Clinical Abbreviation Disambiguation Using Neural Word Embeddings

Yonghui Wu, Jun Xu, Yaoyun Zhang, Hua Xu
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston TX, USA
{Yonghui.wu, Jun.Xu, Yaoyun.Zhang, Hua.Xu}@uth.tmc.edu

Abstract

This study examined the use of neural word embeddings for clinical abbreviation disambiguation, a special case of word sense disambiguation (WSD). We investigated three different methods for deriving word embeddings from a large unlabeled clinical corpus: one existing method called Surrounding based embedding feature (SBE), and two newly developed methods: Left-Right surrounding based embedding feature (LR_SBE) and MAX surrounding based embedding feature (MAX_SBE). We then added these word embeddings as additional features to a Support Vector Machines (SVM) based WSD system. Evaluation using the clinical abbreviation datasets from both the Vanderbilt University and the University of Minnesota showed that neural word embedding features improved the performance of the SVM-based clinical abbreviation disambiguation system. More specifically, the new MAX_SBE method outperformed the other two methods and achieved the state-of-the-art performance on both clinical abbreviation datasets.

1 Introduction

Abbreviations are frequently used in clinical notes and often represent important clinical concepts such as diseases and procedures. However, it is still challenging to handle clinical abbreviations. In a previous study (Wu et al., 2012), we examined three widely used clinical Natural Language Processing (NLP) systems and found that all of them have limited capability to accurately identify clinical abbreviations, especially for ambiguous abbreviations (abbreviations with multiple senses, e.g., "pt" can represent "patient" or "physical therapy"). The prevalence of ambiguous clinical abbreviations is very high. A study (Liu et al., 2001b) examining the abbreviations in the Unified Medical Language System (UMLS) reported that 33.1% of them have more than one sense. In reality, the ambiguity problem of clinical abbreviations could be even higher, as existing knowledge bases (e.g., the UMLS) have low coverage of abbreviations' senses (around 38% to 50%) (Xu, Stetson, et al., 2007).

Clinical abbreviation disambiguation is a particular case of the Word Sense Disambiguation (WSD), which is to "computationally determine which sense of a word is activated by its context" (Navigli, 2009). WSD has been extensively studied in the field of NLP (Lee and Ng, 2002). Researchers have developed different WSD methods including knowledge-based methods (Ponzetto and Navigli, 2010), supervised machine learning methods (Brown et al., 1991) and unsupervised machine learning based methods (Chasin et al., 2014; Yarowsky, 1995) for general English text. As the intrinsic linguistic essentials shared in between, researchers have applied similar methods to biomedical literature and clinical text (Schuemie et al., 2005). For example, researchers have conducted studies to disambiguate important entities in biomedical literature, such as gene names. (Xu, Fan, et al., 2007) Much work has been done for disambiguation of abbreviations in clinical text (Moon et al., 2013; S. Moon et al., 2012; Pakhomov et al., 2005; Wu, Denny, et al., 2013; Xu et al., 2012). Various types of WSD approaches have been proposed for clinical abbreviations, including traditional supervised machine learning based approaches with optimized features (Joshi et al., 2006; Moon et al., 2013; S. Moon et al., 2012), vector space model based methods (Pakhomov et al., 2005; Xu et al., 2012), algorithms based on

Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 171–176,
Beijing, China, July 30, 2015.

hyper-dimensional computing (Moon et al., 2013), as well as recent unsupervised methods based on topic-modeling-based approaches (Chasin et al., 2014). Furthermore, there is also a study to recognize and disambiguate abbreviations in real-time when physicians are authorizing the notes (Wu, Denny, et al., 2013).

Among all these methods, supervised machine learning methods often show good performances, when annotated corpora are available (Liu et al., 2004). A few studies have proposed methods to automatically generate "pseudo" training corpus from biomedical/clinical text, by replacing the expanded long forms by their corresponding abbreviations (Liu et al., 2001a) (Pakhomov, 2002). In the recent 2013 Share/CLEF challenge on clinical abbreviation normalization (Suominen et al., 2013), a hybrid system developed by our group, which combines the supervised machine learning method, the profile-based method, as well as existing knowledge bases achieved the best performance (Wu, Tang, et al., 2013).

Over the last few years, there has been increasing interest in training word embeddings from large unlabeled corpora using deep neural networks. Word embedding is typically represented as a dense real-valued low dimensional matrix M of size V×D, where V is the vocabulary size and D is the predefined embedding dimension. Each row of the matrix is associated with a word in the vocabulary, and each column of the matrix represents a latent feature. Several neural network based training algorithms have been proposed. Bengio (Bengio et al., 2003) and Mikolov (Mikolov et al., 2013) proposed algorithms to train word embeddings by maximizing the probability of a word given by the previous word. Collobert (Collobert et al., 2011) proposed a neural network to train word embeddings using ranking lost criteria with negative sampling. The experimental results showed that the ranking based word embeddings derived from the entire English Wikipedia corpus greatly improved a number of NLP tasks in the general English text. Previous studies have found that the neural word embeddings could represent abundant semantic meanings in the real-valued matrix, which could be useful features for different NLP tasks including WSD. In 2014, Li et al. (Li et al., 2014) proposed two methods to derive word embedding features for WSD, including the "TF-IDF based Embedding" (TBE) feature, and the "Surrounding Based Embedding" (SBE) feature. The experimental results on the MSH collection data and the WISE collection data showed that the SBE method achieved better performance. In the biomedical domain, Tang et al. (Tang et al., 2013) used the popular word2vec package to generate word embeddings and showed that the word embedding features improved the F1-score of a baseline NER system by 0.49% (from 70.0% to 70.49%).

Nevertheless, there is no study that investigates the use of neural word embeddings for WSD in the medical domain, i.e., clinical abbreviation disambiguation. In this study, we developed two new word embeddings methods to generate WSD features from a large unlabeled clinical corpus. We compared them with the existing SBE method proposed by Li et al. for disambiguation of clinical abbreviations in two datasets from Vanderbilt University and the University of Minnesota. Our results showed that clinical abbreviation disambiguation could benefit from a much larger unlabeled corpus and our newly developed embedding features outperformed the SBE. To the best of our knowledge, this is the first study using the word embeddings trained from a large unlabeled clinical corpus to improve the performance of clinical abbreviation disambiguation methods.

2 Methods

2.1 Datasets

This study used the annotated abbreviation datasets from the Vanderbilt University Hospital's (VUH) admission notes, as well as the clinical notes from the University of Minnesota-affiliated (UMN) Fairview Health Services in the Twin Cities. The VUH dataset contains 25 abbreviations. For each abbreviation, up to 200 sentences containing the abbreviation were randomly selected and manually annotated by domain experts. The UMN dataset contains 75 abbreviations and 500 sentences were randomly selected and annotated for each abbreviation. Detailed information for the two datasets can be found in (Wu, Denny, et al., 2013) and (Sungrim Moon et al., 2012) respectively. In order to train the neural word embeddings, we utilized the unlabeled clinical notes from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II corpus (Saeed et al., 2011). The MIMIC II corpus is composed of 403,871 notes from four different note types, including discharge, radiology, ECG and ECHO. Table 1 shows the detailed information about the three datasets.

Dataset	#ABBR	#Sense	Size
VUH	25	103	4,721 sentences
UMN	75	352	37,500 sentences
MIMIC II	N/A	N/A	403,871 notes

Table 1. Statistics of the two abbreviation datasets and the unlabeled clinical corpus

2.2 Supervised machine learning-based WSD method

In this study, we used Support Vector Machines (SVMs), which is a supervised machine learning algorithm that has achieved state-of-the-art performances on a number of WSD datasets. (Cabezas et al., 2001; Hui et al., 2004; Lee and Ng, 2002) We used the implementation of SVMs in the libsvm package [a]. The details of the SVM-based WSD system can be found in our previous study (Wu, Denny, et al., 2013).

2.3 Conventional features

Previous research has identified a number of useful features for WSD. (Wu, Denny, et al., 2013) In this study, we constructed a baseline SVM-based WSD classifier by including the following proven features for clinical abbreviation disambiguation:

1). Word features - words within a window of the target abbreviation. We used the Snowball Stemmer from the python NLTK (Natural Language Toolkit) package to stem the words;

2). Word feature with direction - The relative direction (left side or right side) of stemmed words in feature set 1 towards the target abbreviation;

3). Position feature - The distance between the feature word and the target abbreviation;

4). Word formation features from the abbreviation itself - include: a) special characters such as "-" and "."; b) features derived from the different combination of numbers and letters; c) the number of uppercase letters.

2.4 Word embedding features

This study proposed two new strategies of deriving distributed WSD features from neural word embeddings, including the "MAX" surrounding based embedding features (MAX_SBE) and the Left-Right surrounding based embedding features (LR_SBE). In addition, we compared the two proposed embedding features with the best embedding features reported by Li et al. in 2014 – the surrounding based embedding (SBE) feature.

Surrounding based embedding feature (SBE)

Li et al. proposed the SBE feature, in 2014. The SBE feature for a target word was derived by aggregating the embedding row vectors of the surrounding words within a predefined window size (***k***), as shown in Equation 1.

$$SBE(w) = \sum_{i=j-k}^{j+k} Emb(S(i)) \qquad (1)$$

Where ***w*** is the target word to disambiguate, ***j*** is the index of ***w***, ***S*** is the sentence containing ***w***, ***S(i)*** is the word indexed by position ***i*** in sentence ***S***, and ***k*** is the predefined window size. Previous study from Li et al. showed that the SBE feature achieved the best performance in general English domain.

Left-Right surrounding based embedding feature (LR_SBE)

The LR_SBE is a variation of SBE. Instead of summing up over all of the surrounding word, the LR-SBE composed of the left-side SBE – the SBE from the left-side surrounding words, and the right side SBE – the SBE from the right-side surround words. Previous research has shown that the performance of WSD can be improved by considering the relative word feature with directions (left side or right side). Thus, we assumed that the direction information could help the word embedding feature as well. Equation 2 and 3 show the calculation of LR-SBE embedding features.

$$SBE_{Right}(w) = \sum_{i=j+1}^{j+k} Emb(S(i)) \qquad (2)$$

$$SBE_{Left}(w) = \sum_{i=j-k}^{j-1} Emb(S(i)) \qquad (3)$$

MAX surrounding based embedding feature (MAX_SBE)

The MAX-SBE feature is generated by taking the MAX score of each embedding dimension over all the surrounding words. As each column of the embedding matrix represents a latent feature, the surrounding words that have a high association with a particular semantic meaning are more likely to have a higher score in a particular latent feature. The intuition of MAX_SBE is that the high-score latent features are more important to describe the word semantics. It is more likely that the WSD performance can be improved by

[a] http://www.csie.ntu.edu.tw/~cjlin/papers/ libsvm.pdf

keeping those high-score latent features over all the surrounding words. Equation 4 shows the calculation of MAX_SBE feature, where $\boldsymbol{Emb}_j$ denotes the $\boldsymbol{j}_{th}$ dimension of the embedding matrix.

$$MAX\_SBE(w)_j = MAX\{Emb_j(S(i))\}$$
$$w.r.t.\ j-k \le i \le j+k, S(i) \ne w \quad (4)$$

3 Experiments and evaluation

We implemented the neural network based word embedding algorithm from Collobert et al. (Collobert et al., 2011) and trained the word embedding matrix on the unlabeled MIMIC II corpus. We used the suggested parameters to train the neural network with a hidden layer size of 300, a fixed learning rate of 0.01, and an embedding dimension of 50.

For each abbreviation in a dataset, we trained an SVMs model using the conventional features as the baseline, where the model parameters and the window size were optimized by 10-fold cross validation. To reduce the parameter tuning effort, we select a set of unified model parameters for all the abbreviations. To assess the effect of word embedding features, we added each type of word embedding features (SBE, LR_SBE, or MAX_SBE) to the conventional features and then re-trained the SVM classifier using the optimized parameters. We then reported the (Macro) average accuracy across all abbreviations in either the VUH dataset or the UMN dataset based on the results from 10-fold cross validation.

4 Results

Dataset	Features	Average Accuracy (%)
VUH	Baseline (SVMs)	92.19
	+SBE	92.70
	+LR_SBE	92.86
	+MAX_SBE	**93.01**
UMN	Baseline (SVM)	94.97
	+SBE	95.36
	+LR_SBE	95.46
	+MAX_SBE	**95.79**

Table 2. Average accuracy of the WSD systems using different word embedding features on both VUH and UMN datasets

According to 10-fold cross validation, we set the optimized window size of 3 for both datasets. Table 2 shows the macro average accuracy of using different embedding features on the VUH and the UMN abbreviation datasets. The baseline system (SVMs classifier using conventional features) achieved an accuracy of 92.19% and an accuracy of 94.97% on the VUH and the UMN dataset, respectively. The baseline performance on the VUH dataset is lower than that in the UMN dataset. All three types of embedding features (SBE, LR_SBE, and MAX_SBE) improved the average accuracy when compared with the baseline system, with improvements of 0.51%, 0.67, 0.82% for the VUH dataset and 0.39%, 0.49% and 0.82% for the UMN dataset, for SBE, LR_SBE, and MAX_SBE, respectively. We used Wilcoxon test to compare the embedding features. The test results show that the best embedding features in this study (MAX_SBE) outperformed the SBE feature with a significant p-value of 0.004 on the VUH dataset and 7.05e-05 on the UMN dataset.

5 Discussion

This study demonstrates that the word embedding features derived from a large unlabeled corpus could remarkably improve the performance of the SVM-based clinical abbreviation disambiguation system. To the best of our knowledge, this is the first study that investigates the use of neural word embeddings for WSD in clinical text. The most relevant work is a study by Li et al. (Li et al., 2014), where they utilized the algorithm implemented in word2vec to derive embedding features for WSD on a biomedical literature dataset (MSH collection) and a general English dataset (Science WISE dataset). However, the unlabeled dataset used for training the word embedding was relatively small (7,741 abstracts in the MSH dataset and 2,943 abstracts in the WISE dataset), and the proposed WSD method was to directly calculate the cosine similarity. In this study, we proposed two new embedding features and explored a much larger unlabeled clinical corpus (403,871 notes). Our evaluation showed that the proposed LR_SBE feature and the MAX_SBE feature outperformed the SBE feature by Li et al. Among them, the MAX_SBE embedding feature achieved the best average accuracy on both the VUH and UMN datasets, indicating the potential of this new embedding algorithm in WSD tasks.

In fact, all word embedding features improved the performance of the baseline WSD system that uses conventional features only, indicating the usefulness of neural word embeddings in WSD tasks. The LR_SBE feature outperformed the SBE feature, denoting that it is helpful to consider the relative directions even for the real-valued word embedding features. This is consistent with the findings reported in the supervised machine learning based WSD methods using linguistic features. The MAX_SBE feature outperformed the other two types of embedding features, suggesting that the major dimension of the embedding matrix is more powerful for describing semantic meanings. The MAX_SBE word feature is related to the work from Collobert et al., where they designed a MAX convolutional layer in their deep neural network to weight and select the major dimensions among the context words. Our research shows that simply taking the major dimensions from the embedding matrix of context words works well for clinical abbreviation disambiguation.

The neural word embeddings could represent abundant semantic meanings and capture multi-aspect relations from unlabeled corpora, which may generate novel, useful features for various NLP tasks, as demonstrated in the open domain. (Collobert et al., 2011; Li et al., 2014) This study demonstrates its usefulness for clinical abbreviation disambiguation. In addition to WSD, we believe such word embedding features can benefit other NLP tasks in the medical domain.

This study has limitations. The evaluation datasets are composed of the frequently used abbreviations that have enough training samples. For example, the UMN dataset is a balanced dataset that there are exactly 500 samples for each of the abbreviations. We only used the embedding features from the surrounding words, where some semantically important words out off the window were missed. Similar to the study of capturing long distance conventional features, e.g., the syntactic feature, there are possible approaches that can capture long distance features from embedding matrix. Le et al. (Le and Mikolov, 2014) proposed a distributed representation of sentence and documents, which could be a potential solution. In the future, we plan to investigate different approaches that can capture the sentence level distributed representation feature and paragraph level distributed representation feature. We will also examine the word embedding features using deep neural network based classifiers.

6 Conclusion

This paper examined the neural word embedding features for the disambiguation of clinical abbreviations. We proposed two novel word embedding features and compared them with an existing word embedding feature in an SVM-based WSD classifier. Evaluation using two clinical abbreviation datasets showed that all word embedding features derived from a large unlabeled corpus could improve WSD performance, with MAX_SBE achieving the best performance.

7 Acknowledgement

This study was supported by grants from the NLM 2R01LM010681-05, NIGMS 1R01GM103859 and 1R01GM102282. We would like to thank the University of Minnesota and the 2014 SemEval challenge organizers for the development of corpora used in this study.

Reference

Bengio, Y., Ducharme, R., et al. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137-1155.

Brown, P. F., Pietra, S. A. D., et al. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*. pp.264-270. Berkeley, California.

Cabezas, C., Resnik, P., et al. 2001. Supervised sense tagging using support vector machines. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*. pp.59-62. Toulouse, France.

Chasin, R., Rumshisky, A., et al. 2014. Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods. *J Am Med Inform Assoc*, 21(5):842-849.

Collobert, R., Weston, J., et al. 2011. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.*, 12:2493-2537.

Hui, H., Giles, L., et al. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on* (pp.296-305).

Joshi, M., Pakhomov, S., et al. 2006. A comparative study of supervised learning as applied to acronym expansion in clinical reports. *AMIA Annu Symp Proc*:399-403.

Le, Q., and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of The 31st International Conference on Machine Learning* (pp.1188-1196).

Lee, Y. K., and Ng, H. T. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*. pp.41-48.

Li, C., Ji, L., et al. (2014). Acronym Disambiguation Using Word Embedding. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Liu, H., Lussier, Y. A., et al. 2001a. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *J Biomed Inform*, 34(4):249-261.

Liu, H., Lussier, Y. A., et al. 2001b. A study of abbreviations in the UMLS. *Proc AMIA Symp*:393-397.

Liu, H., Teller, V., et al. 2004. A multi-aspect comparison study of supervised word sense disambiguation. *J Am Med Inform Assoc*, 11(4):320-331.

Mikolov, T., Chen, K., et al. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moon, S., Berster, B., et al. 2013. Word sense disambiguation of clinical abbreviations with hyperdimensional computing. In *AMIA Annu Symp Proc*.

Moon, S., Pakhomov, S., et al. 2012. *Clinical Abbreviation Sense Inventory*. http://purl.umn.edu/137703

Moon, S., Pakhomov, S., et al. 2012. Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations. *AMIA Annu Symp Proc*, 2012:1310-1319.

Navigli, R. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):1-69.

Pakhomov, S. 2002. Semi-supervised Maximum Entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. pp.160-167. Philadelphia, Pennsylvania.

Pakhomov, S., Pedersen, T., et al. 2005. Abbreviation and acronym disambiguation in clinical discourse. *AMIA Annu Symp Proc*:589-593.

Ponzetto, S. P., and Navigli, R. 2010. Knowledge-rich Word Sense Disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. pp.1522-1531. Uppsala, Sweden.

Saeed, M., Villarroel, M., et al. 2011. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. *Crit Care Med*, 39(5):952-960.

Schuemie, M. J., Kors, J. A., et al. 2005. Word sense disambiguation in the biomedical domain: an overview. *J Comput Biol*, 12(5):554-565.

Suominen, H., Salanterä, S., et al. (2013). Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In P. Forner, H. Müller, et al. (Eds.), *Information Access Evaluation. Multilinguality, Multimodality, and Visualization* (Vol. 8138, pp.212-231): Springer Berlin Heidelberg.

Tang, B., Cao, H., et al. 2013. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. *BMC Medical Informatics and Decision Making*, 13(Suppl 1):S1.

Wu, Y., Denny, J. C., et al. (2013). A prototype application for real-time recognition and disambiguation of clinical abbreviations. In *Proceedings of the 7th international workshop on Data and text mining in biomedical informatics* (pp.7-8). San Francisco, California, USA: ACM.

Wu, Y., Denny, J. C., et al. 2012. A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries. *AMIA Annu Symp Proc*, 2012:997-1003.

Wu, Y., Tang, B., et al. (2013). Clinical Acronym/Abbreviation Normalization using a Hybrid Approach. In *Proceedings of CLEF 2013*.

Xu, H., Fan, J. W., et al. 2007. Gene symbol disambiguation using knowledge-based profiles. *Bioinformatics*, 23(8):1015-1022.

Xu, H., Stetson, P. D., et al. 2007. A study of abbreviations in clinical notes. *AMIA Annu Symp Proc*:821-825.

Xu, H., Stetson, P. D., et al. 2012. Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. *AMIA Annu Symp Proc*, 2012:1004-1013.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. pp.189-196. Cambridge, Massachusetts.