Retrieving evidence sentences for BEL statements

Majid Rastegar-Mojarad[1,2], **Ravikumar Komandur Elayavilli**[1], **Hongfang Liu**[1]
[1]*Department of Health Sciences Research, Mayo Clinic, USA*
[2]*University of Wisconsin-Milwaukee, Milwaukee, WI, USA*
Email: {Mojarad.Majid, komandurelayavilli.ravikumar, Liu.Hongfang}@mayo.edu

Abstract: In this paper, we describe our approach to retrieve and rank the evidence statements from PubMed abstracts and full text article for a given Biological Expression Language (BEL) statement towards a sub-task in BEL Track in BioCreative V. Our system comprises two main components, a) retrieving relevant Medline citations for the given BEL statement and b) finding and ranking the evidence sentences in those citations. Our system was able to rank at least one fully relevant evidence sentence in the top 10 retrieved sentences for 72 out of 99 BEL statements. The precision of our system, under full, relaxed, and context criteria, is 0.392, 0.532, and 0.615 respectively.

Keywords. BEL; Information retrieval; Biomedical literature mining; BioNLP

I. Introduction

Biological Expression Language (BEL) [1] and System Biology Markup Language (SBML) [2] are the two main formal representation models of biological network, which is a powerful and expressive in representing the biological information and knowledge. Biomedical literature has been the primary source of information for curating biological networks in BEL or SBML representation. As a measure to bridge the existing gap in the information between the literature and curated source, two subtasks were organized as part of BioCreative V challenge: 1) constructing BEL statements for a given sentence and 2) finding evidence sentences for a given BEL statement. Our group participated in both the tasks. While the first task [3] is more information extraction (IE) centric the second one, which is described in this paper is more centered on the principles of Information retrieval (IR).

The paper is organized as follows. First we discuss our approach to the second task and present the results of the system on the test dataset, which was evaluated manually by the organizers and briefly discuss the performance of the system.

II. Method

Our system has two main components, 1) Retrieve the relevant citations from PubMed abstract and full-text article from PubMed Central (PMC) and 2) Identify the appropriate evidence statements (at most 10) from the citations and rank their relevance to the provided BEL statement.

Retrieval Component

Figure 1. illustrates the outline of the retrieval component of the system. The system translates the given BEL statement into a query by expanding all the individual elements (e.g. entity, function, and relation) expressed in the statement. For an example BEL statement "p(MGI:Tnf) increases p(MGI:Creb1,pmod(P,S,133))" the system will construct a query by including the synonyms of entities 'Tnf', creb1, and "phosphorylat" as a representation for the function "pmod(P)". The query will be augmented by using the appropriate boolean logic between the components.

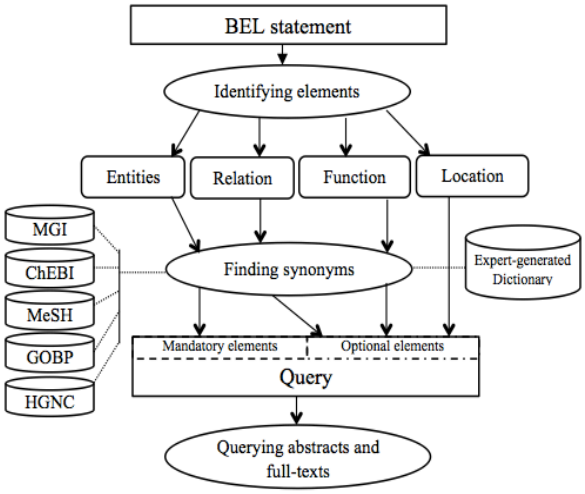

***Figure 1*: The retrieval component**

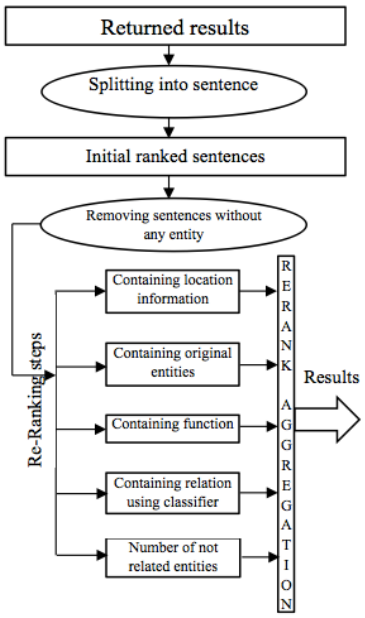

***Figure 2*: The ranking component**

We used the knowledge resources such as Mouse Genome Informatics (MGI) [4], Chemical Entities of Biological Interest (ChEBI) [5], Gene ontology (GOBP) [6], Medical Subject Headings (MeSH) [7], and HUGO Gene Nomenclature Committee (HGNC) [8] for augmenting the query with entity synonyms. For expanding the functions and relations, we created a list of related verbs, their nominalized forms and other synonyms based on the domain expert knowledge. The expanded query is searched against PubMed abstracts and PMC full-texts to retrieve the relevant citations.

Ranking Component

Figure 2 outlines the individual steps in the ranking component of our system. After retrieving the top 1000 citations relevant to the given BEL statement, the system further extracts and rank the sentences that are most relevant to the expanded query. i) First we consider only those sentences that contain all the entities (or their synonyms) mentioned in the query. We trained a binary classifier to classify the sentences into two classes, increase and decrease (two main types of BEL relations) based on the functions and relations represented in the query. For training the classifier, we used the training data provided by the organizers which contain 11,072 BEL statements extracted from 6,358 sentences. We used uni-grams (after removing stop words), bi-grams, entities in the

sentence, and part-of-speech tags of all words between the entities (mentioned in the query) as features. Using 10-fold cross validation, we compared several learning models such as Support Vector Machine (SVM), Naïve Bayes, and Random Forest, on the training dataset. We used the classifier results as one of the parameters for ranking the evidence sentences. The ranking function is directly proportional to the number of elements matched by the evidence sentences in the query. The sentences that matched the maximum components of the query (e.g. entities, functions and relations) are ranked higher than those sentences that matched only a few components.

III. Results

For the blind evaluation, the organizers provided 99 BEL statements and the assigned the task of identifying at most 10 evidence sentences for each statement to the participants. The sentences were manually evaluated for the relevance by the organizers. Table 1 shows the performance of our system under three evaluation criteria, which provided by the organizers. The organizers evaluated the performance on three criteria. i) *Full*: if the identified sentence contains the complete BEL statement. ii) *Relaxed*: The retrieved sentence may not all the evidences for extracting the complete BEL statement. However it may have necessary context and/or biological background to enable extraction of full BEL statement. *Context*: Even though the complete or partial BEL statement cannot be extracted from the sentence, it provides the necessary context for the BEL statement. The entities or their synonyms mentioned in the BEL statement are also identified in the sentence but the context description (function or the relations) may not acurately reflect the actual BEL statement.

TABLE 1: THE SYSTEM PERFORMANCE

Criteria	True positive	False positive	Precision
Full	316	490	0.3920
Relaxed	429	377	0.5322
Context	496	310	0.6153

IV. Discussion

Out of the 99 BEL statements provided in the test data, our system retrieved at least one evidence statement for 98 of them. The performance of the system was best under context criterion. The system was able to retrieve at least one relevant evidence for 81 out of 99 statements. The significant difference in the performance of the system (61% vs 39%) between the Context and Full precision indicate the need of sophisticated evaluation considering the underlying semantics of BEL statements and the retrieved sentences. In the current approach, we extensively relied upon the lexical feature without much consideration for the semantics. We firmly believe that enriching the lexical capabilities of the system with deeper semantic analysis will significantly improve the precision of the system. We plan to achieve this by integrating the information extraction capabilities (system in task 1 of BEL track) into our IR workflow. As an immediate next step, we first plan to process all the evidence statements retrieved by our IR system through the IE engine. Based on the semantic distance between the BEL statements extracted by the IE

system and gold standard we will be able to improve the ranking of the evidence statement. The only statement for which our system failed to find any evidence sentence was "*p(HGNC:IL1A) increases r(HGNC:DEFB4A)*". Preliminary analysis indicates that our IR capabilities may be limited when compared to the Entrez retrieval engine. We also plan to assess the improvements by using the Entrez retrieval engine for IR.

Reference

[1] J. Szostak, S. Ansari, S. Madan, J. Fluck, M. Talikka, A. Iskandar, H. De Leon, M. Hofmann-Apitius, M. C. Peitsch, and J. Hoeng, "Construction of biological networks from unstructured information based on a semi-automated curation workflow," *Database (Oxford)*, vol. 2015, 2015.

[2] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, and SBML Forum, "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, Mar. 2003.

[3] R. Komandur Elayavilli, M. Rastegar-Mojarad, and H. Liu, "Challenges in adapting a rule based information extraction system to the demands of BioCreative V BEL task," in *BioCreative V*, Spain, 2015.

[4] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, and Mouse Genome Database Group, "The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D726–736, Jan. 2015.

[5] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck, "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D456–463, Jan. 2013.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000.

[7] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucl. Acids Res.*, vol. 32, no. suppl 1, pp. D267–D270, Jan. 2004.

[8] K. A. Gray, B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford, "Genenames.org: the HGNC resources in 2015," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D1079–1085, Jan. 2015.