

# Reducing Parsing Error of an Existing Parser for Extracting Biological Relation Events

Hyunchul Jang<sup>1</sup>, Jaesoo Lim<sup>1</sup>, Soo-Jun Park<sup>1</sup>, Seon-Hee Park<sup>1</sup>,  
Kyu-Chul Lee<sup>2</sup>

**Keywords:** Parsing, Biological Relation, Information Extraction

## 1 Introduction.

We are developing an information extraction system for biological literature. We are currently focusing on MEDLINE abstracts and trying to extract named entities and their relationships. We set a goal of finding methods including natural language processing, information extracting, and text processing but neither sentence tagging or parsing.

It's very difficult to develop a new tagger or a new parser specialized in biological document. If well-prepared corpora are supported, it may be easier by using machine learning method. In this paper, we explain a simple way to reduce parsing error of an existing parser.

## 2 Methods

We must define all possible template or rule if we are targeting on extracting more relation events than simple protein-protein interactions which have simple syntactic structures (i.e., Protein-A interacts with Protein-B.). Or we must create a corpus that has morphologically and syntactically tagged sentences to use a machine learning method.

Many sentences in biological literature are more complex syntactically than other documents. For this reason we need a well-trained parser in biological corpus. A tagger can be well trained since there is a splendid corpus [1] (i.e., the GENIA corpus [2,3]). But we are using an existing parser which is not trained in biological domain because we don't have any syntactically tagged corpus. Biological sentences are complex generally and many named entities in them consist of many words those have various morphological tags. When using an existing parser not trained by biological corpus, sometimes parsing is not possible because some sentence is too long or has words tagged not correctly. The following sentence is an example.

“Because activation of the p38 mitogen-activated protein kinase (p38MAPK)-mitogen-associated protein kinase-associated protein kinase (MAPKAPK)-2-heat-shock protein 27 signaling pathway mediates actin polymerization, we explored whether Abeta peptide activates p38MAPK and MAPKAPK-2.”

This sentence above has 52 tokens after tokenized. This sentence can't be processed by the parser we use without training.

---

<sup>1</sup> Bioinformatics Research Team, Electronics and Telecommunications Research Institute(ETRI), Gajeong-Dong, Yusong-Gu, Daejeon, Republic of Korea 305-350, E-mail: janghc, jslim, psj, shp {@etri.re.kr}

<sup>2</sup> Department of Computer Engineering, Chungnam National University, Gung-Dong, Yusong-Gu, Daejeon, Republic of Korea 305-764. E-mail: kcleee@cnu.ac.kr

We use our named entity recognizer to solve this problem. We substitute recognized named entities with one word like: “NEA”, “NEB”, “NEC”. Then we can have a more simplified sentence. As a result of simple measurement, our named entity recognizer identifies 80% of named entities in our corpus. The sentence above is changed like:

“Because activation of the NEA ( NEB ) NEC protein 27 signaling pathway mediates NED , we explored whether NEE activates NEF and NEG .”

This sentence now has 25 tokens. Then the parser can process this sentence.

### 3 Results.

We selected 100 abstracts randomly in our corpus and parsed their sentences. The number of sentences is 938. The parser which not trained by biological corpus can't process 74 sentences among them. But after substituting recognized named entities, it can process 57 sentences more.

Using Named Entity Substitution	# of abstracts	# of total sentences	Parsing	
			# of success	# of fail
Before	100	938	864(92%)	74
After			921(98%)	17

Table 1: Parsing results before and after using named entity substitution

17 sentences can't still be processed because they are still long and complex. The following sentence is an example.

“Recent therapeutic investigations of NEA ( NEB ) have been guided by two seemingly opposed hypotheses : the NEC cascade theory , which favors the NED as the cause of NEE ; and the cholinergic theory , which favors NEF as the cause .”

### References

[3] Kim, Jin-Dong, Tomoko Ohta, Yuka Tateisi and Jun'ichi Tsujii, “GENIA corpus – a semantically annotated corpus for bio-textmining,” *Bioinformatics*, pp i180-i182, Oxford University Press, 2003.

[1] Serguei Pakhomov, Anni Coden, Christopher Chute, “Creating a Test Corpus of Clinical Notes Manually Tagged for Part-of-Speech Information,” *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 62-65, 2004.

[2] Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, and Jun'ichi Tsujii, “The GENIA corpus: An annotated research abstract corpus in molecular biology domain,” *Proceedings of Human Language Technology Conference*, pp. 73-77, 2002.