Acronym Disambiguation Using Word Embedding

Chao Li[1*], **Lei Ji**[2], **Jun Yan**[2]

[1] Dalian University of Technology, Dalian, China, [2] Microsoft Research, Beijing, China
oahci1.dlut@gmail.com, leiji@microsoft.com, junyan@microsoft.com

Abstract

According to the website AcronymFinder.com which is one of the world's largest and most comprehensive dictionaries of acronyms, an average of 37 new human-edited acronym definitions are added every day. There are 379,918 acronyms with 4,766,899 definitions on that site up to now, and each acronym has 12.5 definitions on average. It is a very important research topic to identify what exactly an acronym means in a given context for document comprehension as well as for document retrieval. In this paper, we propose two word embedding based models for acronym disambiguation. Word embedding is to represent words in a continuous and multidimensional vector space, so that it is easy to calculate the semantic similarity between words by calculating the vector distance. We evaluate the models on MSH Dataset and ScienceWISE Dataset, and both models outperform the state-of-art methods on accuracy. The experimental results show that word embedding helps to improve acronym disambiguation.

Introduction

Ambiguous acronyms in the documents are difficult for readers to understand the meanings. For example, the acronym *CMU* refers to universities including *Carnegie Mellon University* and *Central Michigan University*. When talking about *employment rate of CMU*, it is difficult to know what the exact meaning of *CMU* is without the expansion. So it is important to disambiguate these acronyms.

Acronym disambiguation is a subset of the more general problem of Word Sense Disambiguation (WSD), which is to decide the sense of words in context. Many works have been proposed to address this problem in clinical and biomedical domain (Pakhomov, Pedersen and Chute 2005; McInnes et al. 2011) using Unified Medical Language System (UMLS) (Bodenreider 2004), which is a knowledge-base providing semantic information and ontological relationships that can serve as features for machine learning.

In general domain, instead of using the domain-specific knowledge-bases like UMLS, many works tend to consider representing the disambiguating context using syntactic features or bag-of-words (HaCohen-Kerner, Kass and Peretz 2008). Additionally, topic model is also introduced (Zhang et al. 2011).

In this paper, we propose two word embedding based models to address this problem. Word embedding (Mikolov et al. 2013) is an unsupervised learning of word representation. It aims to map a word into a multidimensional space vector with the semantic information, so that similar words will have high similarity in the embedding space even if the string similarity is low.

This paper makes the following contributions:

- Introducing word embedding to address the acronym disambiguation problem
- Proposing two word embedding based models based on (Mikolov et al. 2013)

Models

TF-IDF Based Embedding (TBE)

TBE is based on the assumption that the accurate expansion is highly concerned with the topic of the given context, and the topic can be represented by top TF-IDF words. TF-IDF weight is commonly used correlated with Vector Space Model (VSM) to represent a word in a vector space. Motivated by this, we use top TF-IDF words' embeddings to represent the topic information of the given context.

Based on (Mikolov et al. 2013), we add an output layer in the architecture. For an acronym $\boldsymbol{A}$, the embedding of *TBE(**A**)* is calculated as:

$$TBE(\boldsymbol{A}) = \sum_k \sum_{i=1}^{n} TF\text{-}IDF(w(i), k) * e(w(i)), \quad (1)$$

where k stands for every document containing $\boldsymbol{A}$ (in test data there is only one for we do not know the exact meaning of the acronym), w contains words in the context in decreasing order of TF-IDF weights and e contains the embeddings of all the words in training data trained by (Mikolov et al. 2013). $TF\text{-}IDF(w(i), k)$ is the TF-IDF

*This work was done during Chao Li's internship in Microsoft Research.

weight for word $w(i)$ in document k. We choose top n TF-IDF words in the model.

Surrounding Based Embedding (SBE)

In (Mikolov et al. 2013), sum of embeddings of the surrounding words is used in the hidden layer to calculate the embedding of the certain word. Motivated by this, we propose SBE to represent the word by adding its surrounding words' embeddings, which can enrich the semantic information.

In SBE, we also add an output layer based on (Mikolov et al. 2013). For an acronym $\boldsymbol{A}$, the embedding of $SBE(\boldsymbol{A})$ is calculated as:

$$SBE(\boldsymbol{A}) = \sum_k \sum_{j=k-i}^{k+i} e(w'(j)), j \neq k, \quad (2)$$

where k stands for every position that acronym $\boldsymbol{A}$ appears (maybe not in one single document in training data), i is the word window, w'contains words in the document in the original order and e is the same as that in TBE.

Experiment

Datasets & Experiment Setting

We evaluate our models on MSH Collection and ScienceWISE Collection shared by (Prokofyev et al. 2013) which are both collections of scientific abstracts containing ambiguous acronyms. We apply our models on the same training and test data as (Prokofyev et al. 2013). The MSH dataset contains 7,641 abstracts in training data and 3,631 abstracts in test data, while the ScienceWISE dataset contains 2,943 abstracts in training data and 2,267 abstracts in test data. We do the following pre-treatments:

- For both training and test datasets, we remove all the marks and stop-words
- For training dataset, we replace all the acronym and expansions with the "Acronym+ID," in which ID is the index of the meaning to distinguish different meanings
- For test dataset, we replace all the expansion with the acronym

Then, we apply our models on the datasets to get the embeddings of acronyms (in TBE we use top 20 TF-IDF words, in SBE the word window is 3). We disambiguate the acronyms in test data by calculating cosine similarity to choose the most similar one from the given candidates.

Experimental Results

We compare our models with the results of (Prokofyev et al. 2013), which also use ontologies as knowledge-bases. Table 1 shows the results including the baselines. The reason why the baseline Context Vectors performs so well in MSH dataset can be explained by the relatively low quality of its background ontology which has been automatically constructed while in ScienceWISE the ontology is manually built, as mentioned in (Prokofyev et al. 2013). The results show that our models perform high accuracy than that of (Prokofyev et al. 2013).

Table 1: Precision for models on the two datasets

Approach	MSH	ScienceWISE
Random (Baseline)	46.73	39.37
Most Frequent (Baseline)	43.60	74.46
Context Vectors (Baseline)	**95.29**	74.29
NB (Baseline)	67.31	85.13
Binary CCV (Prokofyev et al. 2013)	90.77	93.34
Binary CCV+NN (with Cat) (Prokofyev et al. 2013)	90.60	94.53
TBE(ours)	90.98	91.56
SBE(ours)	**93.10**	**94.86**

Conclusion

The experimental results show that using word embedding can improve the accuracy of acronym disambiguation without any knowledge-bases, and the results are also very stable in datasets from different domains. We will focus on using word embedding to address the general WSD problem in the future.

References

Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology[J]. Nucleic acids research, 2004, 32(suppl 1): D267-D270.

HaCohen-Kerner Y, Kass A, Peretz A. Combined one sense disambiguation of abbreviations[C]//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Association for Computational Linguistics, 2008: 61-64.

McInnes B T, Pedersen T, Liu Y, et al. Using second-order vectors in a knowledge-based method for acronym disambiguation[C]//Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2011: 145-153.

Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. arXiv preprint arXiv:1310.4546, 2013.

Pakhomov S, Pedersen T, Chute C G. Abbreviation and acronym disambiguation in clinical discourse[C]//AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2005, 2005: 589.

Prokofyev R, Demartini G, Boyarsky A, et al. Ontology-Based word sense disambiguation for scientific literature[M]//Advances in Information Retrieval. Springer Berlin Heidelberg, 2013: 594-605.

Zhang W, Sim Y C, Su J, et al. Entity Linking with Effective Acronym Expansion, Instance Selection, and Topic Modeling[C]//IJCAI. 2011, 2011: 1909-1914.