Advanced Science and Technology Letters
Vol.120 (AIA 2015), pp.793-797
http://dx.doi.org/10.14257/astl.2015.120.156

Wore Representation Analysis of Bio-marker and Disease word

Young-Shin Youn[1,2] , Kyung-Min Nam[1,2], Hye-Jeong Song[1,2], Jong-Dae Kim[1,2], Chan-Young Park[1,2], Yu-Seop Kim[1,2 1]

[1]Department of Convergence Software, Hallym University, Korea
[2]Bio-IT Research Center, Hallym University, Korea
pour657@gmail.com, jkre4030@naver.com,
{hjsong, kimjd, cypark, yskim01}@hallym.ac.kr

Abstract. One of the most important processes in a machine learning-based natural language processing module is to represent words by inputting the module. This can be accomplished by representing words in one-hot form with a large vector size without applying the concept of semantic similarity between words, or by word representation (embedding) with vectors to represent lexical similarity. In this study, classification performance of three word representation models (Word2Vec, canonical correlation analysis, and GloVe) is tested on a corpus that established using the abstracts of 204,674 biomedical articles published in PubMed. Categories include disease name, disease symptom, and ovarian cancer marker. The classification performance of each word representation model for each category is visualized by mapping the results in two-dimensional word representations using t-SNE.

Keywords: disease word, word representation, bio-marker.

1 Introduction

A crucial process in a machine learning-based natural language processing module is representing words by inputting the module. Most related studies have used the one-hot form to represent words that has two main problems: the vector size is too large and the concept of semantic similarity between words is absent.

Word representation processing using vectors to indicate lexical similarity between words has recently attracted considerable attention by improving the performance of machine learning models of natural language processing.

In word representation, unlike one-hot encoding, machine learning occurs at the level of lexical representation reduced to a k dimension. word representation methods have enabled performance upgrades in some natural language processing models, such as syntactic parsing and sentiment analysis (also known as opinion mining).

The more recent study in [3] used Word2Vec and GloVe in the biomedical domain to verify the similarity of word pairs and thus prove the efficiency of word representation.

[1] He is a corresponding author

ISSN: 2287-1233 ASTL

This study extends the result of [3] and verifies the classification performance of word representation in the biomedical domain. To achieve this end, we use a canonical correlation analysis (CCA) [7] model in conjunction with Word2Vec [4, 5] and GloVe [6] models described in [3].

We build a corpus using the abstracts of PubMed articles in the biomedical domain; classify their contents into the categories of disease, symptom, and bio-marker (ovarian cancer markers); and test the classification performance of each word representation model.

2 Word Representation

Word representation or distributed representation is a technology for learning vector representations for all words contained in a given corpus. Among the several word representation models, Word2Vec, CCA, and GloVe models are used in this study.

2.1 Word2Vec

Word2Vec has two main model options: continuous bag-of-words (CBOW) or skip-gram [4, 5]. CBOW model are the neighboring words of the target word. Skip-gram is predicting neighboring words or context for an input word. In other words, skip-gram is a word representation model useful for predicting sentences or neighboring words. This study uses the skip-gram model as a Word2Vec model.

2.2 CCA (Canonical Correlation Analysis)

In a CCA [7], two projection vectors that maximize the correlation are sought. Let x be the word representation in the random variable ($X, Y \in R$), and let y be the context representation associated with that word. Then, k-dimensional projection vector that maximally correlates the two variables is sought.

2.3 GloVe

GloVe [6] refers to a global vector. It is a hybrid-type word representation that considers both global and local contexts of words. Its dot product of W_x and W_y for vocabulary training is proportional to the co-occurrence count. We use a freely accessible GloVe open source tool.

3 Data

In this study, a corpus was built using titles and abstracts of 204,674 PubMed biomedical studies. The corpus was classified into the categories of disease name, disease symptom, and bio-marker [Table 1].

Table 1. Disease, Symptom and bio-marker categories

Disease name	Disease symptom	Bio-marker
Pneumonia	Dizziness	CA125
Glaucoma	Anemia	EGFR
Urethritis	Candidiasis	apoA-I
Gastritis	Osteoporosis	apoC-III
Meningitis	Hypoproteinemia	CRP
Cystitis	Thrombocytopenia	Cortisol
Pneumothorax	Leukopenia	MIF
Leukemia	Arteriosclerosis	Leptin
Adenocarcinoma	Brucellosis	IL-6
Cancer	Septicemia	IL-8
Tumor	Mycoses	Prolactin
Tuberculosis	Candidiasis	OPN
…	…	…

4 Test

We tested the classification performance of word representation of the Word2Vec, CCA, and GloVe models using the categories in the biomedical domain established in Section 3. Fig. 1 present the results of word representation using the Word2Vec skip-gram model, CCA, and GloVe, respectively. Word representations in blue, purple, and green represent disease names, disease symptoms, and ovarian cancer markers, respectively.

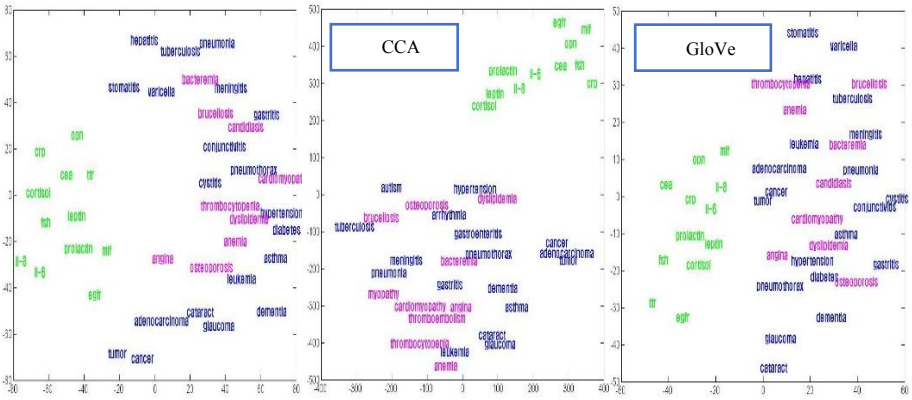

Fig. 1. Result of Word2Vec, CCA, GloVe Model's Word Representation.

From these results, we verified that all three word representation models can distinguish disease names and symptoms from ovarian cancer markers, with related words closely clustered.

5 Conclusion

In this study, a corpus was created using abstracts of biomedical articles published in PubMed. We tested the classification performance of three word representation models based on disease, symptom, and bio-marker (ovarian cancer marker) categories. Experimental results shown in Section 4 demonstrate that all three models could distinguish diseases from ovarian cancer markers and that related diseases were clustered by means of word representation.

We plan to extend this study to identifying new bio-markers for specific diseases by using larger datasets.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (2015R1A2A2A01007333) and by the Ministry of Education, Science and Technology (2010-0010612).

References

1. Collobert, Ronan, et al.: Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 12, (2011)

2. Turian, Joseph, Lev Ratinov, and Yoshua Bengio.: Word representations: a simple and general method for semi-supervised learning. Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, (2010)
3. Muneeb, T. H., Sunil Kumar Sahu, and Ashish Anand.: Evaluating distributed word representations for capturing semantics of biomedical concepts. pp.158, ACL-IJCNLP 2015, (2015)
4. Mikolov, Tomas, et al.: Efficient estimation of word representations in vector space. arXiv preprint arXiv (2013).
5. Mikolov, Tomas, et al.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. (2013)
6. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning.: Glove: Global vectors for word representation. pp.1532-1543, Proceedings of the Empirical Methods in Natural Language Processing (2014)
7. Stratos, Karl, Michael Collins, and Daniel Hsu: Model-based word embeddings from decompositions of count matrices. pp.1282 – 1291, 2015 Proceedings of the Annual Meeting of the Association for Computational Linguistics (2015).

