn, can be learned from the combined data.

3.3 Learning Process and Embedding Quality Evaluation Method

Chinese clinical records were segmented into words by the latest version of Stanford CoreNLP tool[1] with default settings, and adjacent words appearing in our prepared standard medical word dataset would not be segmented (Zhang et al., 2016). Punctuations were removed. Out-domain data went through a similar process but without the second process. We assume that in out-domain data there is no medical words. We directly applied skip-gram model implemented by DeepLearning4J[2] to learn embeddings. Hierarchical SoftMax is used in training process, and context window size and embedding dimensionality are set to 5 and 200 respectively (Choi et al, 2016b).

We used an intrinsic evaluation method, named Chinese Medical Concept Similarity Measure (CMCSM), to distantly measure quality of learned embeddings. CMCSM is defined below:

$$CMCSM = \frac{1}{N}\sum_{i=1}^{N}\frac{2}{C_i(C_i-1)}\sum_{j=1}^{C_i-1}\sum_{k=j+1}^{C_i} s(c_j, c_k) \quad (1)$$

where N is the number of groups of the medical words in the same level of a prepared medical word dataset $\mathbf{C}$, $C_i \in \mathbf{C}$ is one group of the medical words, and c_j and c_k are the jth and kth terms in C_i. $s(c_j, c_k)$ is any commonly used embedding similarity measure (Levy et al., 2015). In this paper, we used the cosine measure.

[1] URL: https://nlp.stanford.edu/software/segmenter.shtml.
[2] URL: https://deeplearning4j.org/.

Dataset		Size
CCRD		25056
ODD		3010739
SMTD	Number of Terms	3617
	Number of Groups	39

Table 1: Detailed Information of the Experimental Datasets.

4 Experiments

4.1 Experimental Data

To validate performance of the proposed method, three experimental datasets were used in this paper, including a Chinese clinical records dataset (CCRD) collected from Teaching Hospital of Chengdu University of Traditional Chinese Medicine, a large scale out-domain dataset (ODD) obtained from the NLPCC 2018 Shared Task 4[3], and a standard medical terminology dataset (SMTD) gotten from WHO[4]. Medical terms in SMTD are organized into a two-layer tree structure. Index of the second layer defines the group id for medical words. Medical words in the same group are more similar. SMTD was used as the prepared medical word dataset $\mathbf{C}$ mentioned previously. The detailed information of these datasets was listed in Table 1.

4.2 Experimental Data

Firstly, we applied skip-gram model to learn embeddings from CCRD and the learned embeddings were evaluated by CMCSM. We sampled 5 sub-datasets from CCRD in order to assess effect of different size of datasets on quality of the learned embeddings. The sizes of the sampled datasets were 80%, 60%, 40%, 20% and 10% of instances in the original CCRD. The sampling process was a recursive sampling without replacement. It implied that more data means more stable learning results of embeddings. Moreover, we ran the above process 10 times to further assess the stability of the results. The results were used as the baseline, and they were shown in Table 2.

We found in Table 2 that the more Chinese clinical records were used for learning embeddings, the smaller variance of CMCSM tended to be achieved. Moreover, an interesting result was that the use of all Chinese clinical records did not nec-

[3] URL: http://tcci.ccf.org.cn/conference/2018/cfpt.php.
[4] We filtered the terminologies which do not appear in CCRD. URL: http://www.wpro.who.int/publications/who_istrm_file.pdf?ua=1.

essarily result in the highest quality of embeddings. It implies that if we only use in-domain data to learn embeddings, we should collect as much training data as possible and also select helpful samples from the collected data.

Secondly, we applied skip-gram model to learn embeddings from combinations of CCRD and ODD with different combination ratios. Results were listed in Table 3, indicating through combining ODD into CCRD, the qualities of the learned embeddings in different conditions were improved dramatically. More ODD data is combined into CCRD, better embeddings would be learned. In the best case (combining the "Time 2-60%" dataset with the "ODD-ALL" dataset), CMCSM increased by 3.8 times.

	10%	20%	40%	60%	80%	100%
Time 1	0.00218	0.00254	0.00238	0.00259	**0.00268**	0.00228
Time 2	0.00210	0.00238	0.00234	**0.00269**	0.00248	
Time 3	0.00183	0.00220	0.00255	**0.00281**	0.00241	
Time 4	0.00188	**0.00254**	0.00225	0.00235	0.00232	
Time 5	0.00132	0.00218	**0.00247**	0.00226	0.00226	
Time 6	0.00229	0.00248	**0.00297**	0.00255	0.00268	
Time 7	0.00134	0.00220	0.00209	**0.00264**	0.00241	
Time 8	0.00189	0.00256	0.00261	0.00242	**0.00263**	
Time 9	0.00141	0.00213	0.00228	**0.00258**	0.00234	
Time 10	0.00199	**0.00269**	0.00255	0.00248	0.00253	
Mean	0.00182	0.00239	0.00245	**0.00254**	0.00247	-
Variance	1.11E-07	3.56E-08	5.32E-08	2.42E-08	2.08E-08	-

Table 2: CMCSM Results of the Embeddings Learned from CCRD by the Skip-Gram Model.

Notably, the highest quality of the learned embeddings in each row of Table 3 was not always achieved when all data in ODD was used. This result was consistent with the result mentioned earlier, indicating that we should collect as much training data as possible and also need to pay attention to reasonably choosing training samples. In addition, the results showed that when the amount of ODD was 1000 times of the basis size of CCRD, optimal embeddings would be achieved.

Moreover, the results suggested that, in practice, the trade-off between quality of embeddings and training time consumption should be considered. Figure 2 displayed that with increasing the amount of the combined ODD, the growth rate of CMCSM of learned embeddings from basis size of CCRD decreased sharply. Furthermore, when the amount of the combined ODD was more than 50 times of the basis size, the growth rate was almost converged. While, as we know, more data were used for learning embeddings by skip-gram model, much more time would be consumed. We should consider whether it is worthwhile to spend a lot of training time in exchange for very little quality improvement. Moreover, little quality improvement sometimes may not improve performance of downstream biomedical applications.

5 Discussion

This paper conducted only intrinsic evaluation and requires further research involving results from extrinsic evaluations. The high quality embeddings from intrinsic evaluations is also essential for enhancing performance in downstream applications.

Experimental results in this paper casted light on the quality improvements of learning embeddings from English clinical records. Most of the existing studies about how to train good embeddings are based on data within the same domain (Chiu et al., 2016; Lai et al., 2016).

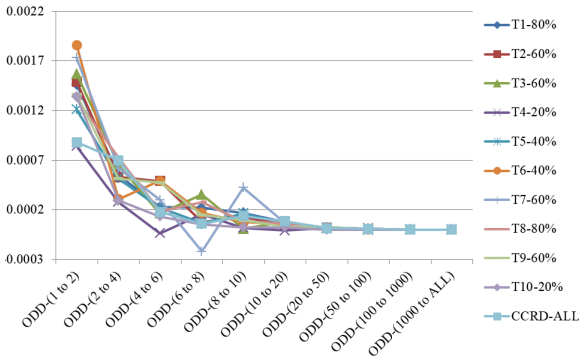

Figure 2: An Example of SMTD.

Further exploration needs to be continued in many aspects. For instance, how to thoroughly understand learning embeddings via complicated neural networks, which is one of current major research hotspots. Only when the complex back-

ground theory is fully interpreted, can we apply this invaluable technology in a flexible way.

	ODD-1	ODD-2	ODD-4	ODD-6	ODD-8	ODD-10	ODD-20	ODD-50	ODD-100	ODD-1000	ODD-ALL
T1-80%	0.0036	0.0050	0.0061	0.0066	0.0070	0.0074	0.0082	0.0087	0.0091	**0.0098**	0.0097
T2-60%	0.0044	0.0059	0.0070	0.0080	0.0081	0.0084	0.0089	0.0095	0.0097	**0.0102**	**0.0102**
T3-60%	0.0041	0.0057	0.0070	0.0074	0.0081	0.0081	0.0089	0.0093	0.0096	**0.0101**	0.0099
T4-20%	0.0056	0.0064	0.0070	0.0069	0.0072	0.0073	0.0072	0.0076	**0.0083**	0.0078	0.0079
T5-40%	0.0046	0.0058	0.0068	0.0073	0.0074	0.0077	0.0084	0.0089	0.0089	0.0089	**0.0092**
T6-40%	0.0050	0.0069	0.0075	0.0085	0.0088	0.0090	0.0095	0.0100	0.0101	0.0102	**0.0103**
T7-60%	0.0041	0.0058	0.0071	0.0077	0.0072	0.0081	0.0088	0.0093	0.0094	**0.0100**	**0.0100**
T8-80%	0.0036	0.0049	0.0063	0.0067	0.0073	0.0074	0.0081	0.0088	0.0092	**0.0098**	**0.0098**
T9-60%	0.0038	0.0051	0.0062	0.0071	0.0074	0.0076	0.0079	0.0088	0.0090	**0.0094**	**0.0094**
T10-20%	0.0061	0.0074	0.0080	0.0083	0.0084	0.0084	**0.0087**	**0.0087**	**0.0087**	**0.0087**	**0.0087**
CCRD-ALL	0.0035	0.0044	0.0058	0.0062	0.0063	0.0066	0.0074	0.0079	0.0083	**0.0091**	**0.0091**
Mean	0.0044	0.0058	0.0068	0.0073	0.0076	0.0078	0.0083	0.0089	0.0091	**0.0095**	**0.0095**

Table 3: CMCSM Results of the Embeddings Learned from the Combinations of CCRD and ODD by the Skip-Gram Model. "T*n*-*X*%" means that "the dataset is the *X*% data of CCRD which is used for learning the highest quality of embeddings in Table 2 at T*n*," and "CCRD-ALL" means that all instances in CCRD are used. "ODD-*n*" means that "the size of ODD currently used is '*n*'×2505." "ODD-ALL" means all samples in ODD are used. 2505 is the basis size of CCRD, and it is approximately equal to the number of 10% of CCRD.

6 Conclusions

This paper presented study on how to learn better embeddings from Chinese clinical records with the supplement of out-domain data in the context of limited in-domain data. Proceeding from the Medical Conceptual Similarity Measure (Choi et al., 2016b), we applied it to distantly evaluate the quality of embeddings. The experimental results showed that a combination use of out-domain and in-domain data could potentially improve the quality of learned embeddings; collecting right amount of out-domain data, trading off between the quality of embeddings and the training time consumption, choosing the good training samples were all essential factors for learning better embeddings. Our results also proved that more data did not necessarily bring more satisfying results, which was consistent with results of Chiu et al. (2016).

Acknowledgments

Authors are pleased to acknowledge the National Natural Science Foundation of China (Grant No. 61501063), the Scientific Research Foundation of Science and Technology Department of Sichuan Province (Grant No. 2016JY0240), the Talent Introduction Project of Chengdu University of Information Technology (Grant No. 376226), and the Scientific Research Funding for Young and Middle-Aged Academic Leaders of Chengdu University of Information Technology (Grant No. J201705).

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, Trevor Cohn. 2017. Cross-Lingual Word Embeddings for Low-Resource Language Modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 937–947.

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137-1155.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267-D270.

Billy Chiu, Gamal Crichton, Anna Korhonen, Sampo Pyysalo. 2016. How to Train Good Word Embeddings for Biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing (BioNLP 2016)*. Association for Computational Linguistics, pages 166-174.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, Jimeng Sun. 2016a. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*. PMLR 56, pages 301-318.

Youngduck Choi, Chill Yi-I Chiu, David Sontag. 2016b. Learning Low-Dimensional Representa-

tions of Medical Concepts. In *Proceedings of the AMIA Summit on Clinical Research Informatics (CRI)*. American Medical Informatics Association.

Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, Peter Bruza. 2014. Medical Semantic Similarity with a Neural Language Model. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. pages 1819-1822.

Samuel G. Finlayson, Paea LePendu, Nigam H. Shah. 2014. Building the Graph of Medicine from Millions of Clinical Narratives. *Scientific Data*, 1: 140032.

Edward Grefenstette, Phil Blunsom, Nando de Freitas, Karl Moritz Hermann. 2014. A Deep Architecture for Semantic Parsing. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*. Association for Computational Linguistics, pages 22–27.

Siwei Lai, Kang Liu, Shizhu He, Jun Zhao. 2016. How to Generate a Good Word Embedding? *IEEE Intelligent Systems*, 31(6): 5-14.

Omer Levy, Yoav Goldberg, Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211-225.

Yun Liu, Collin M. Stultz, John V. Guttag, Kun-Ta Chuang, Fu-Wen Liang, Huey-Jen Su. 2016. Transferring Knowledge from Text to Predict Disease Onset. In Proceedings of the 1st Machine Learning for Healthcare Conference. PMLR 56, pages 150-163.

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv*:1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S Corrado, Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems 26*. pages 19-27.

Jose Antonio Miñarro Giménez, Oscar Marin-Alonso, Matthias Samwald. 2013. Exploring the application of deep learning techniques on medical text corpora. *Studies in health technology and informatics*, 205:584-588.

Riccardo Miotto, Li Li, Brian Kidd, Joel T. Dudley. 2016. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6:26094.

Sunil Mohan, Nicolas Fiorini, Sun Kim, Zhiyong Lu. 2017. Deep Learning for Biomedical Information Retrieval: Learning Textual Relevance from Click Logs. In *Proceedings of the 16th Workshop on Biomedical Natural Language Processing (BioNLP 2017)*. Association for Computational Linguistics, pages 222-231.

Amber Stubbs, Christopher Kotfila, Ozlem Uzuner. 2015. Annotating Longitudinal Clinical Narratives for De-identification: The 2014 i2b2/UTHealth Corpus. *Journal of Biomedical Informatics*, 58:S20-S29.

Martin Sundermeyer, Ralf Schlüter, Hermann Ney. 2012. LSTM Neural Networks for Language Modeling. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012)*. ISCA, pages 194-197.

Lifu Tu, Kevin Gimpel, Karen Livescu. 2017. Learning to Embed Words in Context for Syntactic Tasks. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, pages 265–275

Stephan Tulkens, Simon Suster, Walter Daelemans. 2016. Using Distributed Representations to Disambiguate Biomedical and Clinical Concepts. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing (BioNLP 2016)*. Association for Computational Linguistics, pages 77-82.

Yi Yang, Jacob Eisenstein. 2016. Part-of-Speech Tagging for Historical English. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 1318–1328.

Shaodian Zhang, Tian Kang, Xingting Zhang, Dong Wen, Noémie Elhadad, Jianbo Lei. 2016. Speculation Detection for Chinese Clinical Notes: Impacts of Word Segmentation and Embedding Models. *Journal of Biomedical Informatics*, 60:334-341.