

A Corpus Tagging Tool for Extraction of Biological Relation Events

Hyunchul Jang¹, Hyun-Sook Lee¹, Soo-Jun Park¹, Seon-Hee Park¹,
Kyu-Chul Lee²

Keywords: Tagged Corpus, Named Entity, Biological Relation, Information Extraction

1 Introduction.

We are developing a biological information extraction system. We are currently focusing on MEDLINE abstracts and trying to extract named entities and their relationships. Our system is based on machine learning. To use for training and testing our system, we are creating a tagged corpus from MEDLINE abstracts. We introduce our rules and tagging procedures for tagging biological named entities and their relation events.

2 Rules

Table 1 is tag format for named entity tagging. Identification numbers are omitted. If some words used in common, those are tagged with “<COMMON>”. Combination of other words tagged with “<FACTOR>” and common words are named entities.

Table 2 is tag format for relation event tagging. “%d” means an integer number given automatically, “n%d” for named entities, “a%d” for anaphoric entities and “r%d” for relation events. “AND/OR” tag has same number of identification number as number of “<FACTOR>” tag. Nested tags which have more named entities in themselves have only one identification number because inner named entities have not any role in relation event in general.

```
<NE category="category_string" definition="definition_string">Named Entity</NE>
<NE category="(AND|OR category_string|category_string)"><FACTOR>Factor String 1</FACTOR>
and <FACTOR>Factor String 2</FACTOR><COMMON>Common String</COMMON></NE>
<NE category ="category_string"><NE category ="category_string">Inner Name</NE> Outer
Name</NE>
```

Table 1: Tag format for named entity tagging

```
<NE id="n%d">Named Entity</NE>
<NE id="n%d,n%d" category="(AND category_string1|category_string2)"><COMMON>common
string</COMMON><FACTOR>factor1</FACTOR>,<FACTOR>factor2</FACTOR>
<NE id="a%d" precedence="n%d">Anaphora Entity</NE>
<RE id="r%d" type="type_string" polarity="-|+" negative="true|false" subject="{%n|a%d}"
object="{%n|a%d}" where="{%n|a%d}">Relation Action Word</RE>
```

Table 2: Tag format for relation event tagging

¹ Bioinformatics Research Team, Electronics and Telecommunications Research Institute(ETRI), Gajeong-Dong, Yusong-Gu, Daejeon, Republic of Korea 305-350, E-mail: janghc, lhs63473, psj, shp@etri.re.kr

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

3 Procedures and an Example.

Figure 1 shows the named entity tagging procedure. The relation event tagging procedure is similar in method and process

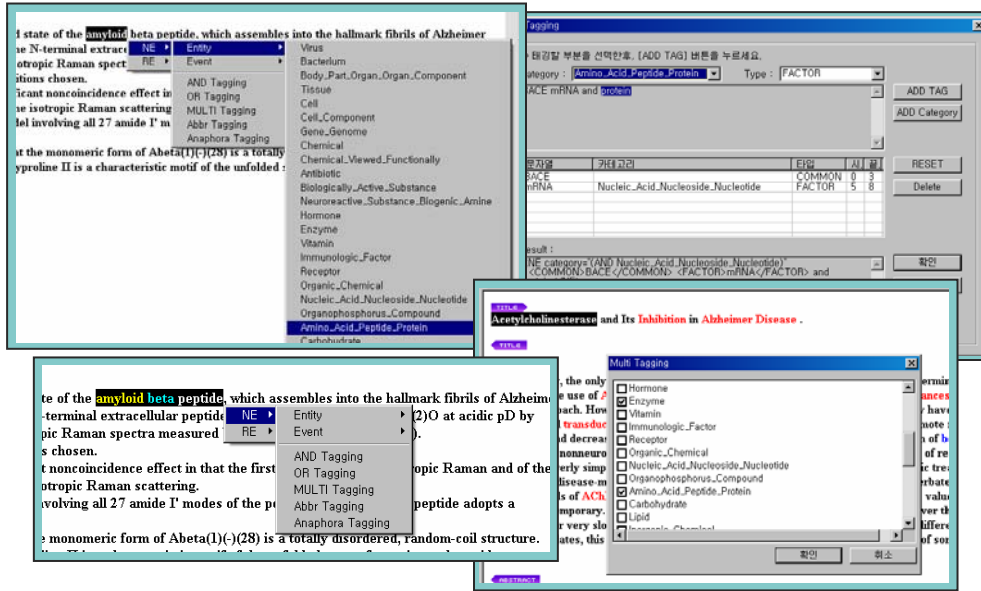


Figure 1: The named entity tagging procedure

```
<NE id="n1" category="Amino_Acid_Peptide_Protein">Abeta</NE> <RE id="r1"
type="interact_with" polarity="+" negativity="false" subject="{%n1}" object="{%n2}"
where="{%n3}">interacts</RE> with <NE id="n2" category="
Amino_Acid_Peptide_Protein">ABAD</NE> in the <NE id="n3"
category="Cell_Component">mitochondria</NE> of <NE id="n4"
category="Disease_Syndrome_Neoplastic_Process" definition="Alzheimer's Disease">AD</NE>
patients and transgenic mice.
```

Table 3: A tagging example

References

- [1] 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.
- [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.