BELMiner – Information extraction system to extract BEL relationships

Ravikumar Komandur Elayavilli[1], Majid Rastegar-Mojarad[1,2], Hongfang Liu[1]
Affiliation: [1]Department of Health Sciences Research, Mayo Clinic, USA
[2]University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Abstract— **In this work, we describe an improved version of BELMiner that extracts Biological Expression Language(BEL) statements from evidence sentences provided for BioCreative BEL 2017 task. The current system builds upon the basic infrastructure of BELMiner; a system was developed to extract BEL relationships from evidence sentences for the BioCreative BEL 2015 task. Training a state of the art machine learning NER architecture using a pooled corpus from diverse shared tasks and marshalling their results with other existing state of the art NER tools, incorporating the latest Stanford parser 3.8 textual entailment functionalities, graph based traversal to extract relations, handling double negations are some of the new functionalities that we incorporated in BELMiner. The system achieved an overall F-measure of 49.6% with gold standard entities, while it achieved a lower performance of 32.45% with the entities extracted by an ensemble of NER systems on blind test data. For relation extraction, the system achieved an F-measure of 66.83% on a blind test data set with gold standard entities. We observe a significant improvement in the state of the art performance in BEL statement extraction by over 14% when compared to the performance of the best system on 2015 data set.**

Keywords—BEL; BELMiner; graph based relation extraction; biomedical information extraction

I. INTRODUCTION

Biomedical literature has been a rich resource for information on biological pathways. Tapping into the information in the literature on signaling pathways is of great importance that will help bridge the gap in the curation of biological pathways. One approach is to crowdsource the human curation of biological events and pathways through intuitive and effective user interfaces such as PubTator [1], BELIEF [2] and other similar tools. This curation initiative can be substantially augmented through text mining efforts.

Formal representation of textual assertions in biological literature is important to fully realize any of the curation and text mining initiatives. There has been considerable attention on Biological Expression Language(BEL) [3] by system biologists in the recent past. It is one of the suitable representation languages to formalize signaling pathways from biomedical literature. BioCreative shared task organizers organized a very important task involving formalizing the relation extracted from biomedical text in BEL framework in 2015 [4]. Multiple teams participated in the task and the state of the art performance achieved F-measure of 35% when the system used the gold standard entities. Our system BELMiner [5] achieved an F-measure of 25.6%. As a continued effort, this year additional test data was provided to validate the ability to extract BEL statements from evidence sentences. In this paper, we describe the performance of an improved version of BELMiner, which we call BELMiner 2.0 on the BioCreative 2017 BEL extraction task. In the following sections, we briefly describe the various improvements we made to the BELMiner, its performance and brief error analysis. Due to space constraints, we provide only a brief description of our system.

II. SYSTEM DESCRIPTION

Figure 1 outlines the overall architecture of BELMiner 2.0. The system consists of the following components executed in sequence to process the evidence statement, with each component incrementally contributing towards BEL statement extraction. 1) Extraction of normalized entities 2) Identify dependency structure 3) graph based traversal to extract causal relationships 4) Formalization of causal relations into BEL statements and 5) Filter out irrelevant BEL statements.

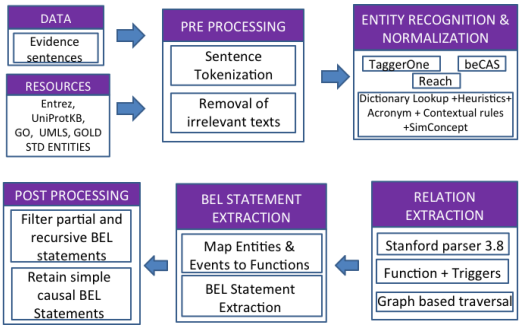

Fig. 1. BELMiner architecture.

We briefly describe the individual steps of BELMiner 2.0 below. In Figure 2 we have illustrated the important steps of BELMiner 2.0 using an example sentence.

A. *Preprocessing*

BELMiner 2.0 pipeline starts with some simple pre-processing such as sentence tokenization. We also have few simple regular expressions to trim unnecessary phrases like "(Fig 2A)", "Figure 1"" from the sentences to avoid errors in the subsequent steps.

B. *Entity recognition and normalization*

In BELMiner, we assembled an ensemble of state of the art entity normalization tools such as PubTator and beCAS [6], further supplemented by dictionary based lookup [7] to

extract normalized entities from the evidence sentences provided for the task. Subsequent analysis of the impact of different NER components showed certain gaps in the NER tools. Most of these tools are trained on the specific corpus and hence do not generalize well for the BEL NER task. In this work, we used TaggerOne [8], a trainable semi-Markov structured linear classifier for any entity types. It also uses supervised semantic indexing for entity normalization. Its unique feature is its ability to jointly model NER and normalization.

We made the following changes to the core components of the TaggerOne framework. First, we replaced the Ab3P acronym detection component with the algorithm of Schwartz and Hearst algorithm [9]. We did joint training for the gene, chemical, and disease entity detection and normalization using diverse corpus from prior BioCreative shared task data sets [10]. While the TaggerOne infrastructure had inherent support for Chemical and Disease entity normalization we did additional implementation to extend its functionality for gene normalization. We used internally generated lexicon integrated from different sources such as UniProtKB [11], Entrez Gene [12], HGNC [13] and MGI [14]. We supplemented the annotations provided by TaggerOne with annotations from other external tools such as beCAS and Reach [15]. Our strategy was similar to the one described in the earlier paper; to build consensus across multiple NER tools for the gene, chemical, and diseases. For Gene ontology term normalization we used dictionary look-up approach where we used only the terms given by the task organizers. In addition to the terms, we included their synonyms from GO ontology.

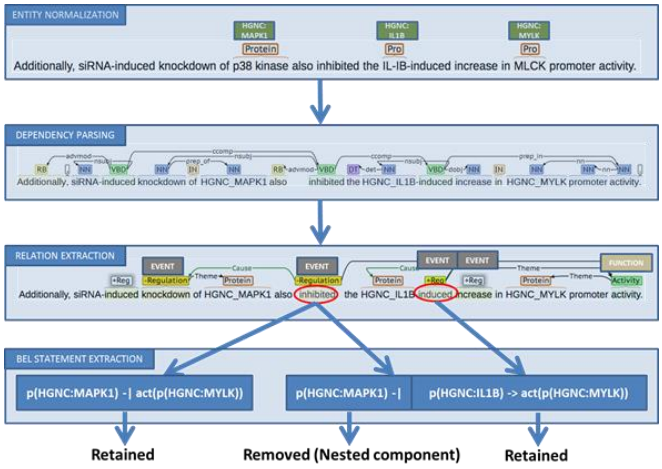

Fig. 2. Illustration of the function of individual components of BELMiner using an example sentence (PMCID: PMC3677168)

In the biomedical text, the presence of coordinated entities is quite common. For example, in the phrase "ERK1 and -2" while it is easily possible to detect ERK1 as gene the second entity ERK-2 is difficult to identify unless the entities are decomposed into two entities ERK1 and EKR-2. We integrated SimConcept [16] into the current BELMiner 2.0 pipeline in order to distinctly identify the composite entities and normalize them to the right database identifiers. Composite names and entity co-ordinations are common in bio-medical literature and their resolution has been shown to improve the performance of gene/disease detection and normalization by 1% [16].

C. *Improved syntactic and semantic features in BELMiner 2.0*

One of the shortcomings in the BELMiner was that its ability to identify long distance relations that occur beyond clausal boundaries are limited. In order to overcome this issue, we used the new Stanford Parser 3.8 [17] that identifies extended dependencies such as anaphora/co-reference resolution, appositives and dependencies that occur beyond clausal boundaries with greater accuracy. Subsequently, we used a list of trigger words to identify the functions and relations (increase or decrease) to identify them in the text.

Our approach to identifying appropriate arguments for function and relations involve a graph based traversal [18]. For example, consider an example sentence shown in Figure 2. The system identified *induced*, *knockdown*, *inhibited* and *increase* as the trigger words that describe bio-medical events. Similarly, the term *activity* was identified as a "function". From each of the event phrases, the relation extraction module traverses along the dependency graphs until it reaches an entity or a function word. During graph traversal, the system considers certain properties of the node such as the semantic type of the node and properties such as negations. It also identifies the successive double negation of events along the traversal path to correctly identify the type of the main event. Consider the phrase "siRNA-induced **knockdown** of ***p38 kinase*** also **inhibited**" in the example sentence shown in Figure 2. There is a possibility to mis-identify "inhibited" as a negative regulation by just considering that word alone. The event "inhibited" takes "knockdown" as the causal argument, which in turn takes entity "p38 kinase" as its argument. Since both events belong to the Negative_Regulation class, the double negation classifies the type of the main event "inhibited" as "Positive_Regulation" class. In this case, the system identifies the correct relation "p(HGNC:MAPK1) ->".

TABLE I. MAPPING NLP FUNCTION AND EVENT TYPES TO BEL FUNCTIONS

Keywords	Type	BEL functions
Physical interaction, binding, complex formation	Function	complex
Gene expression, Transcription	Function	rnaAbundance()
Translocation	Function	tloc()
Phosphorylation, acetylation, methylation	Function	pmod(P/A/M)
Degradation,	Function	deg()
Activity	Function	act()
increases, activation, induce	Event (Relation)	increases

inhibit, block	Event (Relation)	decreases
Directly increases, direct activation, directly induce	Event (Relation)	directlyIncreases
Directly inhibit, directly block	Event (Relation)	directlyDecreases

D. Mapping textual relations to BEL statements

We retained a simple rule-based approach to map the relation extraction output to a formal BEL statement that we have described in our earlier paper [5]. Table 1 lists the mapping between NLP functions and events into BEL functions. We finally filter out the incomplete BEL statements that do not fit the syntax of BEL formalism and recursive BEL statements, as they were not considered for the evaluation in this task.

III. RESULTS AND DISCUSSION

We used the training and test data for BioCreative V BEL track task 1 from 2015 [4] for fine tuning the performance of updated BELMiner. We ran the new BELMiner 2.0 on the test data provided as part of BioCreative BEL task 2017. Similar to the 2015 tasks the performance of systems was performed at different levels namely, Term-Level, Function-Level, Relationship-Level, Full Statement and Overall Evaluation. It was further carried out in two phases i) without named entities and after providing the gold standard entities. Table 2 outlines the results of the system for both runs of the system with and without the gold standard entities respectively. We submitted two runs of BELMiner output during Phase I and Phase II. The only difference between the two runs is that the first run consists of partial BEL statements where the missing entities (arguments) are replaced with PH:Placeholder. The standard metrics namely Precision, Recall, and F-measure was used to evaluate the performance of the system.

TABLE II. PERFORMANCE OF BELMINER ON BIOCREATIVE BEL TASK 2017 (WITH AND WITHOUT GOLD STANDARD ENTITIES)

Class		Entities from Gold standard			Entities from NER		
		Pre (%)	**Rec (%)**	**F-Mes (%)**	**Pre (%)**	**Rec (%)**	**F-Mes (%)**
Term (T)	R1	90.56	**84.92**	**87.65**	78.18	**70.49**	**74.14**
	R2	**90.94**	82.3	86.40	**78.71**	67.87	72.89
Function Secondary (FS)	R1	79.07	**45.33**	57.63	70.83	**45.33**	**55.28**
	R2	**89.19**	44.0	**58.93**	**79.49**	41.33	54.39
Function	R1	77.08	**38.95**	51.75	56.6	**31.58**	**40.54**
	R2	**85.71**	37.89	**52.55**	**63.64**	29.47	40.29
Relation-Secondary (RS)	R1	95.75	**88.65**	**92.06**	89.62	**82.97**	86.17
	R2	**97.55**	86.99	91.92	**93.00**	81.22	**86.71**
Relation	R1	**76.70**	**59.21**	**66.83**	51.81	**37.72**	43.65
	R2	**76.70**	**59.21**	**66.83**	**52.12**	**37.72**	**43.77**
BEL Statement	R1	63.01	**40.35**	49.2	40.67	**26.75**	32.28
	R2	**64.34**	**40.35**	**49.6**	**41.22**	**26.75**	**32.45**
Pre – Precision; Rec – Recall; F-Mes – F-Measure R1 – Run1; R2 – Run2; R3 – Run3							

A. Term level performance

Figure 3 and 4 outlines the performance of the new BELMiner during Phase I and Phase II, respectively. Comparison between the first and second phase between the best systems reveal a 12% difference between Phase I and Phase II. The native BELMiner identified 275 entities in total out of which only 215 were found to be correct when compared against the 305 gold standard entity annotations. With gold standard entities, the system extracted 259 entities correctly. Out of the 305 gold standard annotations, there were 247 protein, 36 chemical, 20 biological process, and 2 diseases annotations. During Phase I the best system extracted 169 protein annotations (F-mes: 72%), while we observed a steep increase during Phase II. We did not observe much of a difference between two phases for chemical entity detection. We saw only 0.5% increase in the overall F-measure for chemical detection. There is a substantial increase in the biological process performance (11% increase) between Phase I and Phase II. The total number of protein annotations was more than the other two and they might have reflected in the overall numbers.

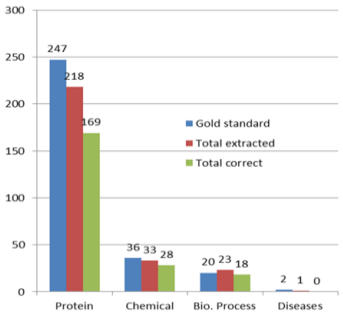

Fig. 3. Term (category wise) level evaluation during Phase I.

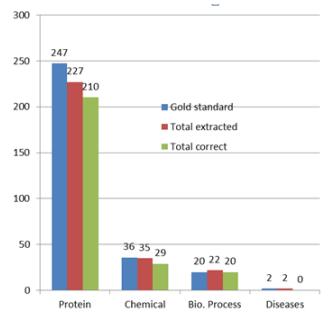

Fig. 4. Term (category wise) level evaluation during Phase II.

B. Performance of BELMiner in identifying functions

The ability of the BELMiner in identifying the functions (secondary functions) seems to be its key bottleneck. There were 75 functions annotated in the gold standard out of which the system identified only 34 correctly. There were totally 6 functional categories namely, activity, complex formation, secretion, protein modification, degradation, and translocation. The performance seems to be the lowest among the activity (F-Mes: 24%). The system really performs well in identifying the post-translational modification (F-mes: 92.31%). We haven't completed our error analysis to understand how to improve the system better in its ability to identify the functions in the text. Figure 5 outlines the performance of BELMiner on secondary functions belonging to different categories.

C. Performance of BELMiner in extracting the relations

The difference in the performance of the system in extracting the relations during Phase I and Phase II is very huge. We observed a huge increase in 23% overall F-measure when we used the gold standard entities for named entities. The increase in the NER performance directly

correlates with the relation extraction. However, the evaluation at secondary functions level revealed that we did not observe a huge difference (increase in 5%) between the two phases.

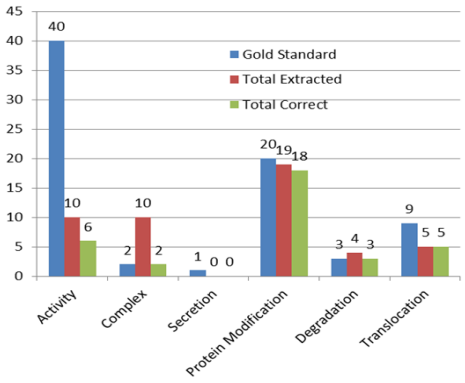

Fig. 5. Secondary function evaluation

D. *Performance of BELMiner in extracting the overall BEL statements*

The trend in complete BEL statement extraction is similar to that of relation extraction between the two phases. We observed nearly a 17% increase in the performance of BEL statement extraction between the two phases. Compared to the performance of BEL statement extraction in 2015, we observed nearly a 15% overall improvement in the performance of BEL statement extraction.

IV. CONCLUSIONS

In this work, we described a generic graph-based traversal on the dependency graphs constructed out of entity normalized sentences. We discussed the overall impact of different changes to the BELMiner on different tasks in BEL statement extraction on BioCreative BEL 2017 extraction task. We observed significant improvement in the state of the art in BEL statement extraction. Similar to the experience we had last time the performance of term level extraction played a major role in the overall improvement of BEL statement extraction. We believe that we can further improve the performance of the system through systematic analysis of the errors.

ACKNOWLEDGMENT

We acknowledge funding from National Institutes of Health (NIH) , grant number NIH-NCATS: 1OT3TR002019-01, "Biomedical Data Translator Technical Feasibility Assessment and Architecture Design", and Mayo clinic Research Foundation that supported this work.

REFERENCES

1. Wei, C.H., Kao, H.Y. and Lu, Z., *PubTator: a web-based text mining tool for assisting biocuration.* Nucleic acids research, 2013: p. gkt441.
2. Madan, S., Hodapp, S., Senger, P., Ansari, S., Szostak, J., Hoeng, J., Peitsch, M. and Fluck, J., *The BEL information extraction workflow (BELIEF): evaluation in the BioCreative V BEL and IAT track.* Database, 2016(Jan 1).
3. Fluck, J., Madan, S., Ellendorff, T.R., Mevissen, T., Clematide, S., van der Lek, A. and Rinaldi, F., *Track 4 Overview: Extraction of Causal Network Information in Biological Expression Language (BEL).* 2015: p. 333-346.
4. Fluck, J., Madan, S. and Ansari, S., *Training corpora for the extraction of causal relationships coded in Biological Expression Language (BEL).* . Database, 2016(This issue).
5. Ravikumar, K.E., Rastegar-Mojarad, M., and Liu, H., *BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences.* . Database, 2017. **1**(baw156).
6. Nunes, T., et al., *BeCAS: biomedical concept recognition services and visualization.* Bioinformatics, 2013. **29**(15): p. 1915-1916.
7. Torii, M., Hu, Z., Wu, C.H. and Liu, H., *BioTagger-GM: a gene/protein name recognition system.* Journal of the American Medical Informatics Association, 2009. **16**(2): p. 247-255.
8. Leaman, R., and Lu, Z., *TaggerOne: joint named entity recognition and normalization with semi-Markov Models.* . Bioinformatics, 2016. **32**(18): p. 2839-2846.
9. Schwartz, A.S.a.H., M. A. *A simple algorithm for identifying abbreviation definitions in biomedical text*. in *Pacific Symposium on Biocomputing*. 2003. Hawaii.
10. Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J. and Sun, C., *Overview of BioCreative II gene normalization.* Genome biology, 2008. **9**(Suppl 2): p. S3.
11. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. and Martin, M.J., *UniProt: the universal protein knowledgebase.* . Nucleic acids research, 2004. **32**(suppl 1): p. D115-D119.
12. Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T., *Entrez Gene: gene-centered information at NCBI.* . Nucleic acids research, 2005. **33**(suppl 1): p. D54-D58.
13. Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., and Wain, H., *The HUGO gene nomenclature committee (HGNC).* . Human genetics, 2001. **109**(6): p. 678-680.
14. Bult, C.J., Eppig, J. T., Kadin, J. A., Richardson, J. E., Blake, J. A., and Mouse Genome Database Group, *The Mouse Genome Database (MGD): mouse biology and model systems.* Nucleic acids research, 2008. **36**(suppl_1): p. D724-D728.
15. Hahn, M.A.V.E.G., and Surdeanu, P. T. H. M. *A domain-independent rule-based framework for event extraction.* . in *ACL-IJCNLP*. 2015.
16. Wei, C.H., Leaman, R., and Lu, Z. *SimConcept: a hybrid approach for simplifying composite named entities in biomedicine*. in *5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2014.
17. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. *The stanford corenlp natural language processing toolkit*. in *ACL (System Demonstrations)*. 2014.
18. Ravikumar, K.E., Wagholikar, K. B., Li, D., Kocher, J. P., and Liu, H., *Text mining facilitates database curation-extraction of mutation-disease associations from Bio-medical literature.* . BMC bioinformatics, 2015. **16**(1): p. 185.