Semantic Representations of Medical Terms: A Comparison Study of Word Embeddings Trained on Clinical Notes and PubMed Articles

Conference Paper · November 2017

CITATIONS
0

READS
66

5 authors, including:

Yanshan Wang
Mayo Clinic
82 PUBLICATIONS **241** CITATIONS
SEE PROFILE

Naveed Afzal
Wolters Kluwer Health
52 PUBLICATIONS **163** CITATIONS
SEE PROFILE

Liwei Wang
Mayo Clinic
66 PUBLICATIONS **196** CITATIONS
SEE PROFILE

Semantic Representations of Medical Terms: A Comparison Study of Word Embeddings Trained on Clinical Notes and PubMed Articles

Yanshan Wang, PhD[1], Naveed Afzal, PhD[1], Liwei Wang, PhD[1], Feichen Shen, PhD[1], Hongfang Liu, PhD[1]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Introduction

With the availability of large quantities of health data, deep learning has become more and more prevalent. The word embeddings learned by deep neural networks have seen tremendous success in various natural language processing (NLP) applications. Word embeddings trained on common domain data have been shown to carry syntactic and semantic meanings of words. Training word embeddings usually requires an adequately large corpus. In medical domain, clinical notes and biomedical articles are two resources that could serve as training data. In this study, we answer the question: Do the word embeddings trained by these two resources perform differently on representing semantics of medical terms?

Methods

In our studies, we utilized the word2vec model, skip-gram, to evaluate the word embeddings. Word2vec is based on the assumption that words occurring in similar contexts tend to have similar meanings. Based on a training corpus, it uses a deep neural network to embed words into a continuous vector space. In our implementation, we set the minimum number of times a word must appear in the corpus to 7, the window size to 5 words, and the number of dimensions in feature space to 60.

Two corpora were utilized to train the word embeddings: EHR Corpus and PubMed Corpus. EHR Corpus contains clinical notes for a cohort of 113k patients receiving their primary care at Mayo Clinic, spanning a period of 15 years (1998-2013). PubMed Corpus contains 1.25 million articles from PubMed Central. Minimum pre-processing was conducted including punctuation removal, digits replacement, etc. We also compared the trained word embeddings with two public pre-trained embeddings: Google News and Wikipedia.

Four datasets were utilized to evaluate the word embeddings for capturing medical term semantics. Dataset 1 contains 30 medical term pairs with manually generated similarity scores; Dataset 2 contains 34 medical term pairs with similarity scores; Data 3 consists of 101 clinical term pairs whose similarity was determined by physicians from the Mayo Clinic; Data 4 consists of 725 clinical term pairs whose similarity was determined by residents from the University of Minnesota Medical School.

On each dataset, the word embeddings were evaluated based on their Pearson correlation coefficient with the similarity scores determined by human experts.

Results

The results of Pearson correlation coefficient on four datasets are listed in Table 1. Overall, the similarity results using word embeddings trained on PubMed Corpus are the closest to human experts' results. However, there is no statistical significance between the word embeddings trained on EHR Corpus and those on PubMed Corpus. Both word embeddings are significantly superior to the pre-trained word embeddings on Wikipedia and Google News.

Table 1. Pearson correlation coefficient of four datasets.

Dataset	EHR Corpus	PubMed Corpus	Wikipedia	Google News
1	0.542	**0.569**	0.334	0.357
2	**0.416**	0.311	0.159	0.243
3	0.295	**0.300**	0.001	0.084
4	0.374	**0.404**	0.190	0.154

Conclusion

There is no statistical significance between the word embeddings trained on EHR Corpus and PubMed Corpus in regards of semantic representations of medical terms. However, the word embeddings trained on medical domain data are superior to those trained on common domain data.