International Journal of Bio-Science and Bio-Technology
Vol.8, No.1 (2016), pp.295-302
http://dx.doi.org/10.14257/ijbsbt.2016.8.1.26

Classification Performance of Bio-Marker and Disease Word using Word Representation Models[1]

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] and Yu-Seop Kim[1,2][2]

[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 processesin 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 (word embedding) with vectors to represent lexical similarity. This has attracted keen research interest by improving the performance of several natural language processing modelssuch as syntactic parsing and sentiment analysis (also known as opinion mining). In this study, classification performance of Word2Vec, canonical correlation analysis (CCA), and GloVeare tested on a corpus that established using the titles and abstractsof 204,674biomedical articles published in PubMed. Categories include disease name, disease symptom, and ovarian cancer marker.Ovarian cancer markers were used as bio-markers.The classification performance of each word representation model for each category is visualized by mapping the results in two-dimensional word representations using t-distributed stochastic neighbor embedding (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. In this method, the word concerned is assigned a value of 1, and all remaining words are assigned a value of 0, with the vector size equal to the vocabulary size [1]. This method 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 [2-11].

In word representation, unlike one-hot encoding, machine learning occurs at the level of lexical representation reduced to a k dimension. Word representations are generated by means of artificial neural networks or explicit representations of contextual words. And Word representations is helpful in many learning algorithm of NLP task, such as machine translation and voice recognition [12]. Such word representation methods have enabled

[1]This paper is a revised and expanded version of a paper entitled [Wore Representation Analysis of Bio-marker and Disease Word] presented at [The 4th International Conference on Artificial Intelligence and Application, Jeju National University International Center, Jeju Island, Korea and December 16-19, 2015].
[2] He is a corresponding author

ISSN: 2233-7849 IJBSBT

performance upgrades in some natural language processing models, such as syntactic parsing and sentiment analysis (also known as opinion mining) [13].

For example, the study in [14] improved the performance of named entity recognition (NER) by using word representation for a conditional random field (CRF) feature. The more recent study in [15] used Word2Vec and GloVe in the biomedical domain to verify the similarity of word pairs and thus prove the efficiency of word representation.

This study extends the result of [16]and verifies the classification performance of word representation in the biomedical domain. To achieve this end, we use a canonical correlation analysis (CCA) [17] model in conjunction with Word2Vec [17, 18] and GloVe [19] models described in [15].

The word representation models used for the experiment in this study can mine both syntactic and semantic properties in the biomedical domain. 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; and test the classification performance of each word representation model. We use ovarian cancer markers as bio-markers. [20]

To test classification performance, the results of k-dimensional word representation are visualized through two-dimensional mapping, using the method of t-distributed stochastic neighbor embedding (t-SNE).

In Section 2, we examine word representation models used in our study. Data used for those models are explained in Section 3. Section 4 presents the experimental methods and processes based on those data, and Chapter 5 reviews experimental results and discusses future research directions.

2. Word Representation

Word representation or distributed representation is a technology for learning vector representations for all words contained in a given corpus.Most related studies have used the one-hot form to represent words. In this method, the word concerned is assigned a value of 1, and all remaining words are assigned a value of 0, with the vector size equal to the vocabulary size. In contrast to the one-hot form, which cannot capture semantic similarity between words, word representation offers vocabulary training as a k-dimensional representation. And they can capture various dimensions of meanings and phrase information relevant to the potential features of words within the vector. Among the several word representation models, Word2Vec, CCA, andGloVemodels are used in this study.

2.1. Word2Vec

Word2Vec has two main model options: continuous bag-of-words (CBOW) or skip-gram [16, 17]. The CBOW model predicts a word based on its neighboring words. Therefore, inputs of the CBOW model are the neighboring words of the target word.Figure 1 presents a flow of the skip-gram for predicting neighboring words or context for an input word (w(t)).

In other words, skip-gram is a word representation model useful for predicting sentences or neighboring words. So, Skip-gram model is a useful word representation in prediction neighboring words in a sentence. Both CBOW and skip-gram models are neural network-based language models in which a big corpus is established using input words. In addition, the word representation of each word contained in the corpus is learned. This study uses the skip-gram model as a Word2Vec model.

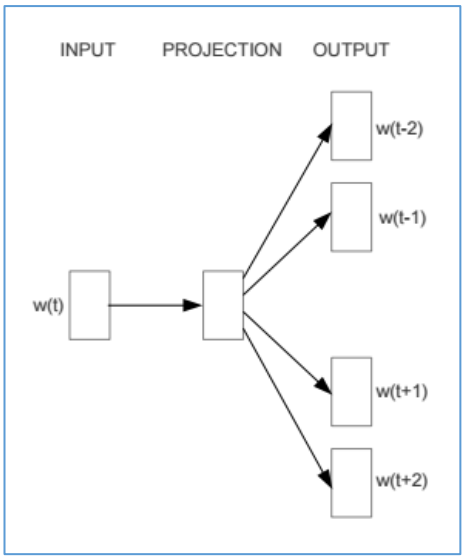

Figure 1. Skip-Gram [17]

2.2. CCA

In a CCA [15], two projection vectors that maximize the correlation are sought. It can be calculated directly from the data set, or calculated as representations similar to covariance matrices. It can be used for dimensional reduction and provides the correlation between two sets at the d dimension and the relevant projection vector. 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. For simplicity, that these variables have zero mean. Then, k-dimensional projection vector that maximally correlates the two variables is sought.

2.3. GloVe

GloVe [18] refers to a global vector.GloVe is an unsupervised learning algorithm to obtain vector representations of words. It is a hybrid-type word representation that considers both global and local contexts of words. The GloVe model learns items that are not 0 in the global and local matrix.Its dot product of w_x and w_y for vocabulary training is proportional to the co-occurrence count that show how often the words. 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. Ovarian cancer markers were used as bio-markers, and for disease symptom, data randomly extracted from the benchmark data of [14] were used. The corpus thus established was used as input data for representation analysis.

Table 1. Disease, Symptom, and Bio-Marker Categories

Disease name	Disease symptom	Bio-marker
Pneumonia	Dizziness	CA125
Cataract	Cardiomyopathy	CA19-9
Glaucoma	Anemia	EGFR

Urethritis	Candidiasis	Myoglobin
Gastritis	Brucellosis	Tenascin-C
Meningitis	Osteoporosis	apoA-I
Conjunctivitis	Hypoproteinemia	apoC-III
Cystitis	Angina	CRP
Stomatitis	Thrombocytopenia	FSH
Pneumothorax	Prostatism	Cortisol
Asthma	Cardiomyopathy	TTR
Leukemia	Leukopenia	CA15-3
Adenocarcinoma	Arteriosclerosis	MIF
Cancer	Bacteremia	Leptin
Tumor	Brucellosis	IL-6
Dementia	Septicemia	CEA
Hepatitis	Mycoses	IL-8
Hypertension	Candidiasis	Prolactin
Diabetes	Prostatism	OPN
Tuberculosis	Dyslipidemia	HE4
Varicella	Brucellosis	MMP-7

4. Test

We tested the classification performance of word representation of the Word2Vec, CCA, and GloVemodels using the categories in the biomedical domain established in Section 3. The results of k-dimensional word representation are visualized through two-dimensional mapping, using t-SNE method.

Figure 2–4 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. The results of three word representation models are shown that all three word representation models can distinguish disease names and symptoms from ovarian cancer markers.

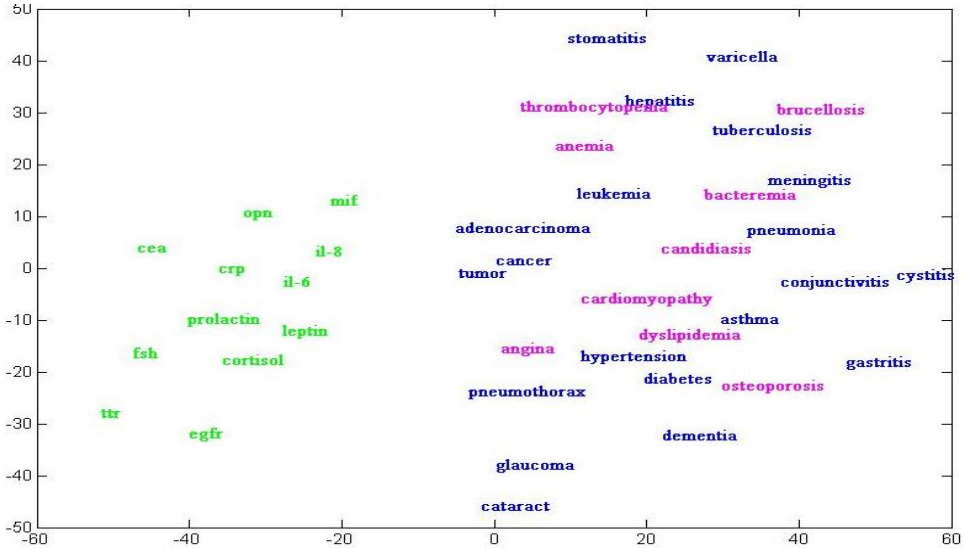

Figure 2. Results of Word2Vec Model's Word Representation

The word representation by the Word2Vec model (skip-gram) effectively distinguishes diseases from markers. Thrombocytopenia-related diseases are leukemia and anemia. Figure 2 shows that the names and symptoms of thesethree diseases (thrombocytopenia, leukemia, and anemia) are clustered. Dyslipidemia, which refers to conditions of increased total cholesterol, LDL cholesterol, triglycerides in blood, and decreased HDLcholesterol are associated with hypertension. Cancer-related words (*e.g.*, cancer, tumor, and adenocarcinoma occurring in gastrointestinal and bronchial mucosa as well as in the pancreas or excretory duct) are also closely clustered. Such clustering of semantically related words demonstrates that Word2Vec is efficient for semantic relation classification.

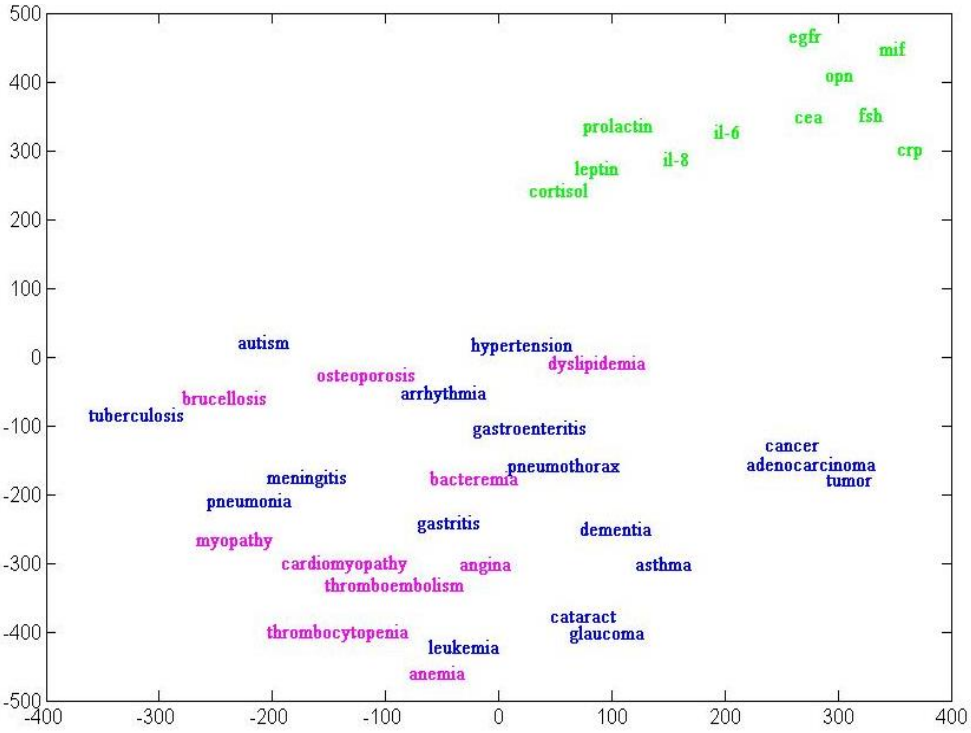

Figure 3. Result of CCA Model's Word Representation

As shown in Figure 3, the CCA model distinguishes diseases from the ovarian cancer markers just as well as does the Word2Vec model. As in Figure 2, Figure 3 also shows that dyslipidemia and hypertension associated with it are clustered, as are thrombocytopenia and anemia. It performed better than the Word2Vec model with respect to cancer-related words by clustering cancer, tumor, and adenocarcinoma more closely together. Furthermore, in contrast to the Word2Vec and GloVe models, it additionally clustered the inflammatory diseases of meningitis, gastritis, and gastroenteritis, as well as the heart disease-related symptoms angina, thromboembolism, and cardiomyopathy.

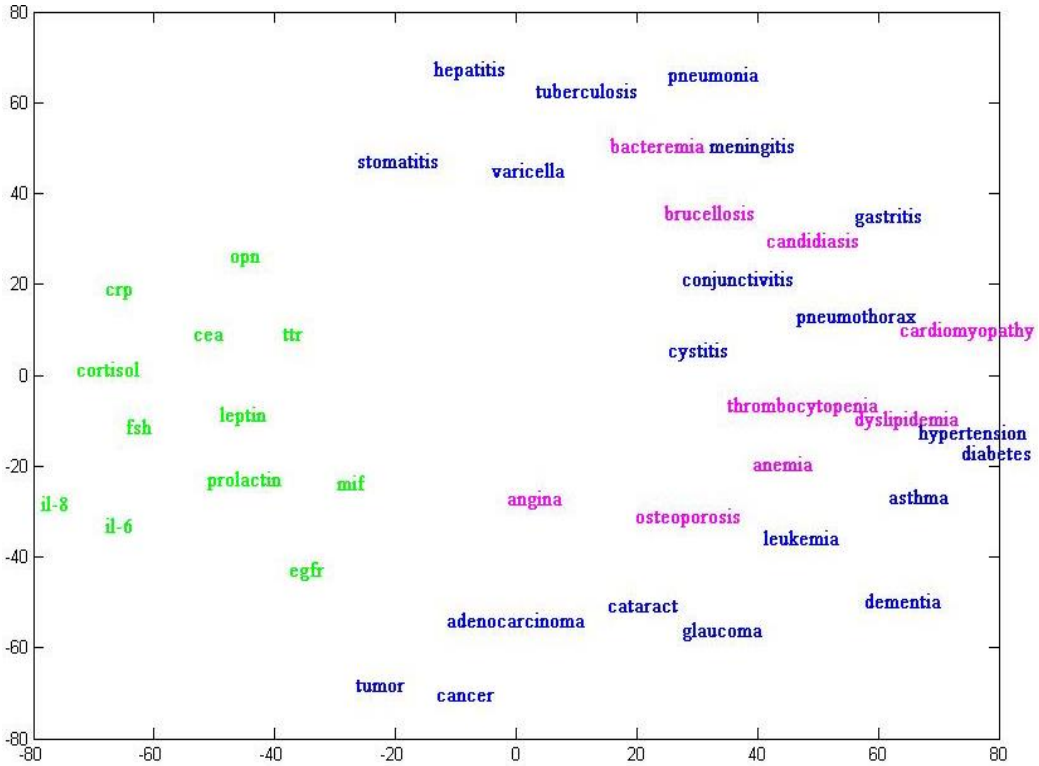

Figure 4. Results of GloVe Model's Word Representation

Figure 4 shows the results of word representation by means of the GloVe model. As in Figure 3, the three cancer-related words are placed fairly close to one another. In addition, eye-related diseases glaucoma and cataractare closely clustered.

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. Of the three models,the Thrombocytopenia-related word, Dyslipidemia-related word, Cancer-related word were closely clustered. The CCA model distinguished inflammatory diseases (meningitis, gastritis, and gastroenteritis) and heart disease and symptoms (angina, thromboembolism, cardiomyopathy), unlike the Word2Vec and GloVe models.

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. The results of k-dimensional word representation were visualized by means of a two-dimensional mapping using t-SNE. 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 and verify that all three word representation models can distinguish disease names and symptoms from ovarian cancer markers.

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

Acknowledgements

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)

References

[1] R. Collobert, J. Weston, L. Bottou, M. Karien, K. Kavukcuoglu and P. Kuksa, "Natural language processing (almost) from scratch", The Journal of Machine Learning Research, vol. 12, **(2011).**
[2] K. Young-Bum, B. Snyder and R. Sarikaya, "Part-of-speech Taggers for Low-resource Languages using CCA Features", Empirical Methods in Natural Language Processing (EMNLP). ACL–Association for Computational Linguistics.
[3] T. Mikolov, W. T. Yih and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations", In HLT-NAACL, **(2013)**, pp. 746-751.
[4] D. Zhou, D. Zhong and Y. He, "Event trigger identification for biomedical events extraction using domain knowledge", Bioinformatics, vol. 30.11, **(2014)**, pp. 1587-1594.
[5] D. Bollegala, T. Maehara, Y. Yoshida and K. I. Kawarabayashi, "Learning Word Representations from Relational Graphs", arXiv preprint arXiv:1412.2378, **(2014)**.
[6] D. Bollegala, T. Maehara, Y. Yoshida and K. I. Kawarabayashi, "Learning sentiment-specific word embedding for twitter sentiment classification", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 1, **(2014)**.
[7] X. Zheng, H. Chen and T. Xu, "Deep Learning for Chinese Word Segmentation and POS Tagging", EMNLP, **(2013)**.
[8] E. H. Huang, R. Socher, C. D. Manning and A. Y. Ng, "Improving word representations via global context and multiple word prototypes", Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, **(2012)**.
[9] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents", arXiv preprint arXiv:1405.4053, **(2014)**.
[10] G. Zhou, T. He, J. Zhao and P. Hu, "Learning continuous word embedding with metadata for question retrieval in community question answering", Proceedings of ACL, **(2015)**, pp. 250–259.
[11] Y. Luan, S. Watanabe and B. Harsham, "A. Efficient learning for spoken language understanding tasks with word embedding based pre-training", In Sixteenth Annual Conference of the International Speech Communication Association, **(2015)**.
[12] L. Qiu, Y. Cao, Z. Nie and Y. Rui, Editors, "Learning Word Representation Considering Proximity and Ambiguity", Twenty-Eighth AAAI Conference on Artificial Intelligence, **(2014).**
[13] S. Li and Y. Jiang, "Semo-supervised Sentiment Classification using Ranked Opinion Words", International Journal of Database Theory & Application, vol. 6.6, **(2013)**, pp. 50-62.
[14] J. Turian, L. Ratinov and Y. 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)**.
[15] T. H. Muneeb, S. K. Sahu and A. Anand, "Evaluating distributed word representations for capturing semantics of biomedical concepts", ACL-IJCNLP, **(2015)**, pp. 158.
[16] K. Stratos, M. Collins and D. Hsu, "Model-based word embeddings from decompositions of count matrices", Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), **(2015)**, pp. 1282 – 1291.
[17] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector

space", arXiv preprint arXiv:1301.3781, **(2013)**.
[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality", Advances in neural information processing systems, **(2013)**.
[19] J. Pennington, R. Socher and D. Cristopher, "Manning. Glove: Global vectors for word representation", Proceedings of the Empirical Methods in Natural Language Processing **(2014)**, pp. 1532-1543.
[20] Y.-S. Youn, K.-M. Nam, H.-J. Song, J.-D. Kim, C.-Y. Park and Y.-S. Kim, "Word Representation Analysis of Bio-marker and Disease word", The 4th International Conference on Artificial Intelligence and Application, **(2015)**.

Authors

Young-Shin Youn, was born in 1993. Department of Convergence Software, Hallym University, 1, Hallym-daeghak-gil, ChunCheon-si, Gangwon-do, Korea.

Kyung-Min Nam, was born in 1991. Department of Convergence Software, Hallym University, 1, Hallym-daeghak-gil, ChunCheon-si, Gangwon-do, Korea.

Hye-Jeong Song, was born in 1966.She received the Ph.D. Degree in Computer Engineering from Seoul National University. She is a Professor in Department of Convergence Software, Hallym University. Her recent interests focus on biomedical system and bioinformatics. Department of Convergence Software, Hallym University, 1, Hallym-daehak-gil, Chuncheon, Gangwon-do, Korea.

Jong-Dae Kim, was born in 1959. He received the M.S. degree and Ph.D. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology, Seoul, Korea, in 1984 and 1990, respectively. He worked for Samsung Electronics from 1988 to 2000 as an electrical engineer. He is a Professor in Department of Convergence Software, Hallym University. His recent interests focus on biomedical system and bioinformatics. Department of Convergence Software, Hallym University, 1, Hallym-daehak-gil, Chuncheon, Gangwon-do, Korea.

Chan-Young Park, was born in 1964. He received the B.S. and the M.S. from Seoul National University and the Ph.D. degree from Korea Advanced Institute of Science and Technology in 1995. From 1991 to 1999, he worked at Samsung Electronics. He is currently a Professor in the Department of Convergence Software, Hallym University. His research interests are in Bio-IT convergence, Intelligent Transportation System and sensor networks. Department of Convergence Software, Hallym University, 1, Hallym-daehak-gil, Chuncheon, Gangwon-do, Korea.

Yu-Seop Kim, was born in 1969. He received his Ph.D. degree fromSeoul National University in 2000. He has been a professor of Hallym University from 2002. His research area is Natural Language Processing, Machine Learning, BioIT Convergence Software, and so on. Departmentof Convergence Software, Hallym University, 1, Hallym-daehak-gil, Chuncheon, Gangwon-do, Korea.

