

Ontology Based Corpus Annotation and Tools

Tomoko Ohta¹

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

Yuka Tateisi²

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

Jin-Dong Kim²

jdkim@is.s.u-tokyo.ac.jp

Hideki Mima²

mima@is.s.u-tokyo.ac.jp

Jun'ichi Tsujii¹

tsujii@is.s.u-tokyo.ac.jp

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

² CREST, Japan Science and Technology Corporation, Honcho 4-1-8, Kawaguchi-shi,
Saitama 332-0012, Japan

Keywords: annotated corpus, ontology, XML, natural language processing

1 Introduction

With the explosion of results in molecular biology there is an increased need for IE to extract knowledge to support database building and to search intelligently for information in online journal collections. We aim to build information extraction systems from biology papers and their abstracts available from the MEDLINE database[1, 3]. As a part of a project on information extraction from the research papers in biology domain, we are creating an expert-tagged corpus of MEDLINE abstracts, which will be used for training and testing the information extraction systems. In this paper, we outline the features of this new corpus, its ontological basis, our annotation scheme, and statistics of its annotated objects. We also show the tagging and tag management tools.

2 Ontology-based Corpus Annotation

The task of annotation can be regarded as identifying and classifying the terms that appears in the texts according to a pre-defined classification. For this purpose, we first built a conceptual model (ontology) of substances and sources (substance location)[2, 4]. Based on this ontology, the names of PROTEINS, DNAs, RNAs, SOURCES, and OTHERS that appear in the abstracts are tagged. These names are considered to be relevant to the description of biological processes, and recognition of such names is necessary for understanding higher level ‘event’ knowledge.

The names in GENIA corpus are tagged using Genia Project Markup Language (GPML). GPML itself is in the form of XML document type definition (DTD) that consists of definitions of several elements. It provides the way of annotating corpora with linguistic information and document structure information. It also supports our strategy to build language resources simultaneously by providing a mechanism of extending existing ontology and lexicon while annotating corpora. The extended parts of ontology will be collected later for complementing the main ontology.

In designing GPML, simplicity and extensibility were regarded highly. For the simplicity, we didn't define such entities that are just useful for potential extension but not used for the current GENIA resources. For example, we didn't define structural entities for chapters or sections because the current GENIA corpus consists of MEDLINE abstracts and there are no such structures. For the extensibility, we tried to secure reliable guidelines instead of relying on imperfect predictions about potential extensions. We decided to maintain the structural part of GPML as compatible as possible with DocBook[5], a markup language for technical documentation maintained by OASIS. It is one

of the most popular XML applications supported by a large base of users and developers. By doing so as stated above, we hope for us, the maintainers of GPML not to lose our way while maintaining GPML, for the users with little experience in XML not to be overwhelmed by GPML itself and for the users with much experience in XML to get accustomed to GPML easily or even to predict the future revolution of GPML.

We have annotated 1,000 abstracts related to the transcription factors in human blood cells. We have marked up around 32,000 names with 36 different semantic classes. Around 9,500 proteins, 3,500 DNAs, 400 DNAs, 7,000 sources, 11,600 others are marked up. We are currently trying to increase the number of abstracts to 3,000. The current corpus is also used to gain the knowledge of how the tagged names are related to each other and other names, in order to enhance the ontology and to build the scheme for annotating more rich information such as biological events and roles.

3 Tools for Corpus Annotation

Although a XML-tagged text can be created by using text editors, semantically annotated corpora must be created by domain experts who are not always familiar with XML tag scheme. A tool for management of the tagged texts is also indispensable for controlling the quality of the corpus and taking the statistics of tag data.

To help annotators, we developed a GUI-based tag definition tool TagEdit and tagging tool JTag in JAVA language. In the tag definition tool TagEdit, definition of new tag-set, refinement of definition, enhancement of the tag-set by adding or removing tags, and enrichment of tags by adding or removing attributes are available. The tagging tool JTag has two frames: one is a tag selection frame, and the other is an annotation frame. In the tag selection frame, a tag-set based on ontology defined by using TagEdit is shown as a concept hierarchy. Tag data can be stored in two forms: An annotated text can be saved as tag-embedded XML document, or the data including the class of tag, the position of tag, and values of attributes can be saved separately from the text to be stored in external database. The latter form allows users to perform various operations on tags, e.g. to compare the annotation by different annotators and to take statistics.

We developed Tag Information Management System (TIMS) Workbench to perform/view tagging on a particular document. Since tag information is stored separately from original documents and managed using an external database software, various different types of tags for the same document can be added. The system also keeps track of the Audit Trail or History, i.e., the date and time, the user or system that performed the tagging etc. TIMS also has facility to search logically for tags, which we call Interval Operations. They are XML specific text/data mining operations which can be performed over TIMS document. The operations can be done over open documents or in batch mode, selecting the files from a list.

References

- [1] Collier N., Mima H., Lee S.Z., Ohta T., Tateishi Y., Yakushiji A., and Tsujii J. The GENIA Project: Knowledge Acquisition from Biology Texts, *Genome Informatics*, 11:448-449, 2000.
- [2] Tateishi, Y., Ohta, T., Collier, N., Nobata, C., and Tsujii, J., Building an Annotated Corpus in the Molecular-Biology Domain, Proc. COLING 2000 Workshop on Semantic Annotation and Intelligent Content, pp. 28-34, 2000.
- [3] <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>
- [4] <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/topics/Corpus/genia-ontology.html>
- [5] <http://www.oasis-open.org/docbook/>