Adapting a rule-based relation extraction system for BioCreative V BEL task

Ravikumar Komandur Elayavilli[1], **Majid Rastegar-Mojarad**[1,2], **Hongfang Liu**[1]

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

We tested a rule-based semantic parser in the BEL statement extraction task of BioCreative V Track4 challenge. While the system achieved an overall F-measure of 21.29% with gold standard entities, it achieved a very low performance of 13.86% with the entities extracted by ensemble of NER systems on test data set. For relation extraction, the system achieved a F-measure of 65.13% on test data set. The limitation in the rule sets to map the textual extractions to BEL function is one of the reasons for our low performance in extracting the complete BEL statement. Besides, the lack of ability to extract long distance relationships, recursive relations and the inability to make certain semantic inferences had significant influence on the overall performance of the system.

Introduction

Elucidation of biological pathway events involving drugs, proteins and diseases through extraction of knowledge from the scientific literature is one of the interesting challenges in biomedical text mining. The automatic extraction of such events will provide insights into the underlying molecular mechanisms of biological macro-molecular interactions and pharmacological dynamics. Despite multiple knowledge acquisition efforts to catalog biological events in databases, a considerable amount of knowledge is still buried in the scientific literature.

On the other hand, there has been considerable work on developing representation standards for pathway information such as Systems Biology Ontology (SBO) (1), Biological pathway exchange language (BioPAX) (2), and Systems Biology Mark Up Language (SBML) (3). There is an urgent need to have a standard representation language to formalize the textual extractions from scientific articles. Such representation language may serve as an intermediate link between the text extractions and the pathway representation standards. Biological expression language (BEL) (4), has been of considerable attention to system biologists in the recent past and it is one of the suitable representation language to formalize the textual extractions.

At an appropriate time, the current BioCreative BEL task has been organized, which involves formalizing the relation extracted from biomedical text in BEL framework. In this paper, we describe the challenges in adapting a rule-based information extraction system (5) in formalizing the biological events extracted from the sentences provided for the task in BEL framework.

System description

Extraction of normalized entities

We used an ensemble of entity normalization tools namely PubTator (6), beCAS (7) and an in-house developed dictionary based lookup (8) to develop a high precision entity extraction and normalization system. For both PubTator and beCAS, we used the REST-API services provided by the respective tools. We also used a number of heuristics to identify certain terms in the biological process. For genes/proteins, chemical and disease names, we preferred the annotations of PubTator to that of beCAS. However, if PubTator has failed to annotate certain genes, chemicals or diseases then we considered the consensus between beCAS and the dictionary based lookup. For detecting Gene ontology (GO) (9) terms, we relied only on beCAS and the dictionary lookup from GO.

Correction of named entity with gold standard entities (Phase 2 submission)

During phase 2, the organizers provided the gold standard for all the entities. First we limited the dictionary lookup to only those entries that were provided in the gold standard. Second, if there is a difference between the gold standard entity and the NER entity tagger system we replaced the entities with the gold standard.

Extraction of biological events

We used a rule-based semantic parser (5), which can handle discourse connectives, entity and event anaphora for effective synthesis and extraction of events both within and across sentences. The system besides identifying the syntactic arguments of verbs also assigns thematic roles. The frame-based semantic rule templates contain nearly 15 verb categories (including causal verbs) and more than 70 verbs and their inflections. While the semantic parser begins as a linear parse, it builds upon its linear structure to handle complex structures such as appositions, selective prepositional phrase attachment, and co-ordinations (include entities, events and clausal). It highly depends on semantic information while linking events across clausal boundaries. Due to space constraints we have given only the summarized description of the system.

Mapping Semantic parser output to BEL annotations

Mapping the extraction output of semantic parser was done at two levels. 1) Mapping certain biological events of the semantic parser to BEL functions. 2) Mapping causal relations (decreases, increases, directlyIncreases and directlyDecreases) that connect BEL functions to complete BEL statements. Table 1 lists some of the examples of how we map the NLP system extractions to BEL functions.

Table1 – Mapping NLP system output to BEL functions

Event/Entities	BEL function
phosphorylation of **PDE3B** on **serine-273**	p(HGNC:PDE3B,pmod(P,S,273))
translocation of **HSF1**	tloc(p(HGNC:HSF1))
expression of **ICAM-1**	act(p(HGNC:ICAM1))
truncal **obesity**	path(MESHD:Obesity)
interaction of **cyclin A1** with **E2F-1**	complex(p(HGNC:CCNA1),p(HGNC:E2F1))
cyclin A1 was complexed with **CDK2**	complex(p(HGNC:CCNA1),p(HGNC:CDK2))
glycerol kinase enzymatic activity	act(p(HGNC:GK))
activates STAT3	increases act(HGNC:STAT3)

There are certain events extracted by the system that could not be directly translated to BEL functions. For example consider the example sentence (SEN: 10000052; 10409724) "**Signaling by the IL-6 receptor** is mediated through the signal transducing subunit gp130 and involves the **activation of Janus-associated kinases (JAKs), signal transducer and activator of transcription 3 (STAT3), and mitogen-activated protein (MAP) kinase.**". The system extracts "signaling(HGNC:IL6-R) activates(HGNC:STAT3) as one of the relations, where the "signaling(HGNC:IL6-R)" doesn't translate to any BEL function. Such functions are dropped from the final BEL statement by further simplifying the statement to p(HGNC:IL6-R), which is consistent with BEL syntax.

After extracting the complete BEL statements we filter out BEL statements, that do not contain four classes of causal verbs namely, increase, decrease, directly increases or directly decreases. We made two submissions for the BEL task. Whenever we did not have high confidence in a namespace we replace them with PH:placeholder so that they are not considered as precision error. In the first submission (Run1 in Table 2), we retained BEL statements that contains PH:Placeholder while in the second submission (Run2 in Table 2), we filtered all the statements containing the placeholder statements altogether.

Results and discussion

In the BioCreative V BEL task, the evaluation was carried at five 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. The system was evaluated on standard metrics namely Precision, Recall and F-measure.

Table 2 – Performance of NLP system on BioCreative BEL task (with and without gold standard entities)

Class		**Entities from Gold standard**			**Entities from NER**		
		Pre (%)	**Rec (%)**	**F-Mes (%)**	**Pre (%)**	**Rec (%)**	**F-Mes (%)**
Term (T)	Run1	91.8	74.67	82.35	82.03	59.33	68.86
	Run2	92.51	70.00	79.70	83.33	50.00	62.5
Function-Secondary (FS)	Run1	51.47	62.50	56.45	50.77	58.93	54.55
	Run2	51.61	57.14	54.24	54.72	51.79	53.21
Function	Run1	25.53	36.36	30.00	27.78	37.88	32.05
	Run2	27.06	34.85	30.46	30.67	34.85	32.62
Relation-Secondary (RS)	Run1	**87.71**	**77.72**	**82.41**	**76.84**	**67.33**	**71.77**
	Run2	**94.38**	**74.75**	**83.43**	**92.37**	**59.9**	**72.67**
Relation	Run1	**77.93**	**55.94**	**65.13**	**69.37**	**38.12**	**49.20**
	Run2	**77.93**	**55.94**	**65.13**	**69.37**	**38.12**	**49.20**
Statement	Run1	32.09	21.29	25.60	26.42	13.86	18.18
	Run2	32.09	21.29	25.60	26.42	13.86	18.18

TP – True positives; FP – False positives; FN – False Negatives; Pre – Precision; Rec – Recall; F-Mes – F-Measure;

From Table 2, we can infer that the performance of the NLP system (row 6 of Table2) in extracting a complete BEL statement is very low. Using gold standard entity instead of our ensemble of NER system resulted in significant improvement in the overall F-measure (nearly 7%). Very low performance of BEL statement extraction is not surprising given that the performance of the system in extracting the BEL function (row 3 of Table 2 and 3) is only 32%. Mapping textual extractions to BEL function is the performance-limiting step of our NLP system. The performance of Function-Secondary (FS) extraction is higher in mid 50%, which may be due to the reason that the system is capable of correctly extracting simpler functions involving entities, while its ability to extract recursive (or) complex functions is limited. On the other hand, the NLP system performs well in identifying the core relation, which is very evident from the row 4 and 5 of Table 2. Being a rule-based system, the precision is reasonably higher (in the mid 70s to 80%) with a reasonable recall. In the Phase 2 evaluation, the performance of relation extraction (Row 4, Column3 of Table2) is even higher with a significant gain of nearly 13% over phase 1. From the above results, we can infer that entity recognition and normalization do have positive influence on relaxation extraction capability.

Conclusion

In this work, we discussed the challenges of a rule-based information extraction system to BioCreative BEL extraction task. The system though achieved very low performance in BEL statement extraction and function detection; it achieved a very balanced performance in the relation extraction task. Lack of rule sets to map textual extraction to BEL formalism, named entity recognition and normalization, lack of methods to infer deeper biological semantics were some of the main for the lower performance of the system. With some fine-tuning we believe that we can address some of the errors of this system to further improve the performance of the system.

References

1. Le Novère N (2006) **Model storage, exchange and integration**. *BMC neuroscience*, **7**(Suppl 1):S11.
2. Demir E, Cary, M.P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J. (2010) **The BioPAX community standard for pathway data sharing**. *Nature biotechnology*, **28**(9):935-942.
3. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A. (2003) **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics*, **19**(4):524-531.
4. **Biological Expression Language** [http://www.openbel.org/]
5. Ravikumar KE, Wagholikar KB, Liu H (2014) **Towards Pathway Curation Through Literature Mining–A Case Study Using PharmGKB**. *Pacific Symposium on Biocomputing*:352.
6. Wei C-H, Kao H-Y, Lu Z (2013) **PubTator: a web-based text mining tool for assisting biocuration**. *Nucleic Acids Research*:1-5.
7. Nunes T, Campos D, Matos S, Oliveira JL (2013): **BeCAS: biomedical concept recognition services and visualization**. *Bioinformatics*, **29**(15):1915-1916.
8. Ravikumar, K.E., Li, D., Jonnalagadda, S., Wagholikar, K. B., Xia, N., & Liu, H. (2013, October). **An ensemble approach for chemical entity mention detection and indexing**. In *BioCreative Challenge Evaluation Workshop* (Vol. 2, p. 140).
9. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: **Gene Ontology: tool for the unification of biology**. *Nature genetics* 2000, **25**(1):25.