A Semantically Annotated Corpus from MEDLINE Abstracts

Tomoko Ohta
okap@is.s.u-tokyo.ac.jp

Yuka Tateisi
yucca@is.s.u-tokyo.ac.jp

Nigel Collier
nigel@is.s.u-tokyo.ac.jp

Chikashi Nobata
nova@is.s.u-tokyo.ac.jp

Katsutoshi Ibushi
k-ibushi@is.s.u-tokyo.ac.jp

Jun'ichi Tsujii
tsujii@is.s.u-tokyo.ac.jp

Department of Information Science, Graduate School of Science, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

1 Introduction

Automatic information extraction is a key technology to help researchers access the information contained in research papers and to extend databases on substances and biological processes. We aim to build information extraction databases [2] from biochemical papers and their abstracts available from the MEDLINE [3] database. To objectively measure the performance of our systems, we built a corpus of expert-tagged abstracts from MEDLINE that can be used as the judgment set. This paper outlines the features of this corpus.

An annotated corpus has several other uses. For example, a large corpus can be a learning data set for statistical programs for information extraction (e.g., [1]). Also, computational linguists can explore semantic processing methods, or get insight into the linguistic patterns of how the relationship between the entities are represented in the texts, using the corpus explicitly marked with such entities.

2 The Annotation Scheme

We chose to mark up the names of *PROTEIN*s, *DNA*s, *RNA*s, and *SOURCE*s that appear in the abstracts in SGML [4] format. These names are considered to be relevant to the description of biological processes in genome domain, and recognition of such names is necessary for understanding higher level 'event' knowledge.

Names of proteins are marked up with `<PROTEIN>` tags. Names of DNAs are marked up with `<DNA>` tags. Names of RNAs are marked up with `<RNA>` tags. `<PROTEIN>`, `<DNA>`, and `<RNA>` tags have an attribute named *unit* which denotes whether the tagged object is a family or a group, a complex, a subunit of a complex, a molecule, a domain or a region, or a substructure. Names of sources are marked up with `<SOURCE>` tags, which have an attribute named *subtype* which denotes whether the source is a multi-cell organism, a mono-cell organism, a virus, a tissue, a cell, a cell-line, or a sub-location of a cell. Another attribute named *id* is assigned to all of the tags to represent synonymy: the same *id* value must be assigned to the names that refers to the same substance or source in a text.

More details of our annotation scheme can be found online [5]. An example of an annotated text is shown in Fig. 1.

3 Tagging Tool

Although a SGML-tagged text can be created by using text editors, semantically annotated corpora must be created by domain experts who are not always familiar with SGML tag scheme. Thus, an easy-to-use tagging tool to help annotators is indispensable for efficiency and accuracy. We have developed a GUI-based annotating tool on Microsoft Word (Fig. 2) and are making a new version in JAVA language.

```
UI - 92406925
TI - The <PROTEIN id=1 unit=fg unsure=ok>Jun family</PROTEIN> members, <PROTEIN id=2 unit=ml unsure=ok>c-Jun</PROTEIN>
and <PROTEIN id=3 unit=ml unsure=ok>JunD</PROTEIN>, transactivate the <DNA id=1 unit=dr unsure=ok>human c-myb
promoter</DNA> via an <DNA id=2 unit=dr unsure=ok>Ap1-like element</DNA>.
AB - The <DNA id=3 unit=ml unsure=ok>c-myb protooncogene</DNA>, which is preferentially expressed in <SOURCE id=1
subtype=ct unsure=ok>hematopoietic cells</SOURCE> at the G1/S boundary of the cell cycle, encodes a transcriptional
activator that functions via DNA binding.
```

Figure 1: Example of Annotated Text.

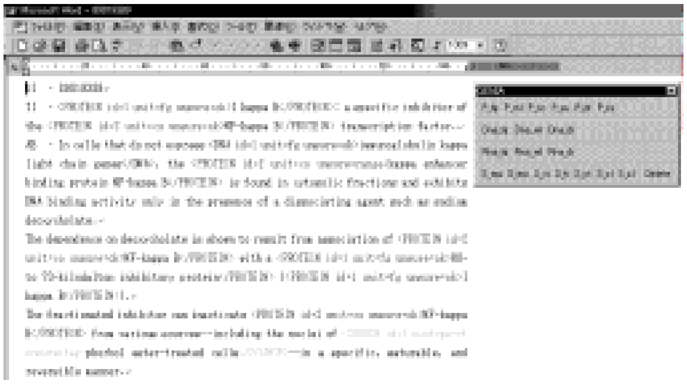

Figure 2: A Screenshot of the Tagging Tool on MS Word.

4 Conclusion

So far we have annotated 200 abstracts related to the transcription factors in human blood cells. We plan to add 600 more annotated abstracts by the end of this year. There are several directions in future work, including enhancement of the tag set by adding other tags, enrichment of tags by adding attributes, and expanding the functionality of the tagging tool.

Acknowledgments

This work is partially funded by Japan Society for Promotion of Science (JSPS-RFTF96P00502).

References

[1] Borthwick, A., Sterling, J., Agichtein, E., and Grishman, R., Exploiting diverse knowledge sources via maximum entropy in named entity recognition, *Proc. 6th Workshop of Very Large Corpora*, 152–160, 1998.

[2] Collier, N., Park, H.S., Ogata, N., Tateisi, Y., Nobata, C., Ohta, T., Sekimizu, T., Imai, H., Ibushi, K., and Tsujii, J., The GENIA Project: Corpus-based knowledge acquisition and information extraction from genome research papers, *Proc. EACL '99*, 271–272, 1999.

[3] National Library of Medicine, PubMed, http://www.ncbi.nlm.nih.gov/PubMed

[4] ISO/IEC 8879, Standard Generalized Markup Language, 1986.

[5] Tateisi, Y., Ohta, T., Collier, N., and Nobata, C., GENIA corpus annotation, http://www.is.s.u-tokyo.ac.jp/˜okap/annotate-bio-new.html
(English translation:http://www.is.s.u-tokyo.ac.jp/˜yucca/docs/genia/annotate-bio-new-e.html)