Constructing Linguistic Verb Source for Relation Extraction

Yong Hwan Kim
Dept. of LIS
Yonsei University,
Seoul, South Korea
kimyonghwan
@yonsei.ac.kr

Seung Han Beak
Institute of Convergence
Yonsei University,
Seoul, South Korea
tourintocell
@gmail.com

Min Song
Dept. of LIS
Yonsei University,
Seoul, South Korea
min.song
@yonsei.ac.kr

Categories and Subject Descriptors
I.2.7 [**Document and Text Processing**]: Document and Text Editing – *Document management*

General Terms
Language, Experimentation.

Keywords
Text mining; big data; Biomedical Verb; PKDE4J

ABSTRACT
In bio-literature mining, relation extraction is one of the important subjects. In RE, verbs are a key to determine a relation type between entities. However, it was not paid much attention to the study related with construction of the biomedical verb list and the study of defining biomedical verbs. Therefore, we attempted to define the biomedical verb, and based on the definition, we established the actual list of verb and constructed the list using PKDE4J. We attempt to construct the verb list with relation type classified by their characteristics and a nominalization form of extracted verbs for usability. Finally, we established the actual list of biomedical verb and constructed the linguistic source out of the entire PubMed records.

1. INTRODUCTION
The important tasks of Bio-literature mining are Named Entity Recognition (NER) and Relation Extraction (RE). Many previous studies focused only on the extraction of entities and whether relation exists. Co-occurrence based approach is wildly used in RE [1-6]. In previous studies, do not describe the relation type between two entities.

Ontology based approach is generally used to extract more precise relations and relation types [7]. If two entities are extracted in a sentence, we can check the validation of this pair and extract the type of relation using ontology such as UMLS. However, this approach has two limitations. One is that the actual relation is not reflected. If two entities have only negative relation "decrease" in ontology, other relations which are presented in the text are disappeared. The result does not include the context information inside the sentence. The other is that an ontology has lower capacity than we need. Most ontologies are made by manual not in an automatic fashion. Therefore, it will be obvious that an ontology does not have sufficient information.

In RE, extracting context information in a sentence is important task because the information help us to determine relation type. Verbs are a key of relation type. Chklovski and Pantel [8] mentioned that "Verbs are the primary vehicle for describing events and expressing relations between entities." In recent study, Song et al. [9] investigated relation type using bio-verb dictionary. They constructed a biomedical verb dictionary using a verb list from Sun and Korhonen's research [10]. They only use 398 verbs to extract relations. Therefore there are limitations due to relation types are only extracted when the sentences include these verbs, while only simple relations such as co-occurrence are extracted when these verbs are not included in the sentences.

In this study, we focus on finding verbs presenting a relation type. If we can find a biomedical verb in linguistic characteristics, this finding will be a great assistance for bio-literature mining or Bio-NLP. Information about biomedical verbs would improve not only the performance of RE but also the syntactic or sematic parsing performance. In our research we focus on constructing a linguistic source that can comprehend the relation between entities extracted from a sentence. Especially, we attempt to construct the verb list with relation type classified by their characteristics. In addition, we provide a nominalization form of extracted verbs for usability.

The rest of the paper is organized as follows. In section 2, we discuss the definition of biomedical verbs. In section 3, we explore works related to biomedical verb list. We then describe our method to construct biomedical verb list in section 4. We analyze and discuss the results of our biomedical verb list in section 5 and 6. In section 7, we conclude the paper and suggest future lines of inquiry.

2. THE CONCEPT OF BIOMEDICAL VERBS
It is difficult to distinguish between biomedical verbs and general verbs. Many studies did not define what a biomedical verb is but simply regarded the verbs used in the biomedical field. Therefore, general verbs such as "have," "use," and "associate" are shown as biomedical verbs in previous studies because these verbs are mostly used in the biomedical field. The definition of the biomedical verb is necessary to overcome this problem.

In our study we define biomedical verb as a verb that is used in the biomedical field and that can explain the relation between two entities. In the bio-literature mining field the most important role of biomedical verb is in presenting the relation type between two entities.

General verbs can present the relation between two entities. Verbs such as "regulate" and "increase" can present the relation between two entities in common fields, while biomedical specified verbs such as "x-ray" is not able to present the relation between entities Therefore as shown in figure 1 verbs that can show the relation among all the general verbs in the biomedical field and among

verbs that can show the characteristics of the biomedical field are considered as Biomedical Verb.

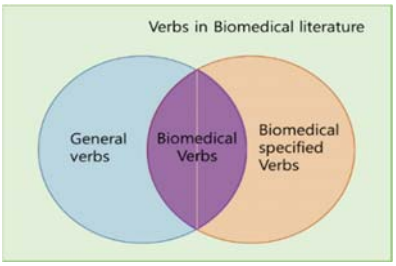

Figure 2. Definition of Biomedical Verbs

3. RELATED WORK

It is difficult to find studies that have constructed a biomedical verb list. In this chapter we examine studies that can identify the biomedical verb list.

Waxmonsky et al. [11] used 3 corpuses which are the general English corpus (newswire), MEDLINE and the sentences that include p53 (MEDLINE sub set) to extract verbs and group them to compare. By using a program called "linguistica" they tried to group similar verbs and their variations as one. They provide 16,601 verbs and its variations extracted from the Pubmed and ranked by appearance frequency and percentage.

To do Semantic role Labeling (SRL) Tsai et al. [12] used specific verbs. The 30 extracted verbs are associated with molecular event, and appear most frequently (not including have, show, use and more) in sentences that mention gene or protein. Though they tried to consider the relation between entities in extracting bio-verbs it does not reflect actual relevance.

Rimell et al. [13] constructed a verb resource that includes the subcategorization frame information and compared the performance. To construct the system they extracted 30 verbs that appear frequently by the standard frequency and used them for comparison.

Most related works do not include verbs to present the relation between entities. In Waxmonsky et al's research [11], they do not apply the lemmatization of the verbs and do not classify them by types therefore when looking at their high ranked verbs by frequency standard, verbs such as "was", "were", "is", "are", and "be" appear showing refinement is needed. Also in all the other studies many verb list do not fully present the relation between entities.

In our research we performed lemmatization for each verb, and classified each verb by their characteristics. Also because we extracted verbs that can show the relation between two entities from a sentence our verb list is specialized for relation extraction which gives its novelty.

4. METHODOLOGY

Using PKDE4J, we extract biomedical verbs. The basic assumption of PKDE4J is that the entity to the left of the main verb has effect on the entity to the right of the main verb. According to this hypothesis we can consider the main verb that is extracted by the PKDE4J as the biomedical verb that can show the relation between the two entities. By using this method we extract the biomedical verbs and by using the semantic relation of UMLS we classified the verbs. Figure 2 shows the overview of our methodology.

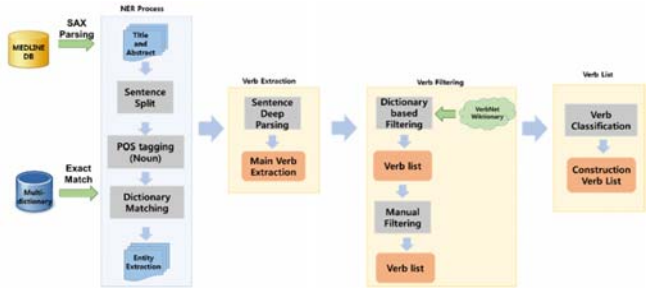

Figure 1. Overview

4.1 Data Collection

4.1.1 Text Collection

We used the entire MEDLINE records to extract main verb. As of 2014 fall, MEDLINE records consist of 23,769,884 articles stored in XML format. Out of these records, we extracted 14,447,667 records that have both title and abstract using SAX parser.

4.1.2 Dictionary data

To extract the main verb by using the PKDE4J we have to set the entity that can be the subject of relation. PKDE4J uses the dictionary based NER method. Therefore to extract diverse biomedical verbs we have to use diverse entities many as possible. In our study to extract relation between entities as shown in Table 1 we use 10 entities which are Gene/protein, Cell, Cellular component, Molecular function, Biological Process, Body part, Disease, Drug, Tissue, and Metabolite.

Table 1. Dictionaries for extraction of main verb

Entity type	Dictionary	# of unique name
Cell	KEGG [14]	1,559
Cellular component	HMDB [15]	672
Molecular function	Gene Ontology [16]	14,857
Biological Process	Gene Ontology [16]	43,391
Gene/protein	Entrez Gene [17]	104,872
Body Part	KEGG [14]	564
Disease	MeSH, KEGG [14]	73,345
Drug	DrugBank [18]	30,703
Tissue	Tiger [16], GDSC [19]	76
Metabolite	HMDB[15]	297,256

4.2 Verb Extraction

The application of PKDE4J is as follows. First, for each paper the title and abstract are split to sentences. A sentence is tokenized by each word and by using the 10 constructed entity dictionaries exact matched entities are extracted from the sentence. Sentences that were extracted for at least two entities were used for extracting main verbs. PKDE4J extracts the main verbs that have dependency relation with two entities that were already extracted and the main verbs that match the Bio-Verb DB. However, in our study we deleted the process that matches the main verb with the Bio-Verb DB and extracted all main verbs.

We extracted total 72,844 verbs from the entire PubMed records. Table 2 shows the top 20 extracted main verbs by frequency standard.

Table 2. Top 20 Verb by frequency

Rank	Verb	Freq.
1	have	2,960,104
2	use	2,744,217
3	associate	2,034,881
4	show	1,987,012
5	increase	1,880,464
6	induce	1,461,874
7	inhibit	1,443,204
8	investigate	1,363,211
9	be	1,327,238
10	study	1,323,839
11	find	1,305,364
12	reduce	1,168,892
13	determine	1,120,401
14	include	1,114,958
15	measure	1,092,983
16	examine	1,071,902
17	observe	983,732
18	evaluate	940,605
19	cause	936,767
20	decrease	847,074

In the top 20 verbs we can find verbs such as “have, use, and associate” that can present the relation while also we can find verbs such as “show, investigate, study, and examine” that do not present the relation between two entities. In addition as the rank becomes lower we can see typographical error and words that are not verbs. In this study we use two filtering processes to solve such problems. VerbNet and Wiktionary was used for the first filtering. By using each dictionary the word was extracted if it had verb meaning. Through this process we extracted total 8855 words. The second filtering process was done by manually and each meaning was determined to extract verbs with relation. The manual filtering was done by 2 people and if they did not agree they decided by mutual consent. By 2 filtering processes total 4524 verbs were extracted for the Verb List.

4.3 Verb Classification

Finally we classified the verbs to specific types. We used semantic relation of UMLS as our classification standard. Semantic relation is the relation of each semantic type in the Semantic Network of UMLS.

Semantic relation is composed of a hierarchy relation and 54 relations. We can find semantic relations that are difficult to distinguish such as “physically_related_to”, “conceptually_related_to”, “functionally_related_to”, “temporally_related_to”, “conceptual_part_of” and “spatially_related_to”. Therefore excluding such relations total 48 semantic relation was used as standard to classify types. Table 3 shows the relation type that was used for our study.

Using these relation types we classified the Verb list. We utilized WordNet and Wiktionary for the classifying process. First, by using JWNL we searched the word in WordNet and extracted the synonyms, and synonyms were extracted after searching the relation type then classified to the type that has the most mutual synonyms. After using JWPL to search for the verb and relation type in the Wiktionary, similarity was calculated and classified to the relation type with the highest similarity. The similarity formula that was used was Cosine-Similarity. Manual inspection was done after to examine the classified words.

4.4 Verb nominalization

In order to create a linguistic source of BioNLP and Bio-literature mining, we included the noun form in addition to an establishment of the List of verbs. A relation between the entities is not necessarily expressed as a verb. If the noun form and gerund form is added to each verb, the relation between entities will not only be expressed as a verb but also it would be capable to extract relations that are in the form of noun or gerund.

5. Result

We have constructed biomedical verb List and constructed a relation type for each word. Also by adding the noun form and the gerund form for each verb we made it possible for using as a linguistic source.

We constructed the Biomedical Verb List of a total of 4,525. This list can be used effectively in expressing a relation between the entities. The following is the example of the established file.

Table 3. Top 10 biomedical verb list

Relation	Rel. type	Verb	Nominal -ization	Freq.
neutral	result_of	have	having	2,960,104
neutral	result_of	use	using	2,744,217
neutral	result_of	associate	associating |association	2,034,881
neutral	location_of	increase	increasing	1,880,464
neutral	part_of	induce	inducing	1,461,874
neutral	branch_of	inhibit	inhibiting	1,443,204
neutral	Isa	be	being	1,327,238
neutral	part_of	find	finding	1,305,364
neutral	result_of	reduce	reducing |reduction	1,168,892
neutral	process_of	include	including	1,114,958

This Verb data can be utilized through the following URL http://informatics.yonsei.ac.kr/tsmm/data/Biomedical_VerbList.xlsx.

6. DISCUSSION

In the process of conducting this study, we were able to discover a discussion point of the ‘typographical error’. The typographical error can be examined in two ways. First, it contains English and American expressions to the corresponding word. In case of “deproteinize” and “deproteinise,” these two incorporate the same contents, but used simultaneously in both English and American expressions. If a particular one is added, the other expression is hardly reflected in the relationship extraction. This study collects and utilizes all of these two expressions.

Second, we discovered the issue of the typographical error. Unexpectedly, a large number of typographical error is visible in MEDLINE. Also, errors such as missing a character or adding other

characters can be found. A method that extracts all of the error types occasionally does not exist. These issues can be the cause of serious errors in the dictionary based approach. Because the words of the typographical error cannot be extracted, which then will not include these words, a decline of an efficiency will occur

In order to solve this problem, a system that solves the errors in the word should be taken into consideration. Text is written by a human, so errors exist. Moreover, each individual has a different way to the expression. Therefore, it is significant to find a solution that comprises all of these issues. Web search engines, for instance, maintain a research of a system that automatically corrects typographical error based on the search terms that people use and actually utilize them. However, despite carrying out the effort to solve typographical error of the entity in the biomedical sector, it is difficult to search for the efforts to modify an error of a verb. Therefore, solving typographical error can be seen as a single study area.

7. CONCLUSION

We developed the study to create a linguistic source. We defined the Biomedical Verb and based on this definition, we established the actual list of verb and constructed the linguistic source by the progress of the classification and by adding nominalization form. This study contained numerous manual works that may create limitations. However, it comprises great significance of creating a linguistic source beyond a plain verb list by establishing a noun form and a gerund form for each word and its classification. For future study, we will conduct a research that enhances the efficiency by utilizing and applying this source to the bio-literature mining.

8. ACKNOWLEDGMENTS

This work was supported by the Bio-Synergy Research Project (NRF-2013M3A9C4078138) of the Ministry of Science, ICT and Future Planning through the National Research Foundation.

9. REFERENCES

[1] Jenssen T-K, Lægreid A, Komorowski J, Hovig E (2001) A literature network of human genes for high-throughput analysis of gene expression. Nature genetics 28, 21-28.

[2] Jelier, R., Jenster, G., Dorssers, L. C., van der Eijk, C. C., van Mulligen, E. M., Mons, B., & Kors, J. A. (2005). Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. Bioinformatics, 21(9), 2049-2058.

[3] Leroy, G., & Chen, H. (2005). Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. Journal of the American Society for Information Science and Technology, 56(5), 457-468.

[4] Li S, Wu L, Zhang Z (2006) Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. Bioinformatics 22, 2143-2150.

[5] Song M, Han N-G, Kim Y-H, Ding Y, Chambers T (2013) Discovering implicit entity relation with the gene-citation-gene network.

[6] Chen G, Cairelli MJ, Kilicoglu H, Shin D, Rindflesch TC (2014) Augmenting microarray data with literature-based knowledge to enhance gene regulatory network inference.

[7] Leroy, G., & Chen, H. (2005). Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. Journal of the American Society for Information Science and Technology, 56(5), 457-468.

[8] Chklovski, T., & Pantel, P. (2004, July). VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In EMNLP (Vol. 4, pp. 33-40).

[9] Song, M., Kim, W. C., Lee, D., Heo, G. E., & Kang, K. Y. (2015). PKDE4J: Entity and relation extraction for public knowledge discovery. Journal of biomedical informatics, 57, 320-332.

[10] Sun, L., & Korhonen, A. (2009, August). Improving verb clustering with automatically acquired selectional preferences. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2 (pp. 638-647). Association for Computational Linguistics.

[11] Waxmonsky, S., Goldsmith, J., & Rzhetsky, A. (2010, December). Discovering and counting biomedical verbs. In Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on (pp. 975-978). IEEE

[12] Tsai, R. T., Chou, W. C., Su, Y. S., Lin, Y. C., Sung, C. L., Dai, H. J., ... & Hsu, W. L. (2007). BIOSMILE: a semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. BMC bioinformatics, 8(1), 325

[13] Rimell, L., Lippincott, T., Verspoor, K., Johnson, H. L., & Korhonen, A. (2013). Acquisition and evaluation of verb subcategorization resources for biomedicine. Journal of biomedical informatics, 46(2), 228-237.

[14] Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27-30.

[15] Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y.\& Bouatra, S. (2012). HMDB 3.0—the human metabolome database in 2013. Nucleic acids research, gks1065.

[16] Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 32(suppl 1), D258-D261.

[17] Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. Nucleic acids research, 39(suppl 1), D52-D57.

[18] Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D. & Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research, 36(suppl 1), D901-D906.

[19] Liu, X., Yu, X., Zack, D. J., Zhu, H., & Qian, J. (2008). TiGER: a database for tissue-specific gene expression and regulation. BMC bioinformatics, 9(1), 271.

[20] Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S. & Ramaswamy, S. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic acids research, 41(D1), D955-D961.