Generating Biological Expression Language Statements with Pipeline Approach and Different Parsers

Po-Ting Lai[1], Ming-Siang Huang[2], Wen-Lian Hsu[1], Richard Tzong-Han Tsai[3,*]
[1] Department of Computer Science, National Tsing-Hua University, Taiwan, R.O.C.
[2] Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.
[3] Department of Computer Science and Information Engineering, National Central University, Taiwan, R.O.C.

***Abstract*—In this paper, we describe our approach for the task 1 of the BioCreative VI Biological Expression Language (BEL) track. Our pipeline system is based on the BEL statement generation system, BelSmile. For the BioCreative VI BEL task, many components of BelSmile were updated. 1) We replaced the Conditional Random Fields (CRFs)-based approach with statistical principle-based approach (SPBA) for gene mention recognition. 2) A new verbal patterns were developed for function classification. 3) To improve SRL, we ensemble different semantic role labeling (SRL) parsers. 4) Our system is able to generate BEL statement even the relation is not described in a subject-verb-object (SVO) format. In the task 1, our best configurations achieved an F-score of 22.99% on the stage 1, and an F-score of 31.79% on the stage 2.**

Keywords—Biological Expression Language; Semantic Role Labeling; Named Entity Recognition; Relation Extraction

I. INTRODUCTION

The goal of the BioCreative VI Biological Expression Language (BEL) task 1 is that given a biological evidence sentence, the participants have to generate its corresponding BEL statement(s). The main challenges of this task are as following. First, the task contains many stages including named entity recognition (NER), named entity normalization (NEN), function classification, and relation classification. Therefore, developing a BEL statement generation system is more complicated than developing a single component. Second, the positions of named entity (NE), function and relation keyword are not provided in the training set. Therefore, the training set cannot be used to tune machine-learning models without appropriate preprocessing.

To tackle above challenges, we developed our system based on our previous pipeline system, BelSmile (1). Table I summarizes the resources and methods used for this task, and the differences with our previous system. The one with '*' symbol means that it was used for this task but not used in BelSmile.

BelSmile and our new system for BioCreative VI BEL task 1 differs in several ways. First, the CRFs-based gene mention recognition component was replaced by SPBA NER (2). Second, the verbal patterns were developed for function classification component. Third, for semantic role labeling, we ensemble our SRL parser, RCBiosmile (3), and a commonly-used parser, Enju (4). Lastly, our system could generate BEL statement even when the relation was presented in temporal and location statement.

This paper is arranged as follows. In Section II, we introduce our system for the BioCreative VI BEL track task 1. The configurations of each submission are described in Section III. In Section IV, we discuss the advantage of our approach. Section V concludes the paper and gives a future work.

TABLE I. THE RESOURCES AND METHOD USED FOR THIS TASK

Component	Method	Training set	Dictionary
Biological Process Recognition	Dictionary		BEL dictionary
Chemical Recognition	CRFs + dictionary	***BioCreative V CEMP**	ChEBI
Disease Recognition	Dictionary		BEL dictionary
Protein Recognition	***SPBA** + dictionary	JNLPBA + ***BioCreative VI CPRO**	Entrez
Function classification	Non-verbal pattern + ***verbal pattern**		
Semantic Role Labeling	RCBiosmile + ***Enju**	BioProp	
Relation Extraction	SRL + ***time/location rules**		

II. SYSTEM DESCRIPTION

In this section, we first introduce the named entity recognition components used in our system, and then introduce our normalization component. Third, we introduce our pattern-based approach for function classification. Then we introduce two different semantic role labeling components used for relation classification, and how we combine them. Lastly, we introduce how to integrate all components for BEL statement generation.

A. Named Entity Recognition

Here we briefly describe our named entity recognition components, which were used in our previous works (1, 2, 5, 6).

Chemical Recognition: We used NERChem (5) for chemical recognition. It was a CRF-based system which used the two-stage tokenization, consisting of GENIA tokenization and symbol tokenization. It used SOBIE (Singleton-, Outside-, Beginning-, Inner- and Ended-Named Entity) tag scheme and empirical feature set, including word, POS, affix, orthographical, word shape, syntax and NE features. We trained NERChem on the BioCreative IV chemical corpus to recognize chemicals.

Disease and Biological Process Recognition: For diseases and biological process recognition, the dictionaries provided by the BEL task were used to develop our dictionary-based recognition component. We used the longest matching algorithm to recognize both disease and biological process.

Gene Mention Recognition: We used SPBA (6) for gene mention recognition. It integrated the advantages of robustness of machine learning model and interpretability of pattern-based approach. SPBA was developed based on our revised version of JNLPBA corpus and BioCreative V.5 Gene and Protein Related Object Recognition (GPRO) corpus (2, 7).

B. Named Entity Normalization

Since named entity (NE) may not exactly match its corresponding dictionary names, the normalization process used heuristic rules to expand the query of a NE. The heuristic normalization rules, such as converting to lowercase and removing symbols and the suffix *'s'*, used in our previous works (1) are employed to expand both NE and dictionary names. Moreover, additional rules were developed to normalize proteins. For example, if a protein complex *"A/B complex"* is failed to be mapped into identifier, we will try to normalize it by separating it into two proteins. Table II shows the resources used in normalization.

TABLE II. THE RESOURCES USED FOR RECOGNITION AND NORMALIZATION

Type	Dictionary
Chemical	ChEBI
Gene mention	Entrez gene (human and mouse)
Biological Process	BEL dictionary
Disease	BEL task

C. Function Classification

Function classification component classifies the molecular activity of the NEs into transcription and phosphorylation activity etc. We used a pattern-based approach to classify the functions of the NEs. Our patterns are divided into two categories: non-verbal and verbal patterns.

In non-verbal pattern, a pattern consists of NE(s) and molecular activity keyword(s). Table III shows some examples of our non-verbal patterns, which were written by our domain experts.

In verbal pattern, each pattern consists of predicate and arguments. In the next section, we will introduced how to generate the predicate-argument-structure (PAS) of a given sentence. Table IV shows some examples of the verbal patterns, and how to transform the PAS into the BEL function statement.

TABLE III. EXAMPLES OF NON-VERBAL FUNCTION PATTERNS

Function	Example Pattern
molecularActivity (act)	<Protein/> activity
complexAboundance (complex)	<Protein/>/<Protein/> complex
degradation (deg)	<Protein/> degradation
proteinModification (pmod)	phosphorylation of <Protein/>
translocation (tloc)	translocation of <Protein/>

TABLE IV. EXAMPLES OF VERBAL FUNCTION PATTERNS

Function	Example Pattern	BEL statement
act	<AgentNE/> <Verb>activates</Verb> <PatientNE/>	<AgentNE/> increase act(<PatientNE/>)
complex	<ProteinA/> is complexed with <ProteinB/>	complex(<ProteinA/>,<ProteinB/>)
pmod	<AgentNE/> <Verb>phosphorylates</Verb> <PatientNE/>	<AgentNE/> increase pmod(<PatientNE/>, "P")

D. Semantic Role Labeling

Before we introduce our relation extraction component, we briefly introduce two SRL components, RCBiosmile and Enju, used in our relation classification component.

RCBiosmile: RCBiosmile (3) is a Markov-Logic-Network (MLN)-based SRL labeler that employs patterns to select candidate semantic roles for each argument and uses MLN (8) to learn and predict the semantic role of each argument. It was trained on BioProp (9).

Enju: Enju is a commonly-used semantic parser (4), which can extract agent and patient arguments. We selected it, since we found that it seems more accurate than other open-source non-biomedical domain SRL parsers in the BEL training set.

E. Combining Different SRL Systems

Here, we describe the procedure that used to combine SRL parsers as follows.

Step l: Both Enju and RCBiosmile were used to parse the predicate-argument-structure (PAS) of a given sentence.

Step 2: The agent argument, patient argument and predicate of the Enju PAS were collected into a list.

Step 3: All arguments and predicate of the RCBiosmile PAS were collected into a list.

Step 4: In the Enju, the arguments like negation, time and location were not labeled as negation, temporal and location arguments. Therefore, if Enju predicts the same predicate with

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

RCBiosmile. For these arguments, we will add these arguments generated by the RCBiosmile into the Enju's argument list.

Step 5: According to our observation on the BEL training set. We found that the RCBiosmile is more accurate than Enju while the predicted predicate is defined in the BioProp corpus. E.g. the predicate *"phosphorylate"*. However, these predicates often appear in the BEL training set. Therefore, we combined the results of two parsers by using predicate. For the predicates which were used in BioProp, we will use the arguments generated by RCBiosmile, and for the rest predicates we will use the arguments generated by Enju.

F. BEL Statement Generation

In this section, we describe how to extract the cause-theme-event relationship from PASs and transform them into BEL relations.

Given a sentence, combined SRL parser is used to parse it, and we will retrieve one or more PAS(s). Each PAS contains the arguments corresponding to the predicate. To extract cause-theme-event relationship, we map the predicate into the verb; map the abundances/processes which are inside the boundaries of the agent argument and the patient argument into the cause and the theme respectively.

However some cause-theme-event relationships are not presented in subject-verb-object format. For example, given a sentence,

"Furthermore, the expression of Bach 2, which can form a heterodimer with mafG protein, was found to be greatly reduced, while Notch 1 expression was increased in mafG-deficient mice." --- PMID:20813153

In SRL, the phrase *"in mafG-deficient mice"* is labeled as a location argument of the predicate *"increased"*;*"Notch 1 expression"* is labeled as the patient argument of the predicate *"increased"*. Although, *"mafG-deficient mice"* and *"Notch 1 expression"* are not an agent-patient pair, but the sentence means that *"mafG-deficient mice"* would increase *"Notch 1 expression"*. Therefore, the BEL statement

"p(EGID:4097) -| p(EGID:4851)" should be generated.

"p(EGID:4097)" refers to *"mafG"*; "p(EGID:4851)" refers to *"Notch 1"*.

To solve such problem, we map the predicate into the verb; map the abundances/processes which are inside the boundaries of the agent/patient argument into theme; map the abundances/processes which are inside the boundaries of the location or temporal arguments into cause.

The BEL relationship type is then determined by the regulation keywords collected from the BioNLP corpora (10) where "Regulation" and "Positive_regulation" types are mapped into the *increases*, and "Negative_regulation" is mapped into the *decreases*. If the surrounding context of NE abundances has certain keywords, like *"inhibition"* and *"knockout"*, then we will reverse the relation type.

III. EXPERIMENT RESULTS

We participated in both stage 1 and 2 of the BioCreative VI BEL track task 1, and three runs were submitted for each stage. The configurations of all runs are as follows:

Stage1:

- Run 1 (our best): it used the combined SRL parser.
- Run 2: it only used Enju for SRL.
- Run 3: it only used RCBiosmile for SRL.

Stage 2:

- Run 1: it used the combined SRL parser.
- Run 2: it used RCBiosmile parser as baseline, and added Run 1 statements. Furthermore, if some sentences only generated function statements, we will also output them.
- Run 3 (our best): it used the Run1 as baseline, and added RCBiosmile-based statement. Furthermore, if some sentences only generated function statements, we will also output them.

Table V and VI show the different level performances of our runs. In stage 1, the Run 1 used the combined SRL parser achieved our best performance on the statement evaluation metric. In stage 2, the Run 3 achieved our highest performance.

TABLE V. THE PERFORMANCES OF STAGE 1

Class		Recall	Precision	F-score
Term	run1	50.49	84.62	**63.24**
	run2	44.59	81.93	57.75
	run3	46.89	88.27	61.24
Function	run1	27.37	44.83	**33.99**
	run2	24.21	43.4	31.08
	run3	25.26	47.06	32.88
Relation	run1	31.58	55.38	**40.22**
	run2	28.07	53.33	36.78
	run3	28.07	56.14	37.43
Statement	run1	17.54	33.33	**22.99**
	run2	15.35	31.82	20.71
	run3	15.35	33.98	21.15

TABLE VI. THE PERFORMANCES OF STAGE 2

Class		Recall	Precision	F-score
Term	run1	72.79	99.11	83.93
	run2	76.07	99.15	**86.09**
	run3	75.08	99.13	85.45
Function	run1	29.47	47.46	36.36
	run2	33.68	50.79	**40.51**
	run3	32.63	49.21	39.24
Relation	run1	46.93	73.29	57.22
	run2	46.49	70.67	56.08
	run3	47.37	73.47	**57.6**
Statement	run1	23.68	46.15	31.3
	run2	23.68	44.63	30.95
	run3	24.12	46.61	**31.79**

IV. Discussion

In this section, we take a sentence as an example to illustrate the advantage of using combined SRL parser .

"Pulse-chase biosynthetic labeling studies showed that AtT-20 cells expressed much less RESP18 than the endogenous prohormone, POMC, but that glucocorticoid treatment lowered POMC and raised RESP18 biosynthetic rates so that they were nearly equimolar." --- PMID:7988462

Two gold BEL statements should be generated.

"a(CHEBI:glucocorticoid) -| p(EGID:5443);"

"a(CHEBI:glucocorticoid) -> p(EGID:389075)"

where *"POMC"* and "*RESP18*" are "p(EGID:5443)" and "p(EGID:389075)" respectively; *"lowered"* and *"raised"* are "-|" and "->" respectively. RCBiosmile failed to generate the BEL statements, since both *"lowered" and "raised"* were tagged as "JJ", and thus caused incorrect SRL results. However, in Enju, both *"lowered" and "raised"* can be tagged as "VBD", and then generated correct agent and patient arguments for both *"lowered" and "raised"*. In contrast, sometimes Enju might generate incorrect SRL results, but RCBiosmile didn't. Therefore, to solve this problem, combined SRL parser could reduce such errors coming from individual SRL components.

V. Conclusion and Future Work

In the future, we would like to apply the SPBA to tackle other NE types like chemical, disease and biological process. Using multiple components, the system performs better than using single component. Therefore, we would like to integrate different state-of-the-art systems in the future. The results of the BEL statement generation are depended on individual components. We also like to use deep learning-based approaches to enhance individual components like semantic role labeling.

References

1. Lai, P.-T., et al., BelSmile: a biomedical semantic role labeling approach for extracting biological expression language from text. *Database*, 2016. p. baw064.

2. Lai, P.-T., et al. Statistical Principle-based Approach for Gene and Protein Related Object Recognition. in *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop*. 2017.

3. Tsai, R.T.-H. and P.-T. Lai, A resource-saving collective approach to biomedical semantic role labeling. *BMC bioinformatics*, 2014. 15(1): p. 160.

4. Matsuzaki, T., Y. Miyao, and J.i. Tsujii, Efficient HPSG parsing with supertagging and CFG-filtering, in *Proceedings of the 20th international joint conference on Artifical intelligence*, 2007, Morgan Kaufmann Publishers Inc.: Hyderabad, India. p. 1671-1676.

5. Tsai, R.T.-H., Y.-C. Hsiao, and P.-T. Lai, NERChem: adapting NERBio to chemical patents via full-token features and named entity feature with chemical sub-class composition. *Database: The Journal of Biological Databases and Curation*, 2016 p. baw135.

6. Tsai, R.T.-H., et al., NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. BMC bioinformatics, 2006. 7(Suppl 5): p. S11-S11.

7. Pérez-Pérez, M., et al. Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: the CEMP and GPRO patents tracks. in *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop*, 2017.

8. Richardson, M. and P. Domingos, Markov logic networks. *Machine Learning*, 2006. 62(1): p. 107-136.

9. Tsai, R., et al., Semi-automatic conversion of BioProp semantic annotation to PASBio annotation. *BMC bioinformatics*, 2008. 9(Suppl 12): p. S18.

10. Kim, J.-D., et al., Overview of BioNLP Shared Task 2011, in *Proceedings of the BioNLP Shared Task 2011 Workshop*, 2011, Association for Computational Linguistics: Portland, Oregon. p. 1-6.