NCU-IISR System for BioCreative BEL Task 1

Po-Ting Lai[1], Yu-Yan Lo[2], Ming-Siang Huang[3], Yu-Cheng Hsiao[2], Richard Tzong-Han Tsai[2,*]

[1] Department of Computer Science, National Tsing-Hua University, Taiwan, R.O.C
[2] Department of Computer Science and Information Engineering, National Central University, Taiwan, R.O.C
[3] Department of Clinical Laboratory Sciences and Medical Biotechnology, College of Medicine, National Taiwan University, Taiwan, R.O.C
s102062802@m102.nthu.edu.tw
103522078@cc.ncu.edu.tw
r00424016@ntu.edu.tw
104522059@cc.ncu.edu.tw
thtsai@csie.ncu.edu.tw
*Corresponding Author

Abstract. Biological networks are important for biologists to represent and understand biological systems. These networks can be represented by languages such as BEL and SBML. Automatically extracting these descriptions and representing them in the biological system languages can improve the efficiency of constructing these networks. In this paper, we expand our previous Named Entity Recognition and Normalization systems for recognizing BEL abundances and processes. We use the Biomedical Semantic Role Labeling to parse the sentences into Predicate-Argument Structures (PASs), and transform these PASs into causal and correlative relationships. As for the BioCreative V BEL task 1, our proposed approach achieved an F-score of 19.66% on stage 1, and 33.08% on stage 2.

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

1 Introduction

The goal of the BioCreative V BEL subtask 1 is that when a biological evidence sentence is provided, a text mining system should extract and return its BEL statement. To complete this task 1, we used a pipeline approach which includes four stages: (1) abundance and process recognition; (2) abundance and process normalization; (3) function classification; (4) causal and correlative relationship classification.
For the first two stages, we expanded our previous gene name recognition, chemical name recognition and gene normalization systems [1-3] for the recognition. Subsequently, a keyword-based approach is used to classify their functions.
For causal and correlative relationship classification, the relationships can be represented in two major ways: the subject-verb-object (SVO) and non-SVO. An example of both representations is shown in Figure 1, in which the subject and object may be

adfa, p. 1, 2011.

the abundances, processes or activities. Our system focuses on extracting the biological events that are represented in the SVO relationship and ignore the non-SVO ones due to their complex nature.
In the SVO relationship, the relation type depends on both the verb and the context of the subject/object. For example, the expression "A up-regulates B" indicates that A *increases* B. However, "A knockout up-regulates B" implies that A *decreases* B. In our submitted results, we used the Biomedical Semantic Role Labeling parser (BioSRL) [4], to represent the sentence with Predicate-Argument Structures (PASs) [5]. In our system, we extract the SVO from the PASs of the sentence, and transform the SVO into the BEL causal and correlative relationship.

SVO Example
Text - "*insulin stimulated whereas dexamethasone inhibited 11beta-HSD1 activity and expression in a time- and concentration-dependent manner"* **Subject-Verb-Object -** *subject*: *"dexamethasone"*; *verb*: *"inhibited"*; *object*: *"11beta-HSD1"* **Predicate-Argument Structure -** *ARG0 (agent)*: *"dexamethasone"*; *predicate*: *"inhibited"*; *ARG1 (patient)*: *"11beta-HSD1 activity and expression"*; *ARGM-MNR (manner)*: *"in a time- and concentration-dependent manner"* **BEL Statement -** a(CHEBI:dexamethasone) *decreases* cat(p(MGI:Hsd11b1))
Non-SVO Example
Text - "*Rephosphorylation of PTP36 seemed to depend on actin polymerization since it was inhibited by cytochalasin D"* **BEL Statement -** a(CHEBI:"cytochalasin D") *increases* phos(p(MGI:Ptpn14))

Fig. 1. The SVO and non-SVO examples for the BEL task

2 Method

2.1 Abundance, Process and Function Identification

Abundance and Process Recognition: Both the Conditional Random Fields (CRFs)-based [1, 3] and dictionary-based Named Entity Recognition (NER) components were used for the recognition of the abundances and processes.
For recognizing the abundance p(), we used the NERBio [3] trained on the JNLPBA corpus [6]. The named entity types of the JNLPBA corpus contain DNA, RNA, Protein, Cell_Line and Cell_Type. We mapped the DNA, RNA and Protein type into p(). In addition, we used the chemical NER system [1] developed on the BioCreative IV chemical corpus to identify the chemical abundance a(). It applied the fine-grained tokenization and SOBIE tag scheme [1].
To determine the biological process of the GO terms bp() and the MeSH disease path()**,** we constructed dictionary-based NER systems that utilize the maximum matching algorithm. The same approach was also applied to distinguish both a() and p(). Table 1 summarizes the approaches and resources used in the recognition of different abundances and processes.
Abundance and Process Normalization: Named entities (NEs) identified as abundance or process must be normalized into their database identifiers. However, the text of

these NEs may be inconsistent with their corresponding names in the database. Therefore, we applied normalization rules such as converting alphabets to lower cases and removing symbols and the suffix "s" to expand both dictionary names and NEs.
If a p() exactly matches an identifier, it will be assigned to it. If two or more matching identifiers were found, we use the Entrez homolog dictionary to normalize homolog identifiers to human identifiers. Since other types of NEs did not contain as many ambiguous IDs as p(),therefore we did not apply the disambiguation process for these types.
Function Classification: The activity of NEs depends on its context. We manually collected both internal and surrounding activity keywords of NEs like "transcription" to classify their functions.

Table 1. The resources and models used for recognizing different abundances and processes.

Type	Model	Corpus	Dictionary	External Resource
a()	CRF+Dict	BCIV CHEMDNER	ChEBI	
p()	CRF+Dict	JNLPBA	Entrez gene	Entrez homolog, GO terms
bp()	Dict	-	BEL task	
path()	Dict	-	BEL task	

2.2 Causal Relationship Classification

When given a sentence, the BioSRL [4] will be used to parse it, and one or more PAS(s) will be retrieved. Each PAS contains a predicate and the arguments corresponding to the predicate. We extract SVO by mapping the predicate to the verb, and the abundances/processes within the boundaries of the agent and patient argument to the subject and the object, respectively.
The BEL relationship types are determined by the regulation keywords collected from the BioNLP corpora [7] and our manually collected keyword list. Both event types "Regulation" and "Positive_regulation" are mapped to the BEL relationship type *increases*, and the event type "Negative_regulation" is mapped to the BEL relationship type *decreases*.
In addition to keywords, relationship types are also determined by the surrounding words of the NEs. We manually constructed a keyword list consisting of words that may alter the relationship type, such as "inhibition", "mutant" and "inactivation". We employed a rule-based approach to adjust the relationship type accordingly. For instance, if the relationship type is *increases*, and the patient argument contains the keyword "inhibition" that is not inside the boundary of the tscript(p(MGI:Stat6)) . Then the p(MGI:Socs1) increases the inhibition of tscript(p(MGI:Stat6)), which implies that the p(MGI:Socs1) may decrease the tscript(p(MGI:Stat6)). Therefore, the relationship type is changed from *increases* to *decreases*.

3 Experiment Results

We participated in both stage 1 and 2 of the BioCreative V BEL task 1, and three runs were submitted for each stage. Table 2 displays the statement-level performances of

our runs for stage 1 and 2. Since our third run for stage 1 did not contain any relation, therefore its statement-level performance is absent in Table 2.

Stage 1 Results: Run 1 obtained a better result of the two, achieving recall/precision/f-score (RPF) of 14.36/31.18/19.66 on the BEL statement while employing the approach described in the Method section. There are two major reasons for the low overall performance. The first is that the false negative terms result in false negative statements. Our third run, which only contains the abundances and processes, achieved a lower RPF 61.0/64.21/62.56 on term-level evaluation. Therefore, the low recall of term-level recognition results in reduced overall performances. The second reason is the poor performance of function identification, which is due to the lack of keywords for activities. The recall of the second run is slightly higher than that of the first, because we returned the NE pairs with less than three NEs in a sentence.

Stage 2 Results: We retrieved the boundaries of abundances and processes by mapping the identifiers to the sentences with their synonyms in the database. If it cannot be mapped to the sentence, we will map it to the NE of stage 1, which has the smallest distance between two database identifiers. The first run with the GO terms biological process bp() improved the RPF from 19.8/68.97/30.77 to 21.78/68.75/33.08. We also submitted a third run, which outputs the pairs in which the total number of the abundances and activities is less than four in a sentence.

Table 2. The statement-level performances of our runs for stage 1 and 2.

Stage 1				Stage2			
Config.	**Recall**	**Precision**	**F-score**	**Config.**	**Recall**	**Precision**	**F-score**
run1	14.36%	31.18%	19.66%	run1	21.78%	68.75%	33.08%
run2	15.35%	26.72%	19.5%	run2	19.8%	68.97%	30.77%
				run3	23.76%	53.33%	32.88%

References

1. Dai, H.-J., Lai, P.-T., Chang, Y.-C., Tsai, R.: Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. Journal of Cheminformatics 7, S14 (2015)
2. Tsai, R.T.-H., Lai, P.-T.: Multistage gene normalization for full-text articles with context-based species filtering for dynamic dictionary entry selection. Food Chemistry 12 Suppl 8, (2011)
3. Tsai, R.T.-H., Sung, C.-L., Dai, H.-J., Hung, H.-C., Sung, T.-Y., Hsu, W.-L.: NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. BMC bioinformatics 7, S11-S11 (2006)
4. Tsai, R.T.-H., Lai, P.-T.: A resource-saving collective approach to biomedical semantic role labeling. BMC bioinformatics 15, 160 (2014)
5. Cohen, K.B., Hunter, L.: A critical review of PASBio's argument structures for biomedical verbs. BMC bioinformatics 7, (2006)
6. Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N.: Introduction to the bio-entity recognition task at JNLPBA. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pp. 70-75. Association for Computational Linguistics, Geneva, Switzerland (2004)
7. Ohta, T., Pyysalo, S., Rak, R., Rowley, A., Chun, H.-w., Jung, S.-j., Jeong, C.-h., Choi, S.-p., Ananiadou, S.: Overview of the Pathway Curation (PC) task of BioNLP Shared Task 2013. (2013)