Distributed Representation of Biomedical Words for Drug Repositioning

Abstract

Distributed Representation of Biomedical Words for Drug Repositioning

Graduate School of
Natural Science & Technology
Kanazawa University

Division of Electrical Engineering
and Computer Science

Student ID No.: 1323112003

Name: Ngo Duc Luu

Chief advisor: Professor Kenji Satou

Date of Submission: January 8th, 2016

Abstract

As a key technology of rapid and low-cost drug development, drug repositioning is getting more essential and popular for all pharmaceutical companies. Since drug repositioning means reuse of approved drugs for another purpose, their safety and method of production have already been confirmed. Besides biomedical experiments, computational methods are developed for drug repositioning. Most of them adopt network-based algorithms and combination of various databases including gene expression and pathway data. On the other hand, it is also suggested that text mining has much potential for drug repositioning. In biomedical text mining, biomedical named entities (e.g. genes, drugs, diseases, etc.) are recognized and the relations between these entities are extracted. Additionally, biomedical ontologies are utilized as the sources of semantic information for recognizing exactly names of biomedical entities.

In this study, a text mining approach to the discovery of unknown drug-disease relation was tested. Starting from over 3 million PubMed abstracts related to cancer, biomedical named entities were first recognized and the relations among them were extracted. Biomedical ontologies such as PharmGKB, MeSH, DrugBank, and CTD databases were utilized for the sources of semantic information. Using a word embedding algorithm, senses of over 1.7 million words were well represented in sufficiently short feature vectors. Through various analysis including clustering and classification, feasibility of our approach was tested. Finally, our trained classification model achieved 87.6% accuracy in the prediction of drug-disease relation in cancer treatment and succeeded in discovering novel drug-disease relations that actually reported in recent studies.

Chapter 1 Introduction

This chapter aims to introduce generally about the content and distribution of our dissertation.

Chương 2 Related works

2.1 Drug repositioning

As a new and effective strategy for drug discovery, drug repositioning (or drug repurposing, reprofiling, etc.) is attracting much interest and expectation from academic researchers and pharmaceutical companies [1]. Briefly saying, drug repositioning is reuse of existing drugs for other purposes.

Besides biomedical experiments, computational methods are developed for drug repositioning. Especially, in rerent years it is also suggested that text mining has much potential for drug repositioning.

2.2 Biomedical text mining

- ✓ **Text mining**

In general, text mining is the process of extracting information or discovering knowledge automatically from textual data or unstructured data. The process of text mining is quite similar to the process of data mining. Text mining is also known as an interdisciplinary area since it is related to various fields such as information retrieval, information extraction, natuaral language processing, data mining.

- ✓ **Biomedical text mining**

Biomedical text mining refers to the application of text mining to biomedical domain. Through biomedical text mining, useful knowledge, which is hidden in biomedical literature, can be extracted.

- ✓ **Information extraction (IE)**

Information extraction is the basic task for text mining. The process of information extraction includes two main tasks: named entity recognition (NER) and relation extraction (RE).

Chương 3 Materials and methods

3.1 Overview of processing pipeline

The processing pipeline in this study is shown in Figure 1. It includes four phases: Text processing; Word embedding; Combination of word vectors and database information; and Generation of drug-disease relation vectors.

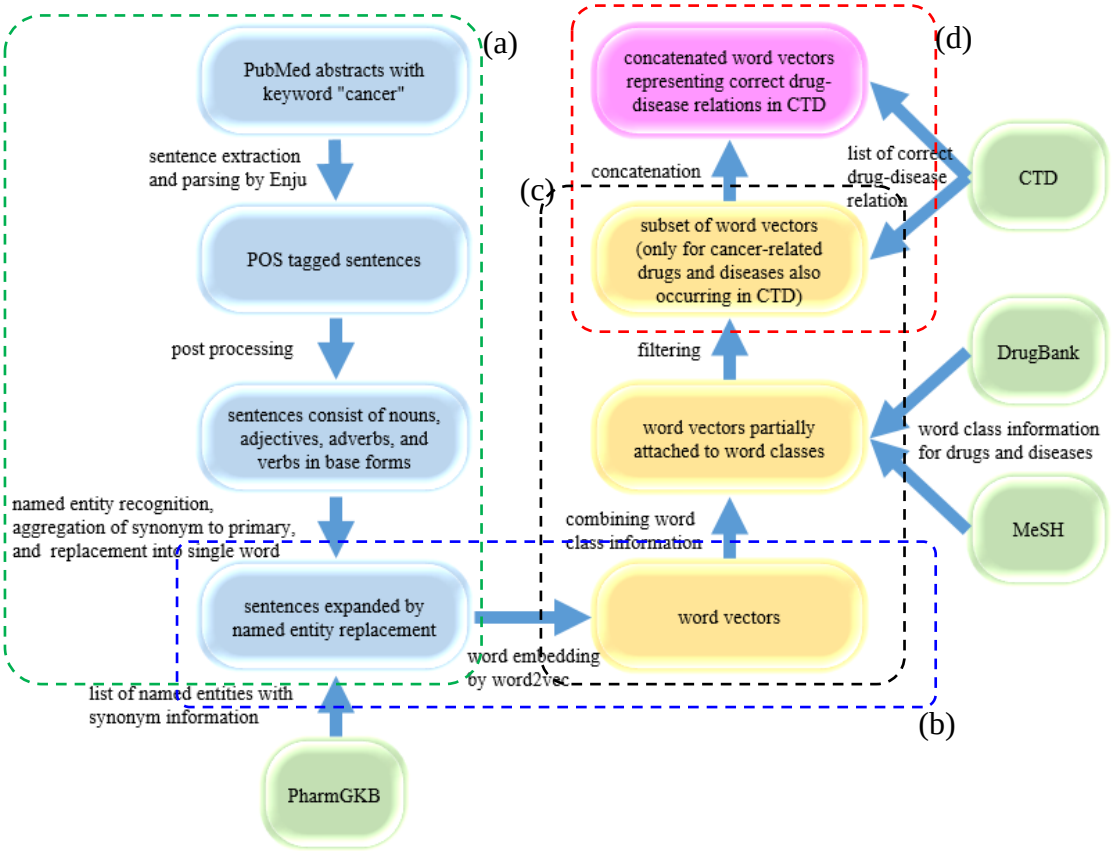

Figure 1. Overview of processing pipeline. Box colors indicate: light blue for corpus, light green for databases, yellow for word vectors, and pink for concatenated word vectors. (a), (b), (c), and (d) corresponding to four phases of processing pipeline.

3.2 Resources and tools

3.2.1 Biomedical databases

- ✓ **PubMed**

  PubMed [2] is widely used in biomedical text mining community. It allows users to access to a large biomedical database including the summary information of published biomedical articles such as author, title, abstract, keywords, etc.

- ✓ **Medical Subject Headings (MeSH)**

  MeSH is a comprehensive controlled vocabulary thesaurus of biomedical terms which is created and updated by the United States National Library of Medical (NLM) [3].

- ✓ **Pharmacogenomics Knowledgebase (PharmGKB)**

  PharmGKB is a pharmacogenomics knowledge resource which is used to aggregate, curate, integrate, and disseminate information about the relationships between human genetic variations and corresponding drug responses [4].

- ✓ **DrugBank**

  DrugBank [5] is a pharmaceutical database which contains knowledge about drugs and drug targets. In DrugBank, Anatomical Therapeutic Chemical (ATC) code is used to classify active ingredients of drugs according to the system or organ.

- ✓ **Comparative Toxicogenomics Database (CTD)**

  CTD provides useful information regarding relationships between different types of biomedical objects such as chemicals, and genes, and diseases [6].

3.2.2 Tools

- ✓ **Enju parser**
  Enju is a tool for parsing English sentences developed at Tsujii Laboratory in Tokyo University. In our study, Enju is used for Part-Of-Speech (POS) tagging and conversion of words into base forms.
- ✓ **word2vec software for word embedding**
  This software provides an effective implementation for computing vector representations of words. It was developed by Tomas Mikolov and his co-workers.

3.3 Data preparation

3.3.1 Cancer-related corpus from PubMed abstracts

As a raw corpus, we used a subset of PubMed abstracts downloaded in October 2013, filtered by the keyword "cancer". From 3,099,076 abstracts, 14,847,050 sentences were extracted.

3.3.2 Parsing sentences

Enju [7] was used for POS recognition of words and conversion into base forms. Since the sentences were extracted from biomedical abstracts, "-genia" option was specified. As a result, POS and base form are recognized for each word.

So that word2vec can differently treat the same word with different POS categories, they were attached right after the base form of words (e.g. "care" -> "care(V)"). For readability, nouns are kept as is. To simplify the input for word2vec, we removed all words except nouns, adjectives, adverbs, and verbs.

3.3.3 Named entity recognition and conversion into single words

Biological terms typically consist of two or more words. In addition, they have many synonyms. Since word2vec basically treats a sentence as a sequence of words, it is needed to recognize biological synonyms, aggregate them into primary terms, and convert them into single words. In this study, primary names and synonyms of drugs, diseases, and genes were extracted from PharmGKB [8] and used for recognition and aggregation (genes are used only for showing distribution of word vectors). For each converted single words, prefixes indicating their semantic categories were attached for later processing. After all conversion, 14,847,050 sentences are expanded to 45,264,480.

3.3.4 Word embedding

In this study, we used word2vec software, a de facto standard implementation of word embedding algorithm, with the following parameters by default. As a result, 1,772,186 words were embedded into word vectors (2,303 for drugs, 3,069 for diseases, 8,703 for genes, and 1,758,111 for others).

3.3.5 Combination of word vectors and database information

For the evaluation of clustering results, ATC codes [9] and MeSH tree numbers [3] were attached to drug and disease names, respectively. ATC codes were extracted from DrugBank [5]. Due to the incompleteness of data annotation, only 1,253 drugs out of 2,303 and 2,745 diseases out of 3,069 have such classification.

3.3.6 Generation of the special vector for drug-disease relations

For the evaluation of difference vectors between drugs and diseases, relations between drugs and diseases occurring in the corpus were extracted from CTD [6]. The figure 2 illustrates the concatenation of drug and disease vectors to create drug-disease relation

vectors. If the relation is written in CTD with therapeutic evidence, it is labeled "TRUE" (otherwise, "FALSE"). Only the 12,462 relations with therapeutic evidences were adopted for obtaining trustable results. In the set of relations, the mapping from drugs to diseases is many-to-many. For example, "drug::gefitinib" is related to 17 different diseases, and "disease::lung_neoplasm" is mapped from 60 different drugs.

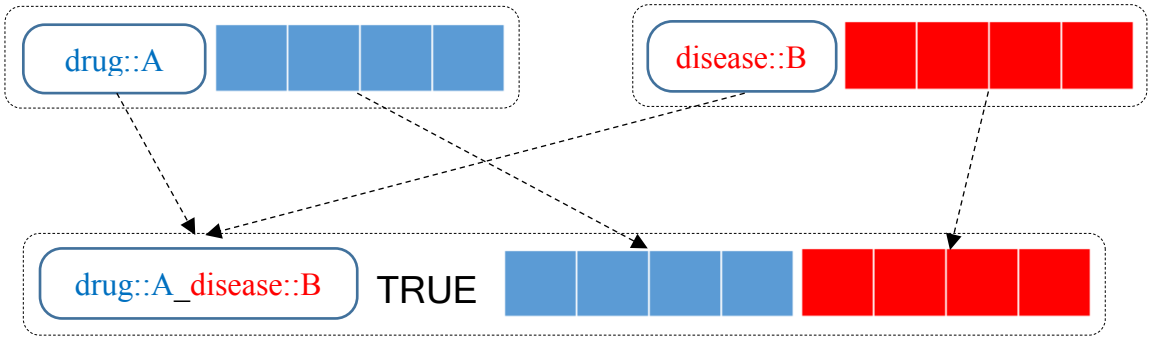

Figure 2. The concatenation of drug and disease word vectors to create drug-disease relation vectors. If the relation is written in CTD with therapeutic evidence, it is labeled "TRUE" (otherwise, "FALSE").

In order to conduct detailed analysis on cancer-related drugs and diseases, 12,462 extracted drug-disease relations were further filtered so that both of drug and disease names in each relation are attached to an ATC code and a MeSH tree number beginning with "L" (Antineoplastic and immunomodulating agents) and "C04" (Neoplasms), respectively. As a result, 1,097 relations consist of 104 anti-cancer drugs and 107 cancer-related diseases were extracted for detailed analysis.

3.4 Algorithms

3.4.1 Checking the distribution of biomedical word vectors

In order to check the distribution of biomedical word vectors for drugs, diseases, and genes, we adopted Principle component analysis (PCA) to convert 200-dimension vectors into 3-dimension vectors and plotted in 3D space. In this study, *prcomp* and plot3d functions included in stats and *rgl* packages respectively for R software were used.

3.4.2 Clustering for drugs and diseases

For visual evaluation of word vector quality, we performed hierarchical clustering with cosine distance and Ward's method [10]. Before the clustering, 2,303 drugs and 3,069 diseases occurring in the corpus were reduced to 1,282 and 1,051, respectively, so that all of them also occur in CTD. In this study, we used *hclust* and *plot* functions included in *stats* and *graphics* packages respectively in R software for hierarchical clustering and visualizing clustering performance.

3.4.3 Classification for drug-disease relations

Support Vector Machine (SVM) was adopted for learning and predicting possible relations between drugs and diseases. As an implementation, *ksvm* function included in *kernlab* package for R software was used with default parameters.

3.4.4 Cross-validation

Cross-validation is a very popular method which is used to evaluate the performance of classification problems. In this study, we performed 100 times 10-fold cross-validation. Then, the accuracies were averaged.

3.4.5 Methods for the class-imbalance problem

To solve class imbalanced problem, we applied the data level method in which corresponding to the number of positive examples (or *correct drug-disease relations*), the same number of negative examples (or *incorrect drug-disease relations*) were randomly selected, and used in each cross-validation.

Chương 4 Experimental results and discussion

4.1 Distribution of drug-disease-gene vectors

Figure 3 illustrates the 3D plot of vectors corresponding to 2,303 drugs, 3,069 diseases, and 8,703 for genes. For visualization, the dimension of vector was reduced from 200 to 3 by PCA. In the top panel of the figure, it is shown that the distributions of word vectors in three categories are clearly separated. In the bottom panel, it is also shown that the frequent words have clear separation, whereas it is relatively difficult to discriminate the categories of rare words.

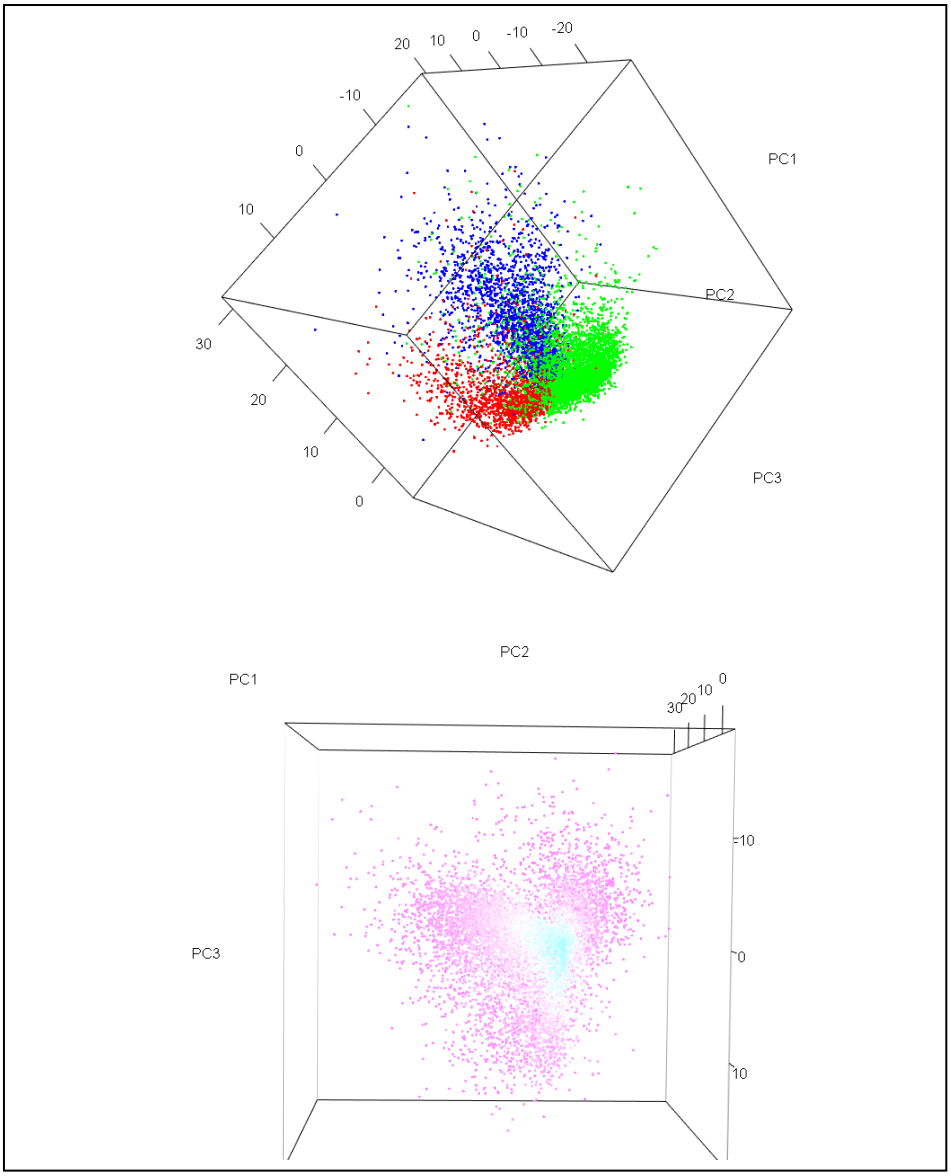

Figure 3. Distribution of word vectors visualized through PCA and 3D plot. Top panel: blue, red, green colors indicate word vectors for drugs, diseases, and genes. Bottom panel: color gradation from light blue to light pink indicates the frequency of words (from rare to frequent).

4.2 Cluster analysis for drug and disease vectors

Figure 4 and 5 show the result of hierarchical clustering for drugs and diseases, respectively. In the right panels of them, entire pictures of clustering results for 1,282 drugs and 1,051 diseases are shown. In the right panel of Figure 4, it can be seen that most of the anti-cancer drugs are condensed in the ninth cluster from the top (left panel for more detail). It indicates that the word vectors for drugs well represent the characteristics of corresponding drugs. Also in Figure 5, we can see that the seventh cluster from the top contains a number of cancer-related diseases, however, also in sixth and ninth clusters. The difference between these results might come from the fact that diseases can be classified from different perspectives (tissues, mechanism, etc.).

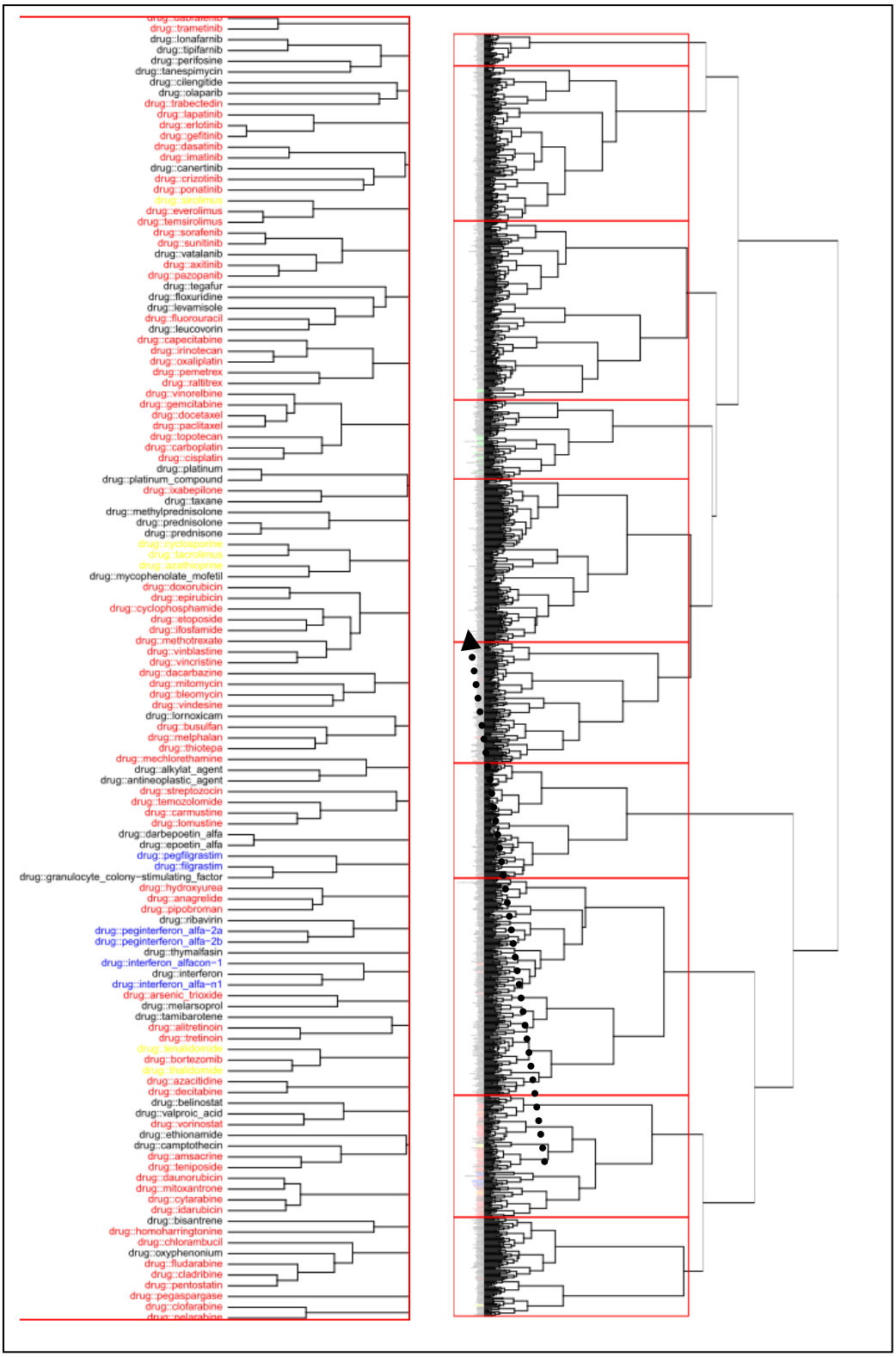

Figure 4. Result of hierarchical clustering on drugs. Red, green, blue, yellow colors for characters indicate that the drugs are classified in ATC codes as "L01:Antineoplastic Agents", "L02:Endocrine Therapy", "L03:Immunostimulants", and "L04:Immunosuppressants", respectively.

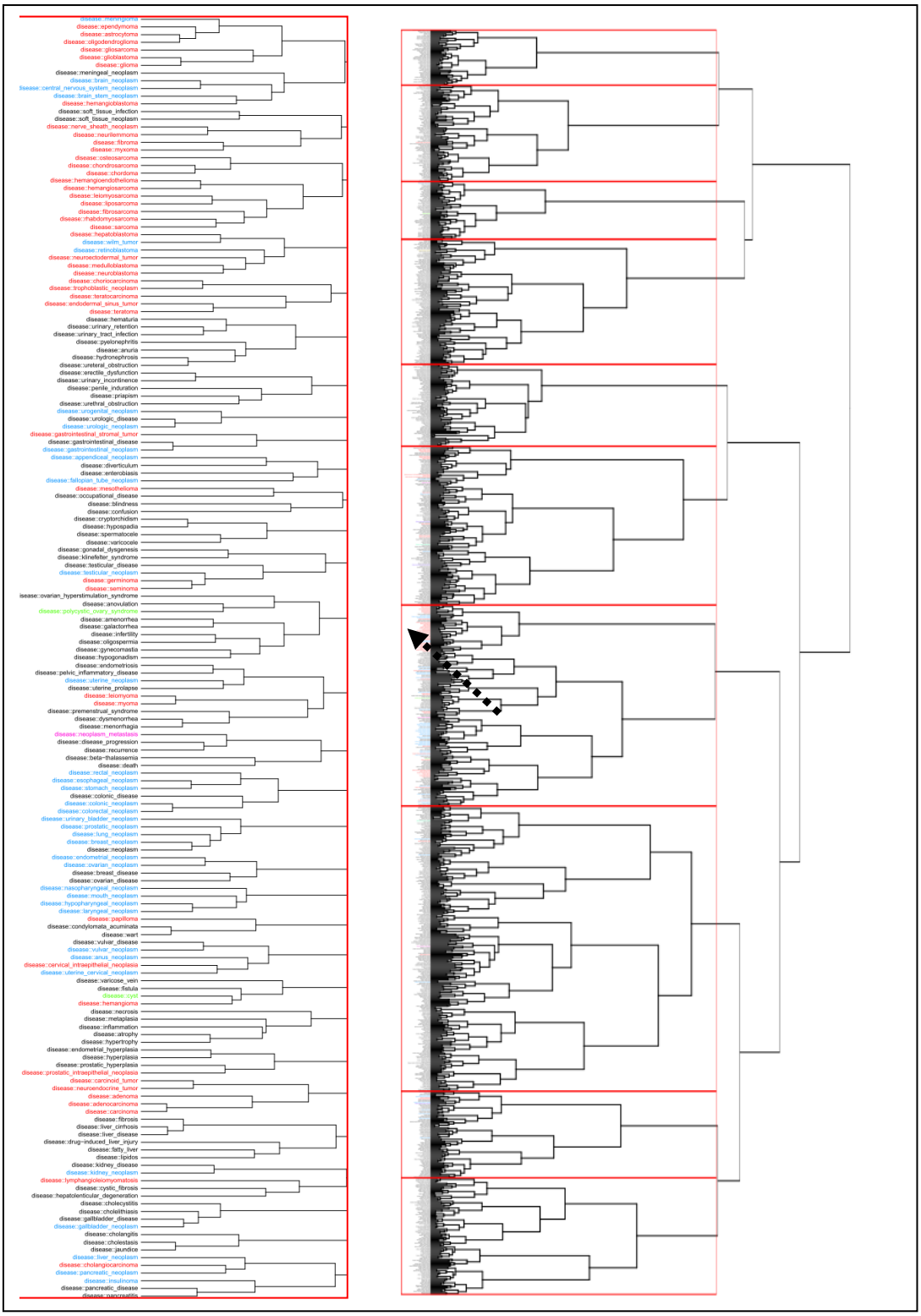

Figure 5. Result of hierarchical clustering on diseases. Green, yellow, red, dodgerblue, pink, blue, and springgreen colors for characters indicate that the diseases are classified in MeSH tree numbers as "C04.182:Cysts", "C04.445:Hamartoma", "C04.557:Neoplasms by Histologic Type", "C04.588:Neoplasms by Site", "C04.697:Neoplastic Processes", "C04.730:Paraneoplastic Syndromes", and "C04.834:Precancerous Conditions".

4.3 Classification of unseen drug-disease relations

The result of performance evaluation is shown in Table 1. Each accuracy is an average of 100 times 10-fold cross-validation with different subsets of negative examples. In this table, it was revealed that the performance of classification is affected by vector size and window size, and the best accuracy was 87.5%. For exploratory use of the classification model to discover candidate drug-disease pairs, it means sufficiently high performance.

Table 1 Accuracy of classifying correct and incorrect drug-disease relations by SVM.

vector size	accuracy			
	window size = 2	window size = 3	window size = 4	window size = 8
50	0.872	0.873	0.875	0.872
75	0.873	0.874	0.874	0.874
100	0.874	0.874	0.874	**0.876**
200	0.874	0.874	0.874	0.874

Finally, we tested all combinations of 2,199 drugs not used in training and 107 cancer-related diseases (in total, 235,293 drug-disease pairs). In case of the classification model trained by 11,128 samples, only 64 test samples were predicted as positive, and all the drugs in the sample were anti-cancer drugs (but not included in 104 anti-cancer drugs used for training). By controlling the degree of class imbalance in training data, it is possible to predict a pair of non-anti-cancer drug and cancer-related disease as positive. For example, using the classification model trained by 1,097 positive and 8,776 negative samples (degree of imbalance is 1:8), 10 times training and test by 235,293 drug-disease pairs discovered the following candidate drugs for repositioning to cancer treatment, where the numbers indicate how many times they were discovered in 10 times training and test.

drug::urokinase(10), drug::photodynamic_therapy(10), drug::oxygen(10), drug::nonoxynol-9(10), drug::nitroglycerin(10), drug::nitrogen(10), drug::l-phenylalanine(10), drug::l-methionine(10), drug::l-glutamine(10), drug::l-cysteine(10), drug::glutathione(10), drug::glucose(10), drug::epoxide(10), drug::enzyme(10), drug::collagenase(10), drug::bisphosphonate(10), drug::amino_acid(10), drug::amide(10), drug::clarithromycin(9), drug::vitamin(8), drug::l-proline(7), drug::vitamin_e(6), drug::xanthophyll(4), drug::phospholipid(4), drug::palifermin(4), drug::ether(4), drug::ethacrynic_acid(4), drug::denosumab(4), drug::egfr_inhibitor(2), drug::pyruvic_acid(1).

Besides too general names like “drug::enzyme” and “drug::amide”, it is notable that the above list includes approved anti-cancer drugs (e.g. “drug::denosumab”), anti-cancer drugs under investigation (e.g. “drug::clarithromycin”, “drug::bisphosphonate”, and “drug::xanthophyll”), and drugs potentially promote cancer (e.g. “drug::urokinase” and “drug::collagenase”). Especially, it should be emphasized that repositioning of clarithromycin to anti-cancer agent has been reported in 2015 [11], despite the fact that the corpus was downloaded in 2013. Though further screening based on expert’s knowledge is necessary, this result demonstrate that the classification of concatenated word vector is a promising approach to in-silico screening of drug-disease relations for drug repositioning.

Chương 5 Conclusion and future works

5.1 Dissertation summary

In this study, we applied text mining techniques to drug repositioning in order to (i) check the distribution of biomedical words; (ii) analyze clustering of anti-cancer drug vectors and cancer-related disease vectors; (iii) find unseen drug-disease relations by classification.

From experimental results, it was revealed that word embedding is effective for representing sense of all words in large amount of cancer-related PubMed abstracts. Furthermore, concatenation of word vectors of drugs and diseases well represents their relations and could be used for finding candidate drugs for repositioning by classification.

5.2 Future works

For better performance of classification, various feature selection and over-sampling algorithms [12] will be tested in the future work. Additionally, we will apply this method to other kinds of diseases as well as new kinds of biomedical relations such as gene-disease relations or drug-gene relations. Finally, if possible we will attempt to cover whole PubMed abstracts.

References

[1] Ashburn, T.T. and Thor, K.B. (2004) Drug repositioning: Identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683.

[2] Website of PubMed [http://www.ncbi.nlm.nih.gov/pubmed].

[3] C.E. Lipscomb. (2000) Medical Subject Headings (MeSH), *Bulletin of the Medical Library Association*, 88 (3), 265.

[4] Whirl-Carrillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B. and Klein, T.E. (2012) Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology & Therapeutics* 92(4): 414-417.

[5] Wishart, D.S., Knox, C, Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, 34(Database issue):D668-72.

[6] Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky D, King BL, Wiegers TC, Mattingly CJ. (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, 43:D914-D920.

[7] Miyao, Y. and Tsujii, J. (2008) Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):35–80.

[8] Whirl-Carrillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B. and Klein, T.E. (2012) Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology & Therapeutics* 92(4): 414-417.

[9] WHO Collaborating Centre for Drug Statistics Methodology, ATC classification index with DDDs, 2015. Oslo 2014.

[10] Xu, R. and Wunsch, D.I.I. (2005) Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645:678.

[11] Pantziarka, P., Bouche, G., Meheus, L., Sukhatme, V. and Sukhatme, V.P. (2015) Repurposing Drugs in Oncology (ReDO)-clarithromycin as an anti-cancer agent. *Ecancermedicalscience* 9, 513.

[12] Dang, X.T., Hirose, O., Bui, D.H., Saethang, T., Tran, V.A., Nguyen, T.L.A., Le, T.T.K., Kubo, M., Yamada,Y., Satou,K. (2013) A Novel Over-Sampling Method and its Application to Cancer Classification from Gene Expression Data. *Chem-Bio Informatics Journal*, 13, 19-29.

平成２８年　２月　２日

学位論文審査報告書（甲）

１．学位論文題目（外国語の場合は和訳を付けること。）

Distributed Representation of Biomedical Words for Drug Repositioning

（ドラッグリポジショニングのための生物医学用語の分散表現）

２．論文提出者　(1) 所　　属　　電子情報科学　専攻

(2) 氏（ふり）　名（がな）　　Ngo Duc Luu（ごー　どぅく　るー）

３．審査結果の要旨（600～650 字）

平成２８年１月２９日に第１回学位論文審査委員会を開催、平成２８年２月２日に口頭発表、その後に第２回審査委員会を開催し、慎重審議の結果、以下の通り判定した。なお、口頭発表における質疑を最終試験に代えるものとした。

薬剤開発のコスト削減を主な目的として、近年では既存の薬剤を従来と異なる病気の治療に用いるドラッグ・リポジショニング（DR）が盛んに研究されている。DR の候補探索には分子シミュレーションを用いることが多いが、本研究では、テキストマイニングの手法を用いた DR について研究を行った。PubMed からダウンロードしたがん関係の論文抄録を元に、ワードの品詞推定、原形変換、類義語の正規化、薬剤名や病気名の認識などの処理を行った上で、word2vec を用いて単語の意味を 200 次元程度の数値ベクトルとして表現し、階層型クラスタリングを行うことにより、がん関連の病気名や抗がん剤の意味が適切に表現されていることを確認した。さらに、がんについて薬剤と病気名の正しい組み合わせを学習し、予測を行うことで、これまで抗がん剤と認識されていない薬剤を DR の候補として発見できることを示した。

以上の研究成果は、コンピュータを用いた新たな DR の可能性を示すものであり、本論文は博士（工学）に値するものと判定した。

４．審査結果　(1) 判　　定（いずれかに○印）　（合　格）・　不合格

(2) 授与学位　　博　士（工学）