iTextMine: Integrated Text Mining System for Large-Scale Knowledge Extraction from Literature

Jia Ren[1], Gang Li[2], Cathy H. Wu[1,2]
[1]Center for Bioinformatics and Computational Biology [2]Department of Computer and Information Sciences
University of Delaware, Newark, Delaware, United States of America

***Abstract*—We describe the iTextMine system with an automated workflow to run multiple text-mining tools on large-scale text for knowledge extraction. We employ parallel processing for dockerized text mining tools with a common JSON output format, and implement a text alignment algorithm to align entity offsets in the text for result integration. iTextMine presently consists of four relation extraction tools and has processed all Medline abstracts. The website (http://research.bioinformatics.udel.edu/itextmine) allows users to browse the text evidence and view integrated results for knowledge discovery through a network visualization.**

Keywords—text-mining, relation extraction, text annotation, knowledge integration

I. INTRODUCTION

With the rapid growth of biomedical literature, text-mining tools help biologists extract useful information quickly. Most text-mining tools are specialized on specific tasks and may be used to recognize certain types of entities or relations. Thus, there is a need to combine results from different text-mining tools to cover a broad range of bioentities and relations for more comprehensive biological knowledge. Combining articles describing multiple relation types, one may identify cross-talk among different entities and relations extracted by different tools.

However, there are challenges in integrating different tools for large-scale processing and knowledge integration: (i) text-mining tools have different run-time dependencies and it is cumbersome to maintain them in the parallel execution engine. Meanwhile we need to make sure each tool can be run in parallel, e.g., two parallel processes will not write to the same file to avoid conflict; (ii) each tool has its own output format to describe the extracted information, and it is hard to store the result with the same database schema; (iii) some text-mining tools modify original text, and the text offset of entity and relation cannot be matched to the original text, making it impossible to directly compare the results from different tools.

Here we describe iTextMine to address the tool integration challenges. The system has been used for processing the entire set of Medline abstracts, combining and disseminating text-mining results from multiple tools developed in our group. Biologists can use this system to perform knowledge discovery via an interactive interface. The system will be run periodically to update the database when new literature is released.

II. METHOD

A. System Overview

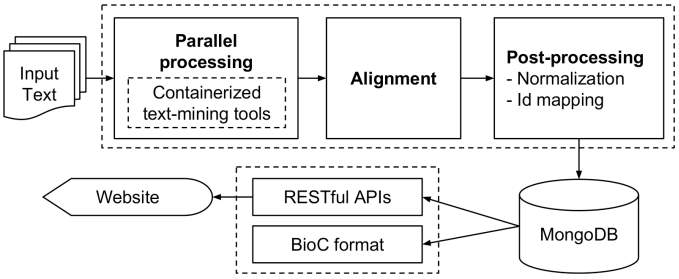

Fig. 1. iTextMine system with an automated workflow to integrate text mining tools and relation extraction results from large-scale text processing

The iTextMine workflow has four major steps (Fig. 1). We first prepare input literature by downloading Medline abstracts from PubMed website. The text is indexed by Lucene[1] on our local machine. Each tool uses a tool-specific entity/trigger word-based query to identify positive abstracts with potential entities or relations of interest. Then a parallel execution engine is set up to run dockerized text-mining tools in parallel. Before importing the text-mining results into the database, text alignment is performed to adjust entity offset. Additional post-processing tasks can be performed afterward, such as entity normalization and ID mapping. The post-processed data is then stored directly into MongoDB[2]. Finally, web services are built to disseminate the results. We created REST APIs to serve the data for web development. The results can be converted to other community standard dissemination formats, such as BioC (1) and brat standoff format[3].

B. Approaches to the Tool Integration Challenges

Dockerization: As text-mining tools may have different dependencies, we build a docker[4] image for each tool. This guarantees that the software is independent of the host machine and operating system, and can be run using a consistent command. Docker also creates an isolated environment for each running instance, allowing the tools to run in parallel without conflicting with each other.

Standardized JSON format: As different text-mining tools may use different output formats, we used one uniform JSON

[1] https://lucene.apache.org
[2] https://www.mongodb.com
[3] http://brat.nlplab.org/standoff.html
[4] https://www.docker.com

format for both input and output of the text-mining components. The basic schema is document-centric: each document contains a doc id field, text field, a list of properties, a hash-table of entities and a hash-table of relations. Each entity contains information such as entity type and offsets, while each relation contains relation type and its arguments.

Text alignment: As some text-mining tools may modify the original text during processing, we use Hirschberg's sequence alignment algorithm (2) to align the modified text and convert back to the original text. During the alignment process, only the character offset of an entity will be changed. After the alignment, the entity offsets in different text-mining results are based on the same text and ready to be merged.

III. Result

iTextMine currently consists of four in-house developed text-mining tools: (i) RLIMS-P (3) for mining protein phosphorylation (kinase-substrate-site), (ii) eFIP (4) for phosphorylation-dependent protein-protein interaction (PPI), (iii) miRTex (5) for miRNA-gene relation, and (iv) eGARD (6) for targeted therapy information from the scientific literature. For gene and other entity normalization, we incorporated results from PubTator (7).

For full-scale processing, we downloaded all Medline abstracts (June 2017) and ran the system pipeline to generate text-mining results. Table 1 summarizes the statistics of each tool—the number of positive abstracts, along with the counts of the specific relations types extracted. Overall, iTextMine identified 300,877 abstracts with at least one relation extracted by its underlying text mining tools.

TABLE I. iTextMine Medline abstract mining summary

Text Mining Tool	# Positive Abstracts		Relation types / counts
	Entities / Triggers	Entities + Relations	
RLIMS-P	289,258	264,163	phosphorylation (kinase-substrate-site): 454,389
eFIP	264,163	23,918	phosphorylation-dependent PPI: 38,814
miRTex	40,032	22,093	miRNA-target: 33,559 miRNA-gene regulation: 44,565 gene-miRNA regulation: 8,426
eGARD*	26,516	17,935	gene-disease-drug-response: 11,233

*The results are based on full-scale processing of 16.8 million abstracts for each tool, except eGARD which is still being processed with partial results only.

The website supports interactive query and display of text-mining results. User query to iTextMine will be sent as a query to PubMed to retrieve a list of PMIDs. Our system will then retrieve text-mining results from the database and generate a network using relations among entities. Redundant entities are merged if they are normalized to the same ID, or by the same text mention if not normalized.

We use the query "Triple negative breast cancer" as an example to demonstrate the integrated network with relations from RLIMS-P, eFIP, miRTex and eGARD (Fig 2). The interface also provides a document-centric view to display detailed text evidence (Fig 3).

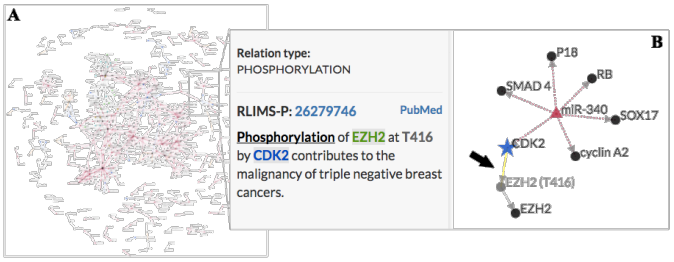

Fig. 2. Triple negative breast cancer network. A) The overall network contains 651 entities and 688 relations extracted by the four text-mining tools. B) Zoom in view. Different types of entity and relations are denoted by distinct node/edge colors and shapes. In this network, miR-340 (red triangle) regulates 5 genes (black circle) and a kinase (blue star). CDK2 phosphorylates EZH2 at Thr-416 and produces a proteoform (gray circle). By clicking the phosphorylation edge, the sentence describing the relation is shown.

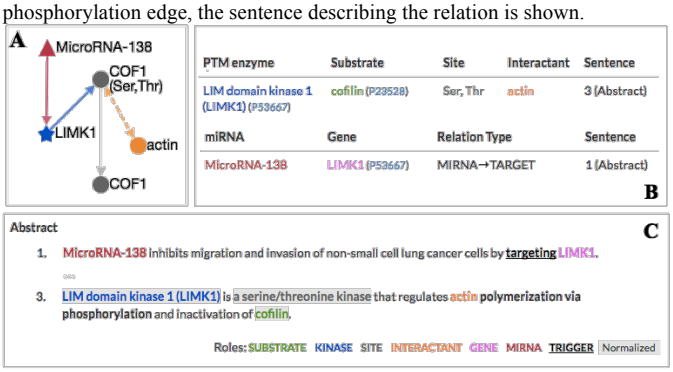

Fig. 3. Document-centric (PMID: 27665963) view. A) A network view summaries all entities and relations extracted from the abstract. RLIMS-P extracts a phosphorylation relation where kinase LIMK1 phosphorylates COF1; eFIP extracts PPI between the COF1 proteoform and actin; and miRTex extracts miRNA-gene regulation between MicroRNA-138 and LIMK1. B) Relation table lists relation arguments and attributes. C) Text evidence section displays the sentences with color-coded entities.

Acknowledgemet

This work was funded in part by grants from the National Institutes of Health (U01GM120953 and U01HG008390).

References

1. Comeau DC, Islamaj Dogan R, Ciccarese P, Cohen KB, Krallinger M, et al. (2013). BioC: a minimalist approach to interoperability for biomedical text processing. *Database* 2013, bat064.

2. Hirschberg DS. (1975). A linear space algorithm for computing maximal common subsequences. *Comm. ACM* 18, 341–343.

3. Torii M, Arighi CN, Li G, Wang Q, Wu CH, Vijay-Shanker K. (2015). RLIMS-P 2.0: A generalizable rule-based information extraction system for literature mining of protein phosphorylation information. *IEEE Transactions on Computational Biology and Bioinformatics* 12, 17–29.

4. Tudor CO, Ross KE, Li G, Vijay-Shanker K, Wu CH, Arighi CN. (2015). Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system. *Database* 2015, bav020.

5. Li G, Ross KE, Arighi CN, Peng Y, Wu CH, Vijay-Shanker K. (2015). miRTex: A Text Mining System for miRNA-Gene Relation Extraction. *PLoS Computational Biology*, 11(9), e1004391.

6. Mahmood AS, Rao S, McGarvey PB, Wu CH, Madhavan S, Vijay-Shanker K. (2017) eGARD: Extracting associations between genomic anomalies and drug responses from text. *PLoS One* (in press)

7. Wei C-H, Kao H-Y, Lu Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research* 41, W518–W522.