Correlation Analysis of Chronic Obstructive Pulmonary Disease (COPD) and its Biomarkers Using Word Embeddings

Byeong-Hun Yoon and Yu-Seop Kim

Hallym University, South Korea
yqudgns1222@gmail.com and yskim01@hallym.ac.kr

Abstract

It is very costly and time consuming to find new biomarkers for specific diseases in clinical laboratories. In this study, to find new biomarkers most closely related to Chronic Obstructive Pulmonary Disease (COPD), which is widely known as respiratory disease, biomarkers known to be associated with respiratory diseases and COPD itself were converted into word embedding. And their similarities were measured. We used Word2Vec, Canonical Correlation Analysis (CCA), and Global Vector (GloVe) for word embedding. In order to replace the clinical evaluation, the titles and abstracts of papers retrieved from Google Scholars were analyzed and quantified to estimate the performance of the word embedding models.

1 Introduction

Chronic Obstructive Pulmonary Disease (COPD) is the fourth leading cause of mortality in the world and the seventh in Korea. It is one of the most respiratory diseases in the elderly, and is known to be a very dangerous disease. Similar to asthma, COPD has symptoms of airway disorders such as dyspnea, cough, and sputum. COPD mainly exacerbates pulmonary function, leading to death. (Vestbo et al., 2013)

Accordingly, research on bio-markers, which serve as an index for predicting the disease and detecting changes in the body, is underway (Aronson et al., 2005). A biomarker is an index that can detect changes in the body using DNA, RNA, metabolites, proteins, and protein fragments. Therefore, many researchers are working hard to find new biomarkers that are deeply related to specific diseases. Recently there have been several attempts to extract such information from documents (Poon et al., 2014) (Poon et al., 2015). (Youn et al., 2016) tried to find biomarkers related to ovarian cancer by using word embeddings. The research is the base step of this paper.

In this study, 26 respiratory-related biomarkers recommended by Chuncheon Kangwon National University Hospital[1] were selected for the first time in order to search for bio-markers related to COPD. We also extracted 800,000 titles and abstracts from Pubmed[2] to construct word embedding. Canonical Correlation Analysis (CCA) (Stratos et al., 2015), Word2vec (Mikolov et al., 2013), and Global Vector (GloVe) (Pennington et al., 2014) were used as word embedding models. With these models, word-embeddings of COPD and respiratory bio-makers are acquired. The word embeddings of COPD and the bio-markers are mapped in two dimensions using t-SNE (t-Distributed Stochastic Neighbor Embedding) (Maaten et al., 2008) and the result is visualized. The relationship between COPD and biomarkers are examined by measuring and comparing the similarity of COPD with biomarkers using cosine similarity.

Clinical trials should follow to validate the methodology presented in this study. However, since this is very difficult in practice, we intend to test this indirectly by analyzing the results of Google Scholars[3]. In other words we analyzed the Google Scholars' search results and compare the pairs with high scores to those with low scores to verify the validity of the proposed methodology.

This paper is composed as follows. Section 2 explains the biomarkers considered in this study. Section 3 explains the models used for word embedding in this study. Section 4 explains the over-

[1] https://www.knuh.or.kr/eng/main/index.asp
[2] https://www.ncbi.nlm.nih.gov/pubmed/
[3] https://scholar.google.co.kr/

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 337–342,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

all flow of this study. Section 5 discusses the experiment and its results. Finally, Chapter 6 discusses conclusions and future research.

2 Bio-markers for COPD

A biomarker is an index that can detect changes in the body using DNA, RNA, metabolites, proteins, and protein fragments. In addition, biomarkers can be used to effectively diagnose various diseases.

The biomarkers listed in table 1 are known to be related to respiratory diseases and are included in the 26 markers recommended by Kangwon National University Hospital.

Table 1 : Bio-marker 26

Bio-marker	SP-D, CC-16, IP-10, IL-2, Eotaxin-1, Leptin Adiponectin, Ghrelin, PAI-1, IL-10, LDL, PON-1 SAA, C9, IGFBP-2, CD105, NF-kB, NSE CYFRA21-1, CEA, TRAIL, DR5, Angiostatin, Endostatin, Calprotectin, rbp

Surfactant, pulmonary-associated protein D, also known as SFTPD or SP-D (Lahti et al., 2008), is a lung-related protein. In addition, SP-D plays an important role in lung immunity and is known to regulate the function of many immune cells.

Clara Cell Secretory Protein (CC-16) (Broeckaert and Bernard, 2000) is known to be a protein distributed in the endocardium and the respiratory bronchus of the lungs and has immunomodulatory and anti-inflammatory effects.

Interferon gamma-induced protein 10 (IP-10) (Dufour et al., 2002) is also known to be CXCL10 or B10. It is involved in the Th1 immune response and is known to be increased in infectious diseases including inflammation of the respiratory tract, immune disorders, and tumors.

Interleukin 2 (IL-2) (Koreth et al., 2011) is an immunoreactive substance involved in anti-inflammatory immune responses, macrophage function to damaged cells, and restoration of the original state. IL-2 plays a major role in the immune system, and if production is reduced, immune defense can be seriously compromised.

In this study, 22 respiratory markers were also recommended and analyzed for their association with COPD.

3 Word-embeddings

Word-embedding is a technique that learns the vector representation of every word in a given Corpus. We can measure the similarity between several words, and perform vector computation with vectorized semantics to enable additional inference. In this paper, we investigate the relationship between COPD and its bio-markers using the following word-embedding models: CCA, Word2vec, and GloVe.

3.1 CCA

CCA (Hotelling and Harold, 1936) is a technique known by Hotelling (1936), which is a technique for examining the correlation of variables. CCA is a statistical method used to investigate the relationship between two words, and it is a technique that simultaneously examines the correlation between variables in a set and variables in another set. In other words, it is a useful tool to grasp the correlation of variable group (X, Y) and to grasp the relationship between two features (Jang et al., 2013).

3.2 Word2vec

Word2vec is a Google-released model in 2013. Word2vec has the premise that words with the same context have close meanings. It is also most commonly used to understand sentences in text. There are CBOW (Continuous Bag of Words) and Skip-grams in the learning method of the word2vec model. The CBOW method predicts the target word using the context that the surrounding word makes. On the other hand, skip-gram (Mikolov et al., 2013) predicts words that can come around in one word. In addition, Skip-grams are known to be more accurate in large datasets. Therefore, we use Skip-gram method in this paper.

3.3 GloVe

GloVe (Pennington et al., 2014) stands for Global Vector, and it is a method of expressing a word as a vector using a non-local learning algorithm. In addition, it is a hybrid model that considers not only the global context but also the local context of the word.

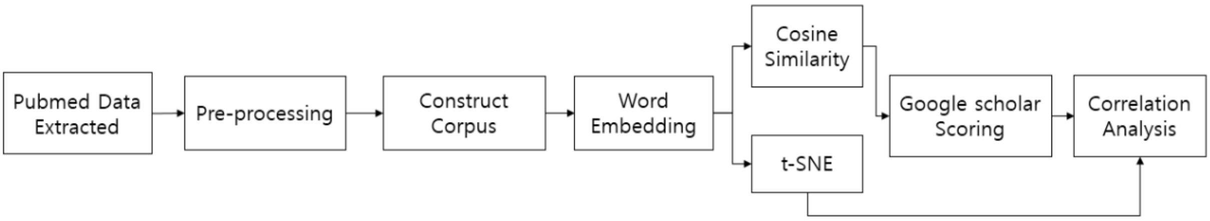

Figure 1 The whole process of proposed research

4 Methodology

In this paper, about 800,000 papers are downloaded from the PubMed site, and word embeddings are applied to only the title and abstract part of each paper. Since there are various expressions for each biomarker in the bio documents, in the preprocessing process, called normalization, the corpus is newly constructed by replacing these various forms into a single form. To do this, various forms of markers, including COPD, must be constructed in dictionary form first. Three word embedding models are used for this corpus to construct two-, five-, ten-, and hundred-dimension lexical vectors. Among them, 100-dimensional vectors are reduced to two dimensions using t-SNE and the result is mapped to a two-dimensional graph. The t-SNE is used to directly look up biomarkers closely related to COPD through the visualization process. Also, the original vectors that are not reduced are also analyzed to calculate the similarity with COPD. By calculating the similarity between the original vectors, we try to more precisely find the related markers.

The process of analyzing the relationship with the last stage, Google Scholars, is for analyzing the validity of the methodology rather than analyzing the relationship between COPD and biomarkers. Figure 1 shows this whole process.

5 Experiment

In this paper, we construct new corpus with nxml documents downloaded from PubMed site. As a result of extracting the nxml document, a total of 807,821 papers were extracted, and the titles and abstracts were collected separately and used as a corpus for word embedding. More than two million words are represented by word embedding. Table 2 shows the most similar words to COPD.

Table 2 The most similar words to '*COPD*.'

Rank	token	similarity
1	exacerbation	0.851586342
2	spirometry	0.819786787
3	obstructive	0.812012613
4	asthma	0.803823590
5	ipf	0.798491836
6	aecopd	0.783644378
7	bronchodilators	0.779349267
8	asthmatics	0.774952054
9	hyperinflation	0.752177000
10	BHR	0.750948250

A word *'exacerbation'* may refer to an increase in the severity of a disease or its signs and symptoms[4]. *'Spirometry'* is the most common of the pulmonary function tests[5]. And doctors may classify lung conditions as *'obstructive'* lung disease or restrictive lung disease[6].

Among two millions word vectors, biomarkers related to pulmonary diseases such as metabolic syndrome and lung cancer are extracted separately to reveal their relationship with COPD. The markers listed in table 1 are recommended by Kangwon National University Hospital.

Twenty six biomarkers are embedded in four cases such as 2-dimensional, 5-dimensional, 10-dimensional, and 100-dimensional. In the case of a high-dimensional vector, the value of a certain element is excessively large, which may interfere with accurate similarity calculation. To solve this problem, a 100-dimensional vector value is reduced to two dimensions using t-SNE and the result is mapped to a two-dimensional space. [Figure 2] shows the results of visualizing two-dimensional mapping of 100-dimensional vectors

[4] http://www.medicinenet.com
[5] https://en.wikipedia.org/wiki/Spirometry
[6] http://www.webmd.com/lung/obstructive-and-restrictive-lung-disease#1

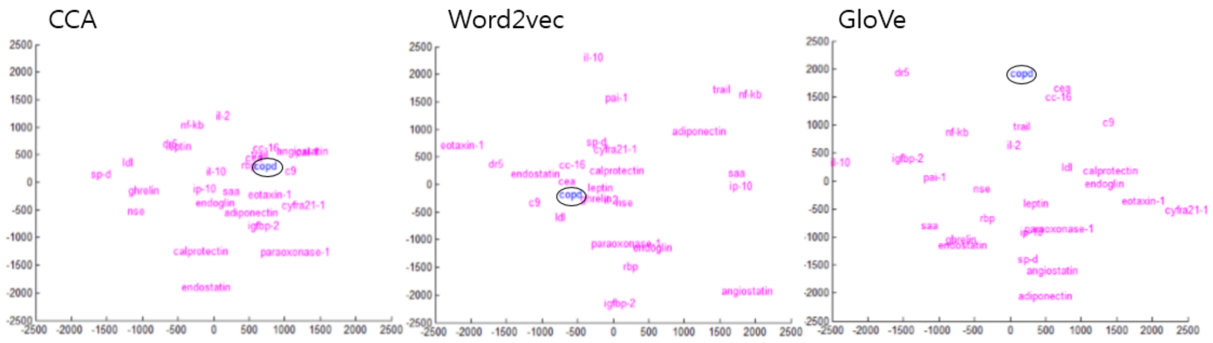

Figure 2 Mapping results of COPD and biomarkers in two-dimensional space.

embedded in CCA, word2vec, and GloVe. In figure 2, all but the COPD are biomarkers.

As shown in figure 2, in the case of CCA and Word2vec, bio-markers are distributed evenly around COPD. However, in the case of GloVe, COPD is found to be far apart. However, in all three models, bio-markers located close to COPD are CC-16 and CEA. In addition, we can see that rbp, ghrelin, and trail are close to COPD on the t-SNE.

Then, the degree of similarity is calculated by applying the cosine similarity to the two-, five-, ten- or 100-dimensional vectors embedded in three ways. Here, it is assumed that the markers with high similarity values are more closely related to COPD. Clinical trials must be conducted to verify the validity of this assumption. However, in reality, clinical trials are very difficult, so this study tries indirect verification through Google Scholars.

Table 3 shows the results for all experiments with 5 markers of the highest similarities and 5 markers of the lowest similarities through Google Scholars. For an indirect evaluation via Google Scholars, we first attempt to search for the keyword "COPD marker_name" pairs. The titles and abstracts of the top 20 papers presented in the search results were analyzed and quantified. Where title_both is the average number of articles having both keywords in the titles and abs_both is the average number of articles having both keywords in the abstracts.

Table 3 shows how well each algorithm discriminates between good markers and bad mark-

Table 3. The average number of articles having both COPD and the biomarker. 'BEST' means the most similar pairs and 'WORST' means the least similar pairs.

algorithm	dimension	BEST		WORST	
		title_ both	abs_ both	title_ both	abs_ both
CCA	2	3.8	7.6	5.2	7.4
	5	1.4	5.4	5.4	7.4
	10	0.4	3.6	3.2	6
	100	1.4	3.8	2.2	5.2
	t-SNE	0.2	2	3.6	7.2
Word2vec	2	4.4	7.8	3.4	4.6
	5	5.4	8.8	1.2	1.8
	10	6.8	9.8	3.2	5.6
	100	4.6	8	1	1.4
	t-SNE	2.6	6.2	3.8	7.2
GloVe	2	3.8	7.4	5.8	10.2
	5	0.6	4.2	4.6	8
	10	4	8.4	3.4	5.4
	100	6	10.8	1.2	2
	t-SNE	3.4	6.6	5	8.2

Table 4. Biomarkers with the highest similarities with COPD

Word2Vec				
biomarker	Similarity	title_ both	abs_ both	words
cc-16	0.57279	11	15	1
eotaxin-1	0.48002	1	3	4
sp-d	0.45937	7	8	2
cyfra21-1	0.40179	0	4	8
ip-10	0.35346	4	10	5
average		4.6	8	4
GloVe				
biomarker	Similarity	title_ both	abs_ both	words
adiponectin	0.15436	11	17	1
cea	0.14091	1	8	1
pai-1	0.11407	1	5	1
saa	0.10816	4	7	1
leptin	0.10062	13	17	1
average		6	10.8	1

ers. In other words, the larger the difference between the values of BEST and WORST, the more favorable it is. As a result, the BEST marker values of CCA were lower than those of the WORST markers. This shows that the CCA does not play a significant role in this problem. On the other hand, Word2vec is a very stable methodology because all the BEST marker values show higher values than the WORST cases, unless the dimension is reduced by t-SNE. However, GloVe showed a stable appearance as the number of dimensions increased, while it didn't in case of 2- and 5-dimension. In the 100 dimension, it showed a bigger difference than word2vec.

Table 4 shows the similarities with the markers analyzed as having the highest correlation values with COPD in Wor2Vec 100 dimension and GloVe 100 dimension. It shows that cc-16 recommended by Word2Vec and adiponectin and leptin recommended by GloVe have already undergone much research on COPD. On the other hand, Word2Vec's eotaxin-1 cyfra21-1 and GloVe's cea and saa have not. This provides a direction for new clinical studies. In other words, among the markers with a high degree of similarity, the markers that are searched with a low frequency in the Google Scholar will be subject to various clinical studies in the future. And if you expand it further, it will help you to find new markers for specific diseases.

In fact, it has been found that cyfra21-1, which has been shown to have a high similarity, not having much research interest so far, to COPD, was found to have a significant correlation with the phenotype of COPD in clinical trial. Currently, Kangwon National University Hospital is conducting experiments to obtain more reliable clinical results.

Table 5. Comparison to Human Selection

Rank (Word2vec)	marker	Rank (Human)
1	CC-16	2
2	Eotaxin-1	9
3	SP-D	1
4	Cyfra21-1	13
5	IP-10	13

Rank (Human)	marker	Rank (Word2vec)
1	SP-D	3
2	CC-16	1
3	Leptin	9
4	Ghrelin	21
5	PAI-1	14

Finally, we compare the most closely related markers recommended by word2vec (100 dimension) to the score made by human researchers. Table 5 shows the rankings that markers recommended by Word2Vec have been given by human researchers. For 26 markers, the rank correlation coefficient value between word2vec and human is 0.44.

Three clinical specialists participated in this experiment. These are professors who have been engaged in research for a long time in university hospitals, but the number of specialists involved should be increased. They are all respiratory medicine specialists, but they are not familiar with all biomarkers used in this research. Therefore, the comparison with the specialists should be seen not only from the performance evaluation of this study but also from the perspective of supplementing the specialists.

6 Conclusion

In this paper, we use word embedding to find markers that are closely related to COPD. For word embedding, we used CCA, Word2Vec, and GloVe. Experimental results show that Word2Vec and GloVe have the best performance when they are 100 dimensions.

In the future, based on this research, we seek to find new markers that are closely related to specific diseases. To do this, it is necessary to construct a corpus that summarizes the various forms of expression that a disease or a marker has. Also, it is necessary to develop various processing algorithms for expressions composed of words. In addition, we will conduct further research on the value of similarity itself as well as the relative ranking of biomarkers.

In this paper, the word embedding is performed in a given corpus, and similarity is calculated by fixed embedding. Later, we will express this problem as deep neural network and develop a model that can learn and predict based on this.

Acknowledgments

This work was supported by a National Research Foundation of Korea grant funded by the Ministry of Science and ICT of the Korea Government (No. 2015R1A2A2A01007333) and by Hallym University Research Fund, 2017 (HRF-201704-013).

References

Jeffrey K. Aronson. 2005. Biomarkers and surrogate endpoints. *British journal of clinical pharmacology*, 59(5):491-494.

Bernard Broeckaert. 2000. Clara cell secretory protein (CC16): characteristics and perspectives as lung peripheral biomarker. *Clinical and Experimental Allergy*, 30(4):469–475.

Jennifer H. Dufour, Michelle Dziejman, Michael T. Liu, Josephine H. Leung, Thomas E. Lane, and Andrew D. Luster. 2002. IFN-gamma-inducible protein 10 (IP-10; CXCL10)-deficient mice reveal a role for IP-10 in effector T cell generation and trafficking*, Journal of Immunology*. 168 (7), 3195–3204.

Harold Hotelling. 1936. Relations between two sets of variates, *Biometrika,* 283(3/4): 321-377.

Min-Ki Jang, Yu-Seop Kim, Chan-Young Park, Hye-Jeong Song and Jong-Dae. Kim. 2013. Integration of Menopausal Information into the Multiple Biomarker Diagnosis for Early Diagnosis of Ovarian Cancer. *International Journal of Bio-Science and Bio-Technology*, 5(4):215-222.

John Koreth, Ken-ichi Matsuoka, Haesook T. Kim, Sean M. McDonough, Bhavjot Bindra, Edwin P. Alyea, Philippe Armand, Corey Cutler, Vincent T. Ho, Nathaniel S. Treister, Don C. Bienfang, Sashank Prasad, Dmitrios Tzachanis, Robin M. Joyce, David E. Avigan, Joseph H. Antin, Jerone Ritz, and Robert J. Soiffer. 2011. Interleukin-2 and Regulatory T Cells in Graft-versus-Host Disease. *New England Journal of Medicine,* 365(22):2055-2066.

Meri Lahti, Johan Löfgren, Riita Marttila, Marjo Renko, Tuula Klaavuniemi, Ritva Haataja, Mika Ramet and Mikko Hallman. 2002. Surfactant protein D gene polymorphism associated with severe respiratory syncytial virus infection. *Pediatric research,* 51(6):696-699.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579-2605.

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint* arXiv:1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *In Advances in neural information processing systems (NIPS2013),* 3111-3119.

Jeffrey Pennington, Richard. Socher and Christopher D. Manning. 2014. Glove: Global vectors for word representation. *In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)*, 1532-1543.

Hoifung Poon, Chris Quirk, Charlie DeZiel, and David Heckerman. 2014. Literone: PubMed-scale Genomic Knowledge Base in the Cloud. *Bioinformatics*, 30(19):2840-2842.

Hoifung Poon, Kristina Toutanova, and Chris Quirk. 2015. Distant Supervision for Cancer Pathway Extraction from Text. *In Proceedings of the Pacific Symposium, Biocomputing 2015*, 120-131.

Karl Stratos, Michael Collins, and Daniel Hsu. 2015. Model-based word embeddings from decompositions of count matrices. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2015),* 1282 – 1291.

Jorgen Vestbo, Suzanne S. Hurd, Alvar G. Agustí, Paul W. Jones, Claus Vogelmeier, Antonio Anzueto, Peter J. Barnes, Leonardo M. Fabbri, Fernando J. Martinez, Masaharu Nishimura, Robert A. Stockley, Don D. Sin, and Roberto Rodriguez-Roisin. 2013. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *American journal of respiratory and critical care medicine,* 187 (4):347-365.

Young-Shin Youn, Chan-Young Park, Jong-Dae Kim, Hye-Jeong Song, and Yu-Seop Kim. 2016. Finding a New Bio-Markers of a Specific Disease using Word Embeddings. *In Proceedings of the fifth International Multi-Conference on Engineering and Technology Innovation 2016 (IMETI-2016).*