RESEARCH

YBio M^{2}CMT : a flexible bio-text mining tool for multi-component multi-target

Min Song*†, Sung Jeon Song, Yong Hwan Kim, Kyu Bin Ahn and In Jun Baek

*Correspondence:
min.song@yonosei.ac.kr
Department of Library and Information Science, Yonsei University, Souel, Korea
Full list of author information is available at the end of the article
†Equal contributor

Abstract

YBio M^{2}CMT, Yonsei Bio-text Mining for Multi Component Multi Target System, is a bio-text mining system that is well suited for datasets that contain bio-named entities. This system use PKDE4J (Public Knowledge Discovery Engine for Java) toolkit, developed by us, as core module to annotate from free text. PKDE4J is a comprehensive biomedical text mining toolkit including NLP, NER and RE. The system was applied to MEDLINE as base text. We extracted the title and abstracts by SAX XML parsing. With these extracted abstracts, we automatically annotated bio-named Entity using PKDE4J. The NER is carried out using a dictionary-based approach, and the detected named entities were written to files as Brat format. These annotation results are available at http://informatics.yonsei.ac.kr/tsmm/bio_textminer.html. In this system, Entity Detection System is working, and Relation Detection System that will be available in the near future is now in developing step.

Keywords: YBio M^{2}CMT System; PKDE4J; NER System; RE System

1. Introduction

1.1. Overview

YBio M^{2}CMT is a bio-text mining system that is well suited for datasets that contain bio-named entities. This system use PKDE4J (Public Knowledge Discovery Engine for Java) toolkit [1], developed by us, as core module to annotate from free text. PKDE4J is a comprehensive biomedical text mining toolkit including Natural Language Processing (NLP), Named Entity Recognition (NER) and Relation Extraction (RE). The system was applied to 2014 MEDLINE®/PubMed® Baseline XML files [2] as base text, contains 22,376,811 records and 101,292,963,034 bytes. Through the FTP server (ftp://ftp.nlm.nih.gov/nlmdata/.medleasebaseline/) of NLM (National Library of Medicine), we acquired the entire data files and extracted the title and abstracts by SAX XML parsing. With these extracted abstracts, we automatically annotated bio-Named Entity using PKDE4J. The NER is carried out using a dictionary-based approach, and the detected named entities were written to files as Brat v1.3 format [3]. These annotation results are available at http://informatics.yonsei.ac.kr/tsmm/bio_textminer.html. In this system, Entity Detection System is working, and Relation Detection System that will be available in the near future is now in developing step. Figure 1 is the main page of YBio M^{2}CMT System.

1.2. Goal of BLAH3

We join BLAH3 with the YBio M^2CMT System that is a biomedical text mining toolkit. Currently, free text including XML, CSV and TXT file format is fed into the system. The system then generates the annotated results in a structured text file containing entity, entity type, relation type, relation verb, relation direction, negation, voice etc. We plan to diversify the input and output in terms of format and style for interoperability with other applications or datasets. Through BLAH3, we aim at identifying the suitable input/output formats and requirements of other applications or datasets that could apply this toolkit, and at developing specific connectors for them. We also expect to improve the performance of the YBio M^2CMT System through the annotation datasets of other participants. Manual curated data sets from PubAnnotation or other participants are useful data for higher performance of our Entity Detection System. We believe that the performance of our relation detection module in developing process also can be improved in BLAH3. Additionally, it is our desire that our PubMed NER annotation results data could be shared through PubAnnotation for other participants, and we plan to add new functions to the system so that the system can be extends to combine with other participants' tools (e.g. event extraction).

2. Detailed description of NER/RE annotation method

We used PKDE4J toolkit for annotation (version 1.1) which we developed for multi-component multi-target extraction. It is an extension of StanfordCoreNLP [4] for performing dictionary based NER and publicly available at http://informatics.yonsei.ac.kr/pkde4j. Figure 2 is the overall process.

2.1. Preprocessing

We parsed the XML files due to our base text data is 2014 MEDLINE®/PubMed® Baseline XML data. After XML parsing, we conducted several preprocessing step. First step is Abbreviation Resolution. If a text which we are going to handle appears in form of abbreviation in paragraphs or abstracts, it will cause inaccurate n-gram matching. To solve this problem, our system module utilizes abbreviation resolution. Second step is String normalization. To reduce the string variations such as case sensitivity and special characters, we replace all uppercase characters into lowercase characters and remove special characters. After going through these preprocessing steps, we used Stanford CoreNLP to conduct Sentence Splitting, Tokenization, Part-of-speech tagging and Lemmatization.

2.2. Entity Extraction

PKDE4J system is designed to make use of several dictionaries. We classified biomedical words into 11 categories, Gene/protein, Disease, Cell, Cellular component, Molecular Function, Biological Process, BodyPart, Drug, Metabolite, Tissue and Organism. Our dictionaries are made from a combination of data collected from open medical databases and contains about 530,000 words. PKDE4J system also supports other types of dictionaries and can be creating annotation dataset using UMLS created by NIH. In the Entity Annotation, we first generate the Word N-gram tokens using the Apache Lucene [5] ShingleWrapper in the N-gram matching

process. Second step is Approximate String Matching, the dictionary data prepared in advance is collated with the string to display the inconsistent. Third step is POS filtering step. We removed the tokens that are judged to be determiner (DT), adverb (RB), comparative adverb (RBR), and superlative adverb (RBS). Finally, in the labeling, we adopted BIO format as labeling scheme, which shows B for the starting of an entity, I for the inside of an entity, and O for the outside of an entity. In version 1.1, we provided a method of filtering candidate entities by UMLS to improve accuracy of entity extraction. In addition, we provided the Conditional Random Field (CRF) based machine learning approach to NER. In the configuration setting, the user can select either mode for NER.

2.3. Post-processing

For further improvement in the quality of extraction, we adopted entity mapping based on regular expressions to the types of entities using Regex NER. It defines cascaded patterns over token sequences, providing a flexible extension of the traditional regular expression language defined over strings. We define a set of rules for each entity types that expresses several patterns of entity mentions by analyzing the corpora, and also those patterns are described. With BIO labels assigned in the previous stage. The rule set is then applied to the pipeline so that PKDE4J can relabel the tokens if any predefined rule is matched.

2.4. Relation Extraction

This module provides relation annotation for the sentence that contains two or more entities. In RE process, extracted sentences were fed to the dependency parsing technique to find the important bio-verbs. If there were two entities on one sentence and if the main verb matched with the verb list from a BioVerb DB, two entities and the main verb were extracted. Our basic assumption is that the entity to the left of the main verb has effect on the entity to the right of the main verb. After extracting the candidate set with the basic assumption mentioned above, we have determined the final relation extraction result through more sophisticated 17 rules. Each rule was constructed to identify whether the entities in a sentence have a direct dependency relation with the main verb. Typical examples are dependency relation sequences between entities and verbs, voice, negation, and distance on parse tree and so on. The result obtained after executing the RE module is provided not only with whether there is a relationship between two entities or not, but also with the direction of relation, relation verb and relation type.

2.5. Flexibility

There are a number of toolkits that perform Named Entity Recognition (NER) and Relation Extraction (RE) with bio data. Recently, the hybrid NER approach has been reported to have a high performance, but it could detect only the entity type that has already been determined. PKDE4J supports both dictionary-based and hybrid approach. Hence, if only the dictionary is configured for the desired entity types, the matched entities and their relation are only extracted. Also, it could provide an easy way to construct the entity-entity relationship network by outputting RE results as predicate form (entity1-relation-entity2). Since PKDE4J can be customized to meet the needs of various users, we can help most of the environment

with automatic annotate needs. This module provides relation annotation for the sentence that contains two or more entities.

3. Web-based YBio M^{2}CMT System

3.1. Search for annotation data

This system provides the NER annotation result which was automatically annotated by the PKDE4J toolkit via web. By supported two search modules, the users are able to obtain the needed annotation data from whole PubMed NER result. Figure 3 show the main page of the YBio M^{2}CMT system. The annotation data can be found through PubMed ID or keyword. The keyword search module provides a search function to narrow down the target PubMed records by query. The indexing and searching utilities are based on Apache Lucene. We indexed the 2014 MEDLINE®/PubMed® Baseline abstract data. The current version of the search function is simple, but we plan to add more advanced search functions so that the user can spot the right PubMed record to work on. Figure 4 shows the keyword search result. By clicking the PubMed ID in the left of search results, you can see detailed annotation information for one document. It is also accessible through PubMed ID search function. As shown in the figure 5, this system provides three items in search result page: 1) visualization of annotation, 2) list of annotated tokens and their basic information, and 3) the proportion of entities by their type. Brat is applied to display NER annotation with abstract text and Google Chart API is applied to draw chart on the web.

3.2. Annotated data format

To make YBio M^{2}CMT properly work, there are two types of input files are required. First one is an annotation file. The annotation format of our NER dataset followed Brat 1.3 format [2]. The Brat annotation format basically consists of two parts: 'Text files (.txt)' and 'Annotation files (.ann)'. Text file contains the base texts (abstracts) to be annotated that are extracted from the MEDLINE XML file. Annotation file contains NER annotations such as entity type, entity text, and entity text alignment. The names of these two files are equally given as PMID, but they could be identified with different file extension (e.g. PMID-15687000.txt, PMID-15687000.ann). The following is an example of the contents of actual annotation files. The second file is a plain text file that is mostly PubMed abstract.

Text file (.txt)

The text file (.txt) contains the abstract texts that correspond to each PMID. NLP is performed in the whole NER process, but the annotation result is displayed on the original abstract text. The following Table 1 is an example of text file.

Annotation file (.ann)

As shown in the Table 2, the NER annotation file (.ann) prints annotation results on a line-by-line basis, with three items: annotation ID, entity annotation, and the text of the span. Each is delimited by a tab delimiter, and the items are as follows.

- Annotation ID: The IDs assigned to each entity extracted from the target article are given in the form of a combination of the letter 'T' and the number in order.

- Entity annotation: It is the section describing the location and the type of the identified entity. The defined entity type is filled in first, followed by the character indexes where the entity is located in the text file after the space. The start index and the end index should be listed and separated by spaces.
- Entity text: It puts the actual text corresponding to the character indexes from the text file.

4. Data Sharing

We look forward to sharing our annotation data on PubAnnotation. Since our data format is not yet in compliance with PubAnnotation, we are in process of converting our data format to TEXTAE friendly JSON format. We anticipate that we can connect our data with others and contribute to the development of biomedical field through PubAnnotation. In addition, we plan to show our data as JSON format on website along with annotation result page. We will make JSON format data reachable via REST API. When the relation detection system is completed in the future, it will be uploaded in PubAnnotation. Through this system, PubAnnotation users can also crosscheck the annotation results.

5. Annotation toolkit (PKDE4J) reltated References

Main Article for Annotation

- 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.

PKDE4J applied Articles

- Kim, E. H. J., Jeong, Y. K., Kim, Y., Kang, K. Y., & Song, M. (2015). Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. Journal of Information Science, 0165551515608733.
- Kim, J. D., Wang, Y., Colic, N., Baek, S. H., Kim, Y. H., & Song, M. (2016). Refactoring the Genia Event Extraction Shared Task Toward a General Framework for IE-Driven KB Development. ACL 2016, 23.
- Verspoor, K. M., Heo, G. E., Kang, K. Y., & Song, M. (2016). Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts. BMC medical informatics and decision making, 16(1), 68.
- Jeong, Y. K., Heo, G. E., Kang, K. Y., Yoon, D. S., & Song, M. (2016). Trajectory analysis of drug-research trends in pancreatic cancer on PubMed and ClinicalTrials. gov. Journal of Informetrics, 10(1), 273-285.
- Lee, D., Kim, W. C., & Song, M. (2016, January). Exploring perceptual differences of experts and the public on diabetes. In 2016 International Conference on Big Data and Smart Computing (BigComp) (pp. 3-9). IEEE.

6. License policy

Competing interests
The authors declare that they have no competing interests.

Author's contributions
All authors have equal contribution.

Acknowledgements
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.

References

1. Song, M., Kim, W.C., Lee, D., Heo, G.E., Kang, K.Y.: Pkde4j: Entity and relation extraction for public knowledge discovery. Journal of biomedical informatics **57**, 320–332 (2015)
2. 2014 MEDLINE®/PubMed® Baseline Distribution. https://www.nlm.nih.gov/bsd/licensee/2014_stats/baseline_doc.html
3. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: a web-based tool for nlp-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 102–107 (2012). Association for Computational Linguistics
4. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: ACL (System Demonstrations), pp. 55–60 (2014)
5. Białecki, A., Muir, R., Ingersoll, G.: Apache lucene 4. In: SIGIR 2012 Workshop on Open Source Information Retrieval, p. 17 (2012)

Figures

Figure 1 Main page of YBio M^{2}CMT System

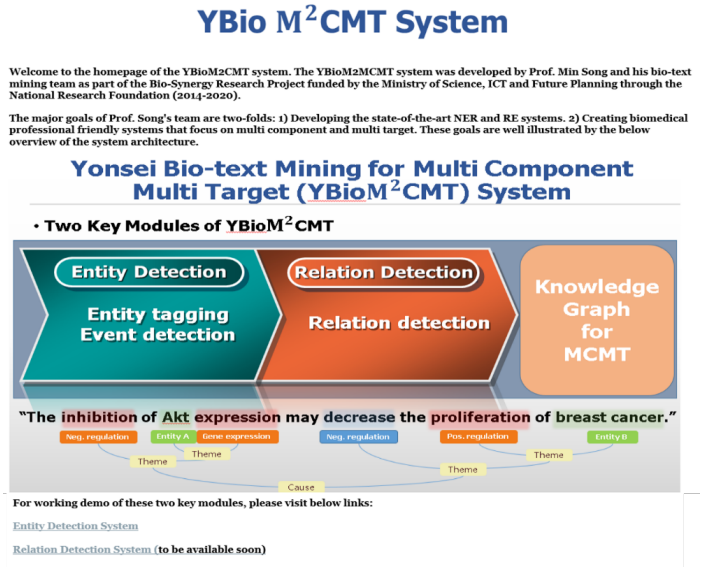


Figure 2 Overall Process of NER Annotation

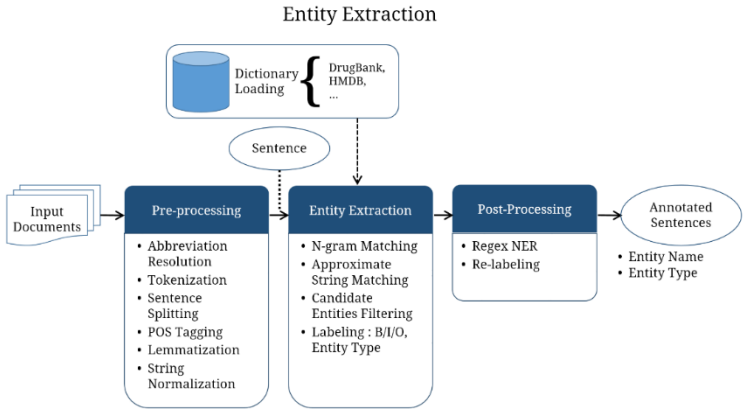


Figure 3 Main search Page of YBio M^{2}CMT NER System

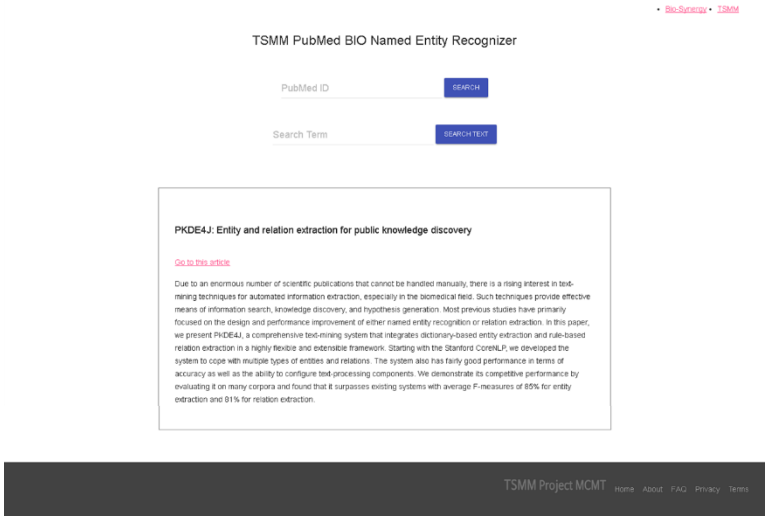
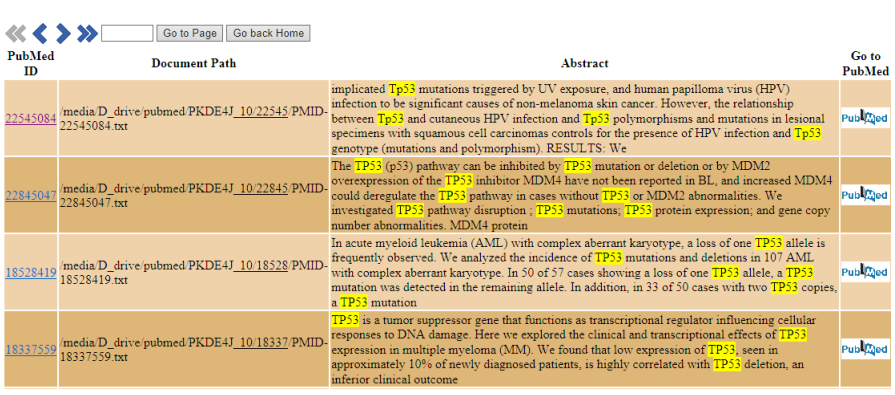


Figure 4 An example of the keyword search result page

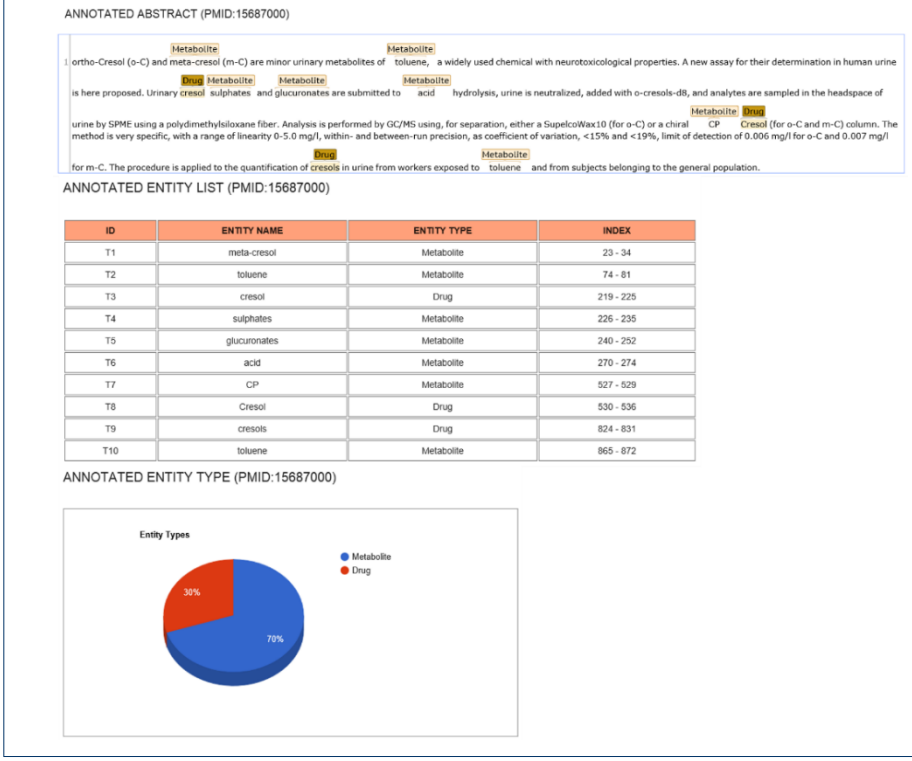


Search Results

Go to Page | Go back Home

PubMed ID	Document Path	Abstract	Go to PubMed
22545084	/media/D_drive/pubmed/PKDE4J_10/22545/PMID-22545084.txt	implicated Tp53 mutations triggered by UV exposure, and human papilloma virus (HPV) infection to be significant causes of non-melanoma skin cancer. However, the relationship between Tp53 and cutaneous HPV infection and Tp53 polymorphisms and mutations in lesional specimens with squamous cell carcinomas controls for the presence of HPV infection and Tp53 genotype (mutations and polymorphism). RESULTS: We	PubMed
22845047	/media/D_drive/pubmed/PKDE4J_10/22845/PMID-22845047.txt	The TP53 (p53) pathway can be inhibited by TP53 mutation or deletion or by MDM2 overexpression of the TP53 inhibitor MDM4 have not been reported in BL, and increased MDM4 could deregulate the TP53 pathway in cases without TP53 or MDM2 abnormalities. We investigated TP53 pathway disruption ; TP53 mutations; TP53 protein expression; and gene copy number abnormalities. MDM4 protein	PubMed
18528419	/media/D_drive/pubmed/PKDE4J_10/18528/PMID-18528419.txt	In acute myeloid leukemia (AML) with complex aberrant karyotype, a loss of one TP53 allele is frequently observed. We analyzed the incidence of TP53 mutations and deletions in 107 AML with complex aberrant karyotype. In 50 of 57 cases showing a loss of one TP53 allele, a TP53 mutation was detected in the remaining allele. In addition, in 33 of 50 cases with two TP53 copies, a TP53 mutation	PubMed
18337559	/media/D_drive/pubmed/PKDE4J_10/18337/PMID-18337559.txt	TP53 is a tumor suppressor gene that functions as transcriptional regulator influencing cellular responses to DNA damage. Here we explored the clinical and transcriptional effects of TP53 expression in multiple myeloma (MM). We found that low expression of TP53, seen in approximately 10% of newly diagnosed patients, is highly correlated with TP53 deletion, an inferior clinical outcome	PubMed

Figure 5 Example of Annotation Result page



ID	ENTITY NAME	ENTITY TYPE	INDEX
T1	meta-cresol	Metabolite	23 - 34
T2	toluene	Metabolite	74 - 81
T3	cresol	Drug	219 - 225
T4	sulphates	Metabolite	226 - 235
T5	glucuronates	Metabolite	240 - 252
T6	acid	Metabolite	270 - 274
T7	CP	Metabolite	527 - 529
T8	Cresol	Drug	530 - 536
T9	cresols	Drug	824 - 831
T10	toluene	Metabolite	865 - 872

Tables

Table 1 Example of text file

ortho-Cresol (o-C) and meta-cresol (m-C) are minor urinary metabolites of toluene, a widely used chemical with neurotoxicological properties. A new assay for their determination in human urine is here proposed. Urinary cresol sulphates and glucuronates are submitted to acid hydrolysis, urine is neutralized, added with o-cresols-d8, and analytes are sampled in the headspace of urine by SPME using a polydimethylsiloxane fiber. Analysis is performed by GC/MS using, for separation, either a SupelcoWax10 (for o-C) or a chiral CP Cresol (for o-C and m-C) column. The method is very specific, with a range of linearity 0-5.0 mg/l, within- and between-run precision, as coefficient of variation, <15% and <19%, limit of detection of 0.006 mg/l for o-C and 0.007 mg/l for m-C. The procedure is applied to the quantification of cresols in urine from workers exposed to toluene and from subjects belonging to the general population.

Table 2 Example of text file

T1	Metabolite	23	34	meta-cresol
T2	Metabolite	74	81	toluene
T3	Drug	219	225	cresol
T4	Metabolite	226	235	sulphates
T5	Metabolite	240	252	glucuronates
T6	Metabolite	270	274	acid
T7	Metabolite	527	529	CP
T8	Drug	530	536	Cresol
T9	Drug	824	831	cresols
T1	0 Metabolite	865	872	toluene