Extraction of disease-related genes from PubMed paper using word2vec

Takahiro Koiwa
Tokyo University of Science
Japan
7416619@ed.tus.ac.jp

Hayato Ohwada
Tokyo University of Science
Japan

ABSTRACT

Finding disease-related genes is important in drug discovery. Many genes are involved in the disease, and many studies have been conducted and reported for each disease. However, it is very costly to check these one by one. Therefore, machine learning is a suitable method to address this problem. By extracting study results from research papers by text mining, it is possible to make use of that knowledge. In this research, we aim to extract disease-related genes from PubMed papers using word2vec, which is a text mining method. The method extracts the top 10 genes whose known disease genes and vectors are close to those obtained by word2vec. Based on these, genes other than known disease-related genes are extracted and used as disease-related genes. We conducted experiments using schizophrenia, and confirmed the likelihood of this disease-related gene using xgboost. Pattern 1: Only known genes. Pattern 2: Pattern 1 plus disease-related genes extracted in this study. Pattern 3: Pattern 1 plus the same number of random genes. Using these three patterns, we performed a xgboost with microarray data and compared the classification accuracy. The result was that Pattern 2 had the highest accuracy. Therefore, we could extract genes with using genes related to disease by our method.

KEYWORDS

Machine Learning, Text mining, Word2vec

CCS CONCEPTS

• **Computing methodologies → Machine learning approaches**

ACM Reference format:

Takahiro Koiwa and Hayato Ohwada. 2017. Extraction of disease-related genes from PubMed paper using word2vec. *In CSBio '17: 8th International Conference on Computational Systems-Biology and Bioinformatics, December 7–8, 2017, Nha Trang City, Viet Nam*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3156346.3156355

1 INTRODUCTION

The application of machine learning in drug discovery is increasing more and more. Due to the enormous number of genes, it is difficult to find disease-related genes, and it is very expensive in terms of time and expense to conduct each experiment. By discovering regularities such as the expression and structure of genes and compounds using machine learning, important new knowledge useful for drug discovery, such as the discovery of genes related to specific diseases, is obtained. Discovering genes related to certain diseases has a very important meaning in medicine discovery support. This is because of the need for clarifying genes to be targeted when developing drugs and preparing medicines acting on the genes. Therefore, in this study, genes related to disease are extracted using text mining. With the recent development of microarray technology, it has become possible to obtain enormous amounts of genetic data (microarray data). Many studies have been done in this area, and there are many reports of genes related to disease. However, it is costly to manually summarize those reports and judge whether they are correct.

Therefore, despite being reported, many genes that record genes involved in disease are not listed in the database. In this situation, we cannot use the knowledge gained. Therefore, studies on text mining are being promoted for biomedical papers. For example, Al-Mubaid H. et al. performed a syntactic analysis of biomedical articles for extracting proteins related to disease. They extracted relevant proteins for various diseases such as Alzheimer's, dengue fever, and lung cancer. They then extracted six proteins related to Alzheimer's disease [1]. However, there are many genes related to disease, and it is important to extract more genes. Also, due to the development of machine learning in recent years, the field of text mining has greatly advanced and can perform analyses that take into account the meanings of words and sentences. It is also being applied to the analysis of literature in the biomedical field. For example, Shahin Mohammadi et al. proposed a method for extracting similar articles by looking at the similarity of sentences in papers from PubMed [2].

Word2vec is a the text-mining method which makes semantic analyses and is attracting the most attention recently [3][4][5]. This software predicts the words appearing before and after a given word using a neural net. The weighting of the network at that moment becomes a vector of words. This method is superior in terms of understanding the meanings of words that have been considered difficult in the past, and it is possible to extract words having similar

CSBio '17, December 7–8, 2017, Nha Trang City, Viet Nam
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5350-2/17/12…$15.00
https://doi.org/10.1145/3156346.3156355

meanings based on similarities between the vectors of words. Many studies have been conducted to apply this method to the analysis of literature in the biomedical fields. For example, a study to extract drugs and disease relations from a paper has been conducted, and it was reported that it can be extracted appropriately [6][7]. Word2vec has two types of prediction models, skip-gram and CBOW. Skip-gram may have higher performance in analyzing medical theses [8]. In addition, hyperparameter is very important for the performance in machine learning, Chiu et al [9]. Investigated the influence of parameter change and showed that there are different optimum parameters for genetic similarity and gene relevance It was. In this study, we aim to extract unknown genes (genes with possibility of disease association) similar to known disease-related genes, so we adopt optimal parameters for gene similarity.

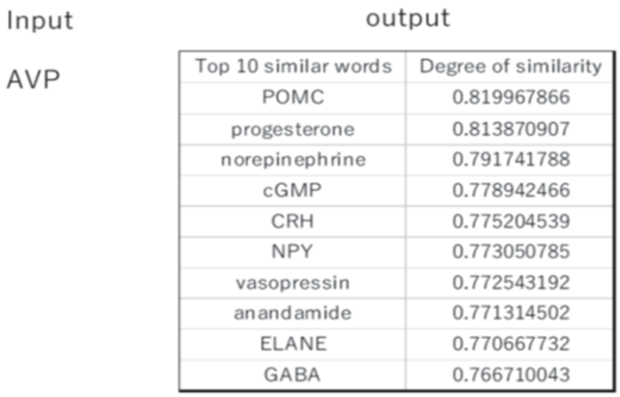
Therefore, in this study, we aim to perform a semantic analysis using word2vec for biomedical papers and extract genes that are not yet systematized from the papers.

2 METHODS

In this chapter, we describe the methods used in this study. This research consists of three steps: a pretreatment method, a synonym extraction method and a related gene extraction method. There will be explained in order below.

2.1 Pretreatment Method

Information on genes include gene symbol and gene name, but in this study, we focus only on gene symbols. In biomedical papers, the same gene may be written with a different Symbol for each paper. We followed the NCBI gene list and converted all genetic symbols in each paper into the official symbols. This prevents synonyms from appearing as similar genes, and it is impossible to calculate a vector with respect to words with very few occurrences. In addition, because the target sentence contains information such as author name and citation, it was deleted to prevent erroneous extraction.



Input
AVP
output

Top 10 similar words	Degree of similarity
POMC	0.819967866
progesterone	0.813870907
norepinephrine	0.791741788
cGMP	0.778942466
CRH	0.775204539
NPY	0.773050785
vasopressin	0.772543192
anandamide	0.771314502
ELANE	0.770667732
GABA	0.766710043

Figure 1: Word2vec output example.

2.2 Synonym Extraction Method

We extracted 10 genes with high similarity using word2vec with one of the disease- related genes listed in the database as input. This is taken as a candidate word. This is done for all words in the database, and 10 candidate words are extracted for each gene. Fig.1 presents an example of the output when the known disease-related gene AVP is input. Ten words with high cosine similarity as determined by Word2vec are output.

2.3 Related Gene

Upon first examination of the candidate word, there are many known genes (genes contained in the database) and words other than genes. Therefore, we first extract only gene names. Using the gene list of NCBI, we extract only the words matching the official symbol. Next, genes not matching the genes contained in the database are extracted. These are potentially disease-related genes that could be extracted by this study. An example is summarized in Table 1. In the example, "ELANE" is extracted as a disease-related gene.

Table 1: Examples extracted as disease-related genes

Top 10 similar words	Gene Symbol or not	known related gene or not
POMC	Gene Symbol	Known
progesterone	Not	Not
norepinephrine	Not	Not
cGMP	Not	Not
CRH	Gene Symbol	Known
NPY	Gene Symbol	Known
vasopressin	Gene Symbol	Known
anandamide	Not	Not
ELANE	**Gene Symbol**	**Not**
GABA	Not	Not

3 EXPERIMENT

In this chapter, we describe the data used in this research, the parameter settings of word2vec and the verification of the method. In this study, we extract the disease-related genes of schizophrenia.

3.1 Data

First, the text is only the title and abstract of all the articles that can be searched by the search word "schizophrenia" in PubMed of NCBI [10]. Since there are some papers are not posted, there are also articles with only titles. The number of articles retrieved was 120,904.

Next, we searched two large-scale datasets, the GENE database [11] and the DisGeNet database [12], for known disease-related genes. There were 1928 related genes of "schizophrenia" in GENE and 1833 in DisGeNet, the overlap between the two databases was 951. Therefore, there were 2810 known disease-related genes. For the list of genes, we used the gene name list of NCBI. This list contains OrganismName, NCBIGeneID, GeneSynbol, GeneName, AnotherSymbol, and AnotherName for each gene. OrganismName and NCBIGeneID are not used because organisms are all homo sapiens and NCBIGeneID is only number data. Multiple words are used for gene names, so we will not use them this time (ex. "ZZ-type containing 3"). This list has 59,545 genes.

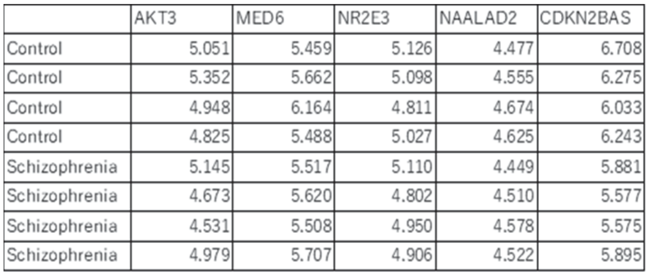
3.2 Experiment Setting

Word2vec has two methods for vectorization and two methods for speeding up calculation, but vectorization is skip gram and algorithm for speedup of calculation uses negative sampling. The other parameters are as shown in the table2. This is a parameter shown to be optimal for extracting similarity of genes by Chiu et al [9].

Table 2 : Word2vec parameters

Dimensions	200
Window size	2
Sub-sampling	1e^{-4}
Negative Sampling	5
Min-count	5
Learning Rate	0.05

Table 3 : Microarray data excerpts



	AKT3	MED6	NR2E3	NAALAD2	CDKN2BAS
Control	5.051	5.459	5.126	4.477	6.708
Control	5.352	5.662	5.098	4.555	6.275
Control	4.948	6.164	4.811	4.674	6.033
Control	4.825	5.488	5.027	4.625	6.243
Schizophrenia	5.145	5.517	5.110	4.449	5.881
Schizophrenia	4.673	5.620	4.802	4.510	5.577
Schizophrenia	4.531	5.508	4.950	4.578	5.575
Schizophrenia	4.979	5.707	4.906	4.522	5.895

3.3 Verification

To demonstrate the plausibility of the genes extracted by this research, we verified them with a xgboost. We used GSE92538, which is microarray data on schizophrenia that can be obtained from the NCBI database. This data has 228 samples, the "Control" has 175 samples and schizophrenia has 53 samples. Another sample of the disease was included, but it was deleted. We classified the microarray into Control and Schizophrenia. The microarray data is data as seen in Table 3.

Three patterns of features to be used were prepared, and their accuracy was compared. Regarding Pattern 1, only 2810 related genes in the GENE database and DisGeNet are featured. Regarding Pattern 2, the related gene extracted by this method is added to the 2810 of pattern 1 to prepare a feature. Regarding Pattern 3, the same number of genes as the related genes extracted by Pattern 1 by this method are randomly sampled to obtain the feature. It is like Fig.2 when I draw it in easy to understand. The parameters of the xgboost are selected from the parameters of the Table 4 for hyperopt. There are 10,000 trees. We use 10-fold cross-validation to obtain the accuracy of the classification.

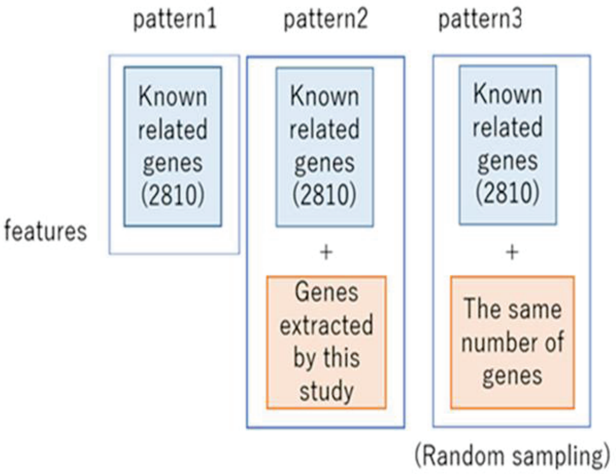

pattern1
pattern2
pattern3
features
Known related genes (2810)
Known related genes (2810)
+
Genes extracted by this study
Known related genes (2810)
+
The same number of genes
(Random sampling)

Figure 2: Three patterns used for verification. Pattern 1 is a known gene only Pattern 2 and 3 each add 62 genes. Pattern 3 is performed five times considering randomness.

Table 4 : Hyperopt parameters

Hyperparameters	Values
Max depth of tree	1 to 10 step1
Minimum weight required for child nodes	1 to 10 step1
Extraction ratio of training data when generating subsample	0.5 to 1 step0.05
Minimum value of loss reduction	0 to 1 step0.05

4 RESULT AND DISCUSSION

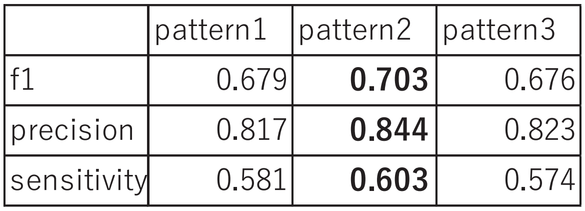
In this chapter, we will summarize the results obtained by the method discussed in this study and the knowledge obtained from them. First, when extracting the top 10 words with an appearance pattern close to each other using word2vec using each of 2810 known disease-related genes as an input, it was possible to extract a total of 9240 words. When extracting only the gene symbol from among them, the result contained 523 kinds of genes. This, it was confirmed that another gene can be extracted by extracting words having a similar appearance based on disease-related gene names. Since 461 known schizophrenia related-genes were included among the 523 kinds of genes, unknown genes that were not found in 62 databases could be extracted. This is indicated in Table 5. Among the 523 genes obtained with word2vec, 62 new disease-related genes were found. That is, 461 known genes are close to vectors of other known genes. Therefore, it can be seen that other disease-related genes can be extracted by extracting genes close to known disease-related genes using word2vec. Briefly, genes close to vectors in disease-related genes are likely to be disease-related genes themselves. From the above, it is highly likely that a gene extracted by this study is a disease-related gene.

The result of the xgboost verification is given in Table 6. The calculation of the t-test showed that this change was statistically significant at the p = 0.05 level. Classification was repeated ten times for the three patterns, and the average accuracy was calculated. The average was taken because the random nature of the xgboost changes the results slightly for each trial. Even random sampling of Pattern 3 is sampled five times, considering bias due to randomness. In other words, Pattern 3 is an average using 10 xgboost for 5 sampling. The accuracy of Pattern 2 is the highest in all indicators. By adding the gene extracted by the proposed method as a feature, it is possible to more accurately distinguish it. Therefore, there is a difference in expression level between the control and schizophrenia in these genes. There is thus a high possibility that it is a disease-related gene.

Table 5 : Related gene list (extracted in this research)

PLOD1	SPARC	CA1	MARS
KLHL1	UNC13B	SLC27A5	ADK
GATA3	CHN1	NKX2-1	FUT1
SH2B1	CPZ	CSTB	SPINT2
WAS	AFP	C6orf15	ELANE
WFS1	SULT2A1	KNG1	ARHGDIB
CHD8	RASL12	STK39	HOXA1
ADAR	CD44	SCN10A	RIC3
DECR1	AMPH	C2	PRB3
PAM	SPI1	TNFRSF25	ARNT2
RGN	C6orf48	CA3	SART3
CFD	FOXK2	HR	IL5
PGRMC1	GCG	CYP2C9	SP3
WNK1	CA4	PPIG	ERBB2
HOXD13	ARFGEF1	ARNT	
SYNE1	TNC	CD80	

Table 6 : The average result of performing xgboost ten times for each pattern



	pattern1	pattern2	pattern3
f1	0.679	**0.703**	0.676
precision	0.817	**0.844**	0.823
sensitivity	0.581	**0.603**	0.574

5 CONCLUSION

In this study, we aimed to extract disease-related genes from biomedical papers by text mining. We performed text mining considering the meaning of words using word2vec. Using this, we could extract a vector close to known disease-related genes, i.e., genes with similar meanings. There are 523 such genes including 62 genes other than known disease-related genes. In addition, by verification we could show the likelihood of those genes. From the above, genes with a high possibility of disease association can be extracted by this method. However, we must recognize that whether or not these genes are truly disease-related genes will be revealed only by biomedical experts.

REFERENCES

[1] Al-Mubaid H, Singh RK.(2005). A new text mining approach for finding protein-to-disease associations. Am J Biochem Biotechnol, pp. 145-152.

[2] Shahin Mohammadi, Sudhir Kylasa, Giorgos Kollias(2016). Context-specific Recommendation System for Pre-dicting Similar PubMed Articles. Data Mining Workshops (ICDMW)

[3] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word represen-tations in vector space. arXiv preprint arXiv:1301.3781.

[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed represen-tations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pages 3111–3119.

[5] Goldberg, Y. and Levy, O. (2014). word2vec explained: deriving mikolov et al.'s nega-tivesampling word-embedding method. arXiv:1402.3722 [cs, stat]. arXiv: 1402.3722.

[6] Segura-Bedmar, I., Suarez-Paniagua, V., Martinez, P. (2015). Exploring word embedding for drug name recogni-tion. In: Sixth Workshop on Health Text Mining and Information Analysis.

[7] Segura-Bedmar, I., Suarez-Paniagua, V., Martinez, P. (2015). Exploring word embedding for drug name recogni-tion. In: Sixth Workshop on Health Text Mining and Information Analysis.

[8] J. A. Mi˜narro-Gim´enez, O. Mar´ın-Alonso, and M. Samwald(2015). Applying deep learning techniques on med-ical corpora from the world wide web: a prototypical system and evaluation. arXiv preprint arXiv:1502.03682.

[9] Chiu, B., Crichton, G., Korhonen, A. and Pyysalo, S.,(2016). How to train good word embeddings for biomedical NLP. ACL 2016, p.166.

[10] PubMed.The National Center for Biotechnology Information(NCBI)http://www.ncbi.nlm.nih.gov/pubmed

[11] Gene. The National Center for Biotechnology Information(NCBI). https://www.ncbi.nlm.nih.gov/gene

[12] DisGeNET.http://www.disgenet.org/