The Strategy of Reproducing Database-based Knowledge by Using Text Mining System

Tianyi Liu[1], Xuan Qin[2], Danbing Zou[1], Zhen Yu[1], YingYing Jiang[1], Jingbo Xia[2,3,4*]

[1]College of Science, Huazhong Agricultural University, Wuhan, China

[2]College of Informatics, Huazhong Agricultural University, Wuhan, China

[3]Hubei Key Laboratory of Agricultural Bioinformatics, Wuhan, China

[4]Institute of Applied Mathematics, Huazhong Agricultural University, Wuhan, China

*Corresponding author, mailto: xiajingbo.math@gmail.com

Keywords: BioNLP, DGidb, Drug-gene interaction, Reproducibility.

Abstract. To meet the aim of large-scale knowledge discovery, biomedical natural language processing is regarded as an effective tool to curate hidden information in research paper or related texts. Being a typical form of knowledge in bioinformatics, the interaction of drug and target genes plays vital role in medical-clinical research. Several NLP tools are used to extract drug/gene information from texts, and henceforth generate separate knowledge entries. Furthermore, the depth and coverage of NLP-generated knowledge entries are compared among different tools. The effectiveness of BioNLP tools is assessed eventually by evaluating their knowledge reproducibility ability for know data set, DGidb, a database with known drug-protein interaction pairs.

Introduction

Knowledge discovery based on text mining technique has played key roles in the known knowledge curation and novel relation extraction in the field of bio-medical research. Swanson's ABC model [1] was an innovative research, in which latent links between drugs and clinical purposes was mined out manually and it showed fish oil's new therapeutic effect upon Reynold' syndrome. Swanson's job was long regarded as a monumental research in drug repurposing, while text mining has been widely used in various applications including protein protein interaction (PPI) [2], drug drug interaction (DDI) [3], cancer genetics [4], and some other bioinformatics-related computation and inference [5-7].

Until now, dozens of BioNLP tools have been developed to incorporate the usage of bio text data and ontologies, and various task settings were designed. Co-occurrence strategy by associating co-occurred entities was extensively used in the context of literature-based knowledge discovery, so as to create inter-connection between genes, protein, and other biological entities [8]. The hypothesis is that a more frequent co-occurrence of entity pair leads to a higher chance for the relevance among them. Unfortunately, this strategy relied heavily on the quality of NER (Named entity recognition) output, and it took years for NLP community to develop sophisticated tools like ABNER [9], Banner [10], Pubtator [11], and tmchem [12]. Here, Pubtator was regarded as a standard bio-term NER tool, which annotate drugs, proteins, genes, mutations and species from abstract retrieved from NCBI Pubmed abastracts.

After co-occurrence strategy, sophisticated semantic relationship miner and algorithms were developed to enhance the text mining methodologies. Unlike co-occurrence strategy, which lack sufficient semantic evidence to support biomedical

findings, semantic strategy usually performed a high precision knowledge curation. A typical evaluation-aiming shared task (ST) was BioNLP 2009/2011 ST [13], in which the Turku Event Extraction System [14] was the best performing tool that was designed for the extraction of events and relations from biomedical text. In its strategy, entities were annotated first, followed with sentences parsing, and then an event detector was used to detect trigger words and relationship type between entities. By doing this, TEES determined event-level relationships between entity pairs. The results turned to be not only co-occurred but also semantic-oriented, and thus make it a sufficient reasonable target for performance evaluation.

The purpose of this research is to propose a new metric named Reproducibility Possibility (RP metric) to assess the possibility that the tool could reproduce knowledge entries in a known data set dgIDB [15]. Performance comparison was taken between two typical BioNLP tools, i.e., Pubtator for co-occurrence in a form of rough knowledge representation of biomedical entities, and TEES in a form of fine-grained knowledge representation. The RP metric were also utilized to compare the knowledge discovery rate in terms with the fine-grained extent of the task setting.

Result obtained from our experiments verified the common sense that rough text mining system, e.g., Pubtator, achieved very high recall rate while its positive precision rate dropped enormously, while fine-grained text mining system like TEES suffered from low recall rate but harvested high recall rate. Furthermore, RP metric was calculated for each strategy and the corresponding probability of reproducing knowledge were concerned, in the form of drug-gene pair. Henceforth, the hybrid strategy was also evaluated by computing PR metric, and the eventual result led to a better prospect of the hybrid strategy. Furthermore, the result in this manuscript also showed a potential prospect of looking for a better trade-off in Bio knowledge reproducibility by the evaluation of PR metric.

Material and Method

Material

Taking the concern of drug knowledge reproducibility, drugs and their targeted gene info were obtained from dgIDB (http://dgidb.genome.wustl.edu/), as listed in Table 1.

Table 1. Two repurposed drug with their targeted genes

Drug	Targeted Gene in DGidb	Pubmed Paper	Initial Indication	Repurposed Usage
Plerixafor	CXCR4, ACKR3, CCR4	1,013	HIV	Stem cell mobilizing
Rapamycin	MTOR, PIK3CA, TSC1	31,118	Transplant anti-resistance	pancreatic cancer

Two cases of drug-protein pairs curation were set as evaluation task, each of which presented a repurposed drug and its targeted genes. In detail, plerixafor was inititally an anti-HIV drug, and later found to be therapeutic active for stem cell mobilization. Furthermore, there were three targeted gene CXCR4, ACKR3, and CCR4 in dgIDB database. Similarly, rapamycin, with three target genes info in dgIDB, was previously a drug for curing transplant anti-resistance, but later found to be effective for pancreatic cancer.

All the abstracts were extracted from literature database of NCBI PubMed (https://www.ncbi.nlm.nih.gov/pubmed/). Here, 1,013 abstracts for plerixafor and other 31,118 for rapamycin. Taking concern that the drug-target gene relation is far not a straightforward observable knowledge among texts, it made the two cases illuminative targets for knowledge reproducibility testing in our research.

Method

Metric for Performance Evaluation of Text Mining System

Two traditional metrics, i.e., recall rate and precision (https://en.wikipedia.org/wiki/Information_retrieval#Precision) are introduced to evaluate the robustness of text mining system. Here

$$\text{Recall rate}=\frac{\#\{\{Retrieved\ Entries\}\cap\{Relevant\ Entries\}\}}{\#\{Relevant\ Entries\}}.$$

$$\text{Precision}=\frac{\#\{\{Retrieved\ Entries\}\cap\{Relevant\ Entries\}\}}{\#\{Retrieved\ Entries\}}.$$

Metric for Knowledge Reproducibility Evaluation of Text Mining System

Assume that there are n knowledge entries awaiting retrieving, the Probability of Reproducing (PR) the i-th knowledge entry from results obtained by text mining system, TMSys, is defined as

$$PR_{TMSys}(i)=\frac{\#\{\{Retrieved\ Entries\}\cap\{Related\ Entries\}\}}{\#\{Retrieved\ Entries\}}\times Prec(TMSys)\times REL(TMSys),$$

where $Prec(TMSys)$ refers to the system precision of $TMSys$,which is an empirical value for the whole text mining system, while $REL(TMSys)$ is a prior knowledge of reliability of the mined results.

Furthermore, the probability of reproducing m knowledge entries out of the n entries is defined as:

$$PR_{TMSys}(\bigcup_{i=1}^{m} i),\ m=1,2,3,\dots,n.$$

For instance, a PR metric for reproducing 3 knowledge entries is calculated as: $PR_{TMSys}(\bigcup_{i=1}^{3} i)=\overline{P(1)}P(2)P(3)+P(1)\overline{P(2)}P(3)+P(1)P(2)\overline{P(3)}$, where $P(i)$ is short for $PR_{TMSys}(i)$. Alternatively, the logarithm computation is taken into account so as to convert the original output to a proper numeric value. In addition, a plotting graph is also a visual option for the systematic evaluation of the whole system.

Result

Though being with different ways for knowledge presentation, average precisions are compared between Pubtator and TEES. For Pubtator, the correct co-occurrence of gene and drug in one sentence is regarded as true positive, and then the average precision of the system is 0.99. While for TEES, the report of the retrieved result is a semantic relationship between entities like gene and protein, and the manual annotation is carried to verify whether the semantic relation is corrected picked or not. Thus, the TEES has a low average precision value, i.e., 55%, which is a moderate result if compared with Pubtator, but still acceptable.

First, recall rate of gene-drug pair is compared between TEES and Pubtator. For the drug plerixafor, there are three target genes in DGidb, ie., CXCR4, ACKR3, and CCR4, and the hitting rate of TEES and Pubtator is in Table 2. While TEES found only CXCR4-Plerixafor pair info along with 56 relevant entries, and zero drug-gene pair for other two targets genes, Pubtator found every gene-drug pair info, which lead to higher recall than TEES. For the drug Rapamycin, all of the three drug-gene pairs were found both in Pubtator and TEES.

Table 2. Hit number of genes for TEES and Pubtator

Drug Target	Related entries found in TEES	Entries found in TEES	Related entries found in Pubtator	Entries found in Pubtator
CXCR4 for Plerixafor	56	136	1058	4636
ACKR3 for Plerixafor	0	136	2	4636
CCR4 for Plerixafor	0	136	4	4636
MTOR for Rapamycin	913	8322	30610	158609
PIK3CA for Rapamycin	4	8322	228	158609
TSC1 for Rapamycin	45	8322	481	158609

For the quality of the found entry, the composition of CXCR4-Plerixafor pair out of all filtered pair is 56/136 in TEES, while this composition value is 1058/4636 in Pubtator. This result shows similar performance for reproducibility. But cases varied among six chosen gene-drug pairs, so PR metrics are calculated for the two systems so as to evaluate the possibility of knowledge rediscovery. Here the systematic comparison of PR values are listed in Table 3, where $Prec(TMSys)$ of TEES is 0.55, and that of Pubtator is 0.99. Based on the prior knowledge that people relied lightly on co-occurrence result but counted more on semantic relationship, the value of $REL(TMSys)$ of TEES and Pubtaor are setted to be 0.1 and 0.01, respectively.

Table 3. PR value for the simulation results

PR value	m=0	m=1	m=2	m=3	m=4	m=5	m=6
Pubtator	9.96-01	4.22-03	4.55-06	2.51-10	4.48-15	3.07-20	6.73-26
TEES	9.74-01	2.61-02	1.21-04	3.34-08	7.34-13	1.47-63	7.34-115

PR value is to calculate how possible people will re-discover the drug-gene pair knowledge based on the observation of the text mining system report. It is clearly shown in table 3 that the Pubtator and TEES performed in different rate in reproducibility. In detail, the possibility of reproducing nothing (m=0) is both huge in TEES and Pubtator, i.e., 0.974 vs. 0.996. In a subtle comparison phage, when m=1, the possibility of re-discovery either one gene-drug pair is 2.61% for TEES, and 0.422% for Pubtator. This shows that text mining system with fine-grained task setting prevails the simple co-occurrence task. In both systems, to reproduce 2 out of 6 gene-drug knowledge has slight chance, not to mention more knowledge entries, which means that it is still remain a challenge for text mining system to carry on knowledge discovery in an ab initio principle.

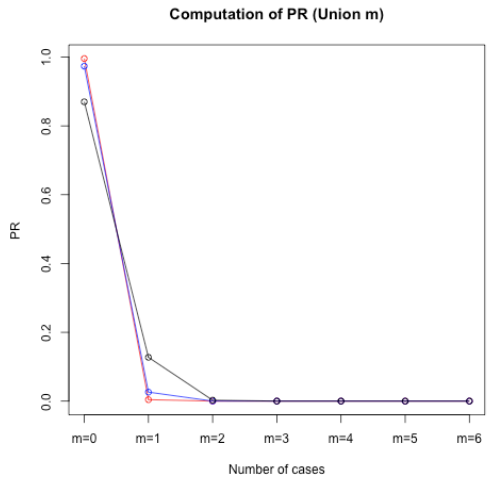

Figure 1. Comparison of PR value for Pubtator, TEES and Hybrid method in terms of knowledge re-discovery and data entry reproducibility in DGidb

As an application of the PR metric, we analyze the hybrid strategy of TEES and Pubtator by compare their PR value. The experiments are done by calculating the common hit of two text mining system, and the comparison of PR value is shown in Fig. 1. From the figure, it is clearly stated that hybrid strategy prevails the other two strategies. First, the PR value of m=0 case drops, while that in m=1 case increases.

Conclusion

As introduced the PR metric and the simulation experiments, several conclusion towards text mining system and knowledge reproducibility are obtained as below:

First, it remains a big challenge to reproduce most knowledge entries by purely using text mining systems. And a general hypothesis obtained from the simulation is that merely using a text mining system is not sufficiently capable of re-discovering the knowledge in known bioinformatics database. This is also an alert for BioNLP community to pay more attention to the integration of various Omics data when handling issues of bioinformatics discovery, besides the enhancement of the current methodologies of BioNLP theorems and toolkits.

Second, comparison of two TMsys (such as TEES and Pubtator) unveils the truth that the task setting of the system from scratch affects the performance of knowledge reproducibility substantially. It is possible that the more fine-grained system has the lower recall rate and the higher precision, and it doesn't always lead to high PR value. Neither a text mining system with rough knowledge mining purpose, nor a fine-grained text mining system achieved sufficient knowledge reproducibility. Actually, the reason for the narrow hope of large scale bio-knowledge reproducibility partly stem from the low recall of a fine-grained text mining system and a high false positive rate of a rough knowledge oriented text mining system.

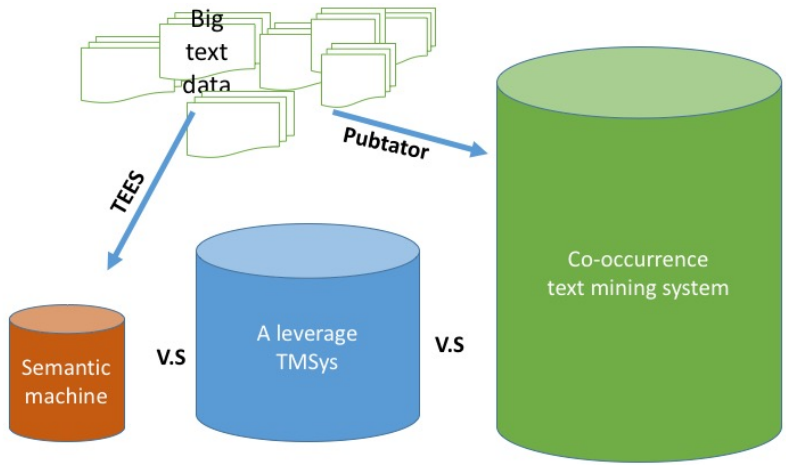

Figure 2. The existence of a leverage text mining system between Pubtator and TEES

Third, PR metric analysis is a reliable way for text mining system evaluation. As shown in Figure 1, when comparing a number of TMsys, hybrid strategy turns to be a potential leverage for excessive low recall and low precision. This leads to a hypothesis that there exists a comparatively reliable text mining system between co-occurrence system and semantic machine, e.g., Pubtator and TEES, which serves better in knowledge discovery and reproducibility. In the era of big data, a leveraging text

mining system will be suited for bioinformatics data curation, compared with those traditional knowledge engineering tools, as shown in Figure 2.

Acknowledgement

This research is funded by the National Natural Science Foundation of China (Grant no. 61202305) and the Fundamental Research Funds for the Central Universities (Project No. 2013PY120).

References

[1] Swanson, Don R. "Medical literature as a potential source of new knowledge." *Bulletin of the Medical Library Association* 78.1 (1990): 29.
[2] Sætre, R., Miwa, M., Yoshida, K., & Tsujii, J. I. (2009, June). From protein-protein interaction to molecular event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task* (pp. 103-106). Association for Computational Linguistics.
[3] Björne, J., Airola, A., Pahikkala, T., & Salakoski, T. (2011). Drug-drug interaction extraction from biomedical texts with svm and rls classifiers. *Proceedings of DDIExtraction-2011 challenge task*, 35-42.
[4] Pyysalo, S., Ohta, T., & Ananiadou, S. (2013, August). Overview of the cancer genetics (CG) task of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop* (pp. 58-66). Association for Computational Linguistics.
[5] Xia, J., Fang, A. C., & Zhang, X. (2014). A novel feature selection strategy for enhanced biomedical event extraction using the turku system. *BioMed research international*, *2014*.
[6] Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., & Lu, Z. (2016). A survey of current trends in computational drug repositioning. *Briefings in bioinformatics*, *17*(1), 2-12.
[7] Huang, C. C., & Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, *17*(1), 132-144.
[8] Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, *7*(2), 119-129.
[9] Settles, B. (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, *21*(14), 3191-3192.
[10] Leaman, R., & Gonzalez, G. (2008, January). BANNER: an executable survey of advances in biomedical named entity recognition. In *Pacific symposium on biocomputing* (Vol. 13, pp. 652-663).
[11] Wei, C. H., Kao, H. Y., & Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, gkt441.
[12] Leaman, R., Wei, C. H., & Lu, Z. (2015). tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics*, *7*(1), S3.
[13] Kim, J. D., Ohta, T., Pyysalo, S., Kano, Y., & Tsujii, J. I. (2009, June). Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task* (pp. 1-9). Association for Computational Linguistics.
[14] Björne J, Salakoski T. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task.
[15] Griffith, M., Griffith, O. L., Coffman, A. C., Weible, J. V., Mcmichael, J. F., & Spies, N. C., et al. (2013). Dgidb - mining the druggable genome. *Nature Methods, 10*(12), 1209.