Syntax annotation for the GENIA corpus

Yuka Tateisi[1] **Akane Yakushiji**[2] **Tomoko Ohta**[1] **Jun'ichi Tsujii**[2,3,1]

[1] CREST, Japan Science and Technology Agency
4-1-8, Honcho, Kawaguchi-shi, Saitama 332-0012 Japan
[2] Department of Computer Science, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[3] School of Informatics, University of Manchester
POBox 88, Sackville St, MANCHESTER M60 1QD, UK
{yucca,akane,okap,tsujii}@is.s.u-tokyo.ac.jp

Abstract

Linguistically annotated corpus based on texts in biomedical domain has been constructed to tune natural language processing (NLP) tools for bio-textmining. As the focus of information extraction is shifting from "nominal" information such as named entity to "verbal" information such as function and interaction of substances, application of parsers has become one of the key technologies and thus the corpus annotated for syntactic structure of sentences is in demand. A subset of the GENIA corpus consisting of 500 MEDLINE abstracts has been annotated for syntactic structure in an XML-based format based on Penn Treebank II (PTB) scheme. Inter-annotator agreement test indicated that the writing style rather than the contents of the research abstracts is the source of the difficulty in tree annotation, and that annotation can be stably done by linguists without much knowledge of biology with appropriate guidelines regarding to linguistic phenomena particular to scientific texts.

1 Introduction

Research and development for information extraction from biomedical literature (bio-textmining) has been rapidly advancing due to demands caused by information overload in the genome-related field. Natural language processing (NLP) techniques have been regarded as useful for this purpose. Now that focus of information extraction is shifting from extraction of "nominal" information such as named entity to "verbal" information such as relations of entities including events and functions, syntactic analysis is an important issue of NLP application in biomedical domain. In extraction of relation, the roles of entities participating in the relation must be identified along with the verb that represents the relation itself. In text analysis, this corresponds to identifying the subjects, objects, and other arguments of the verb.

Though rule-based relation information extraction systems using surface pattern matching and/or shallow parsing can achieve high-precision (e.g. Koike et al., 2004) in a particular target domain, they tend to suffer from low recall due to the wide variation of the surface expression that describe a relation between a verb and its arguments. In addition, the portability of such systems is low because the system has to be re-equipped with different set of rules when different kind of relation is to be extracted. One solution to this problem is using deep parsers which can abstract the syntactic variation of a relation between a verb and its arguments represented in the text, and constructing extraction rule on the abstract predicate-argument structure. To do so, wide-coverage and high-precision parsers are required.

While basic NLP techniques are relatively general and portable from domain to domain, customization and tuning are inevitable, especially in order to apply the techniques effectively to highly specialized literatures such as research papers and abstracts. As recent advances in NLP technology depend on machine-learning techniques, annotated corpora from which system can acquire rules (including grammar rules, lexicon, etc.) are indispensable

resources for customizing general-purpose NLP tools. In bio-textmining, for example, training on part-of-speech (POS)-annotated GENIA corpus was reported to improve the accuracy of JunK tagger (English POS tagger) (Kazama et al., 2001) from 83.5% to 98.1% on MEDLINE abstracts (Tateisi and Tsujii, 2004), and the FraMed corpus (Wermter and Hahn, 2004) was used to train TnT tagger on German (Brants, 2000) to improve its accuracy from 95.7% to 98% on clinical reports and other biomedical texts. Corpus annotated for syntactic structures is expected to play a similar role in tuning parsers to biomedical domain, i.e., similar improvement on the performance of parsers is expected by using domain-specific treebank as a resource for learning. For this purpose, we construct GENA Treebank (GTB), a treebank on research abstracts in biomedical domain.

2 Outline of the Corpus

The base text of GTB is that of the GENIA corpus constructed at University of Tokyo (Kim et al., 2003), which is a collection of research abstracts selected from the search results of MEDLINE database with keywords (MeSH terms) *human, blood cells* and *transcription factors*. In the GENIA corpus, the abstracts are encoded in an XML scheme where each abstract is numbered with MEDLINE UID and contains title and abstract. The text of title and abstract is segmented into sentences in which biological terms are annotated with their semantic classes. The GENIA corpus is also annotated for part-of-speech (POS) (Tateisi and Tsujii, 2004), and coreference is also annotated in a part of the GENIA corpus by MedCo project at Institute for Infocomm Research, Singapore (Yang et al, 2004).

GTB is the addition of syntactic information to the GENIA corpus. By annotating various linguistic information on a same set of text, the GENIA corpus will be a resource not only for individual purpose such as named entity extraction or training parsers but also for integrated systems such as information extraction using deep linguistic analysis. Similar attempt of constructing integrated corpora is being done in University of Pennsylvania, where a corpus of MEDLINE abstracts in CYP450 and oncology domains where annotated for named entities, POS, and tree structure of sentences (Kulick et al, 2004).

2.1 Annotation Scheme

The annotation scheme basically follows the Penn Treebank II (PTB) scheme (Beis et al, 1995), encoded in XML. A non-null constituent is marked as an element, with its syntactic category (which may be combined with its function tags indicating grammatical roles such as -SBJ, -PRD, and -ADV) used as tags. A null constituent is marked as a childless element whose tag corresponds to its categories. Other function tags are encoded as attributes. Figure 1 shows an example of annotated sentence in XML, and the corresponding PTB notation. The label “S” means “sentence”, “NP” noun phrase, “PP” prepositional phrase, and “VP” verb phrase. The label “NP-SBJ” means that the element is an NP that serves as the subject of the sentence. A null element, the trace of the object of “studied” moved by passivization, is denoted by “ <NP NULL="NONE" ref="i55"/>” in XML and “*-55” in PTB notation. The number “55” which refers to the identifier of the moved element, is denoted by “id” and “ref” attributes in XML, and is denoted as a part of a label in PTB.

In addition to changing the encoding, we made some modifications to the scheme. First, analysis within the noun phrase is simplified. Second, semantic division of adverbial phrases such as “–TMP” (time) and “–MNR” (manner) are not used: adverbial constituents other than “ADVP” (adverbial phrases) or “PP” used adverbially are marked with –ADV tags but not with semantic tags. Third, a coordination structure is explicitly marked with the attribute SYN=”COOD” whereas in the original PTB scheme it is not marked as such.

In our GTB scheme, “NX” (head of a complex noun phrase) and “NAC” (a certain kind of nominal modifier within a noun phrase) of the PTB scheme are not used. A noun phrase is generally left unstructured. This is mainly in order to simplify the process of annotation. In case of biomedical abstracts, long noun phrases often involve multi-word technical terms whose syntactic structure is difficult to determine without deep domain knowledge. However, the structure of noun phrases are usually independent of the structure outside the phrase, so that it would be

easier to analyze the phrases involving such terms independently (e.g. by biologists) and later merge the two analysis together. Thus we have decided that we leave noun phrases unstructured in GTB annotation unless their analysis is necessary for determining the structure outside the phrase. One of the exception is the cases that involves coordination where it is necessary to explicitly mark up the coordinated constituents.

In addition, we have added special attributes "TXTERR", "UNSURE", and "COMMENT" for later inspection. The "TXTERR" is used when the annotator suspects that there is a grammatical error in the original text; the "UNSURE" attribute is used when the annotator is not confident; and the "COMMENT" is used for free comments (e.g. reason of using "UNSURE") by the annotator.

<S><PP>In <NP>the present paper </NP></PP>, <NP-SBJ id="i55"><NP>the binding </NP><PP>of <NP>a [125I]-labeled aldosterone derivative </NP></PP><PP>to <NP><NP>plasma membrane rich fractions </NP><PP>of HML </PP></NP></PP></NP-SBJ><VP>was <VP>studied <NP NULL="NONE" ref="i55"/></VP> </VP>.</S>

(S (PP In (NP the present paper)), (NP-SBJ-55 (NP the binding) (PP of (NP a [125I]-labeled aldosterone derivative)) (PP to (NP (NP plasma membrane rich fractions) (PP of HML)))) (VP was (VP studied *-55)).)

Figure 1. The sentence "In the present paper, the binding of a [125I]-labeled aldosterone derivative to plasma membrane rich fractions of HML was studied" annotated in XML and PTB formats.

2.2 Annotation Process

The sentences in the titles and abstracts of the base text of GENIA corpus are annotated manually using an XML editor used for the Global Document Annotation project (Hasida 2000). Although the sentence boundaries were adopted from the corpus, the tree structure annotation was done independently of POS- and term- annotation already done on the GENIA corpus. The annotator was a Japanese non-biologist who has previously involved in the POS annotation of the GENIA corpus and accustomed to the style of research abstracts in English. Manually annotated abstracts are automatically converted to the PTB format, merged with the POS annotation of the GENIA corpus (version 3.02).

3 Annotation Results

So far, 500 abstracts are annotated and converted to the merged PTB format. In the merging process, we found several annotation errors. The 500 abstracts with correction of these errors are made publicly available as "The GENIA Treebank Beta Version" (GTB-beta).

For further clean-up, we also tried to parse the corpus by the Enju parser (Miyao and Tsujii 2004), and identify the error of the corpus by investigating into the parse errors. Enju is an HPSG parser that can be trained with PTB-type corpora which is reported to have 87% accuracy on Wall Street Journal portion of Penn Treebank corpus. Currently the accuracy of the parser drops down to 82% on GTB-beta, and although proper quantitative analysis is yet to be done, it was found that the mismatches between labels of the treebank and the GENIA POS corpus (e.g. an –ing form labeled as noun in the POS corpus and as the head of a verb phrase in the tree corpus) are a major source of parse error. The correction is complicated because several errors in the GENIA POS corpus were found in this cleaning-up process. When the cleaning-up process is done, we will make the corpus publicly available as the proper release.

4 Inter-Annotator Agreement

We have also checked inter-annotator agreement. Although the PTB scheme is popular among natural language processing society, applicability of the scheme to highly specialized text such as research abstract is yet to be discussed. Especially, when the annotation is done by linguists, lack of domain knowledge might decrease the stability and accuracy of annotation.

A small part of the base text set (10 abstracts) was annotated by another annotator. The 10 abstracts were chosen randomly, had 6 to 17 sentences per abstract (total 108 sentences). The new annotator had a similar background as the first annotator that she is a Japanese non-biologist who has experiences in translation of

technical documents in English and in corpus annotation of English texts.

The two results were examined manually, and there were 131 disagreements. Almost every sentence had at least one disagreement. We have made the 'gold standard' from the two sets of abstracts by resolving the disagreements, and the accuracy of the annotators against this gold standard were 96.7% for the first annotator and 97.4% for the second annotator.

Of the disagreement, the most prominent were the cases involving coordination, especially the ones with ellipsis. For example, one annotator annotated the phrase 'IL-1- and IL-18-mediated function' as in Figure 2a, the other annotated as Figure 2b.

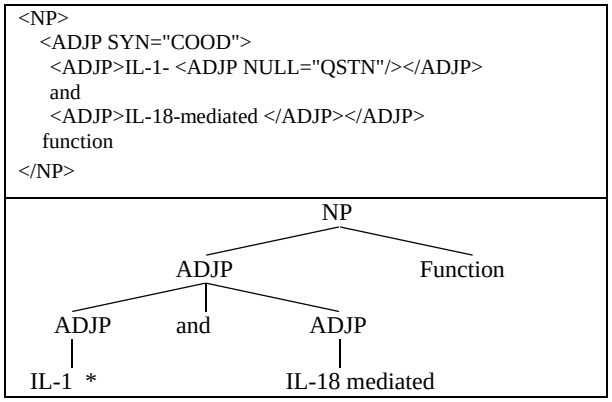

Figure 2a. Annotation of a coordinated phrase by the first annotator. A* denotes a null constituent.

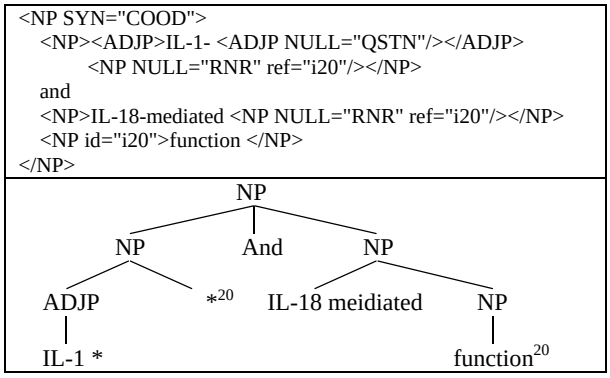

Figure 2b. Annotation of the same phrase as in Figure 2a by the second annotator. A * denotes a null constituent and '20' denotes coindexing.

Such problem is addressed in the PTB guideline and both formats are allowed as alternatives. As coordination with ellipsis occurs rather frequently in research abstracts, this kind of phenomena has higher effect on decrease of the agreement rate than in Penn Treebank. Of the 131 disagreements, 25 were on this type of coordination.

Another source of disagreement is the attachment of modifiers such as prepositional phrases and pronominal adjectives. However, most are 'benign ambiguity' where the difference of the structure does not affect on interpretation, such as 'high expression of STAT in monocytes' where the prepositional phrase 'in monocytes' can attach to 'expression' or 'STAT' without much difference in meaning, and 'is augmented when the sensitizing tumor is a genetically modified variant' where the *wh*-clause can attach to 'is augmented' or 'augmented' without changing the meaning. The PTB guideline states that the modifier should be attached at the higher level in the former case and at the lower case in the latter. In the annotation results, one annotator consistently attached the modifiers in both cases at the higher level, and the other consistently at the lower level, indicating that the problem is in understanding the scheme rather than understanding the sentence. Only 15 cases were true ambiguities that needed knowledge of biology to solve, in which 5 involved coordination (e.g., the scope of 'various' in 'various T cell lines and peripheral blood cells') .

Although the number was small, there were disagreements on how to annotate a mathematical formula such as 'n=2' embedded in the sentence, since mathematical formulae were outside the scope of the original PTB scheme. One annotator annotated this kind of phrase consistently as a phrase with '=' as an adjective, the other annotated as phrase with '=' as a verb. There were 6 such cases. Another disagreement particular to abstracts is a treatment of labeled sentences. There were 8 sentences in two abstracts where there is a label like 'Background:'. One annotator included the colon (':') in the label, while the other did not. Yet another is that one regarded the phrase 'Author et al' as coordination, and the other regarded 'et al' as a modifier.

Other disagreements are more general type such as regarding '-ed' form of a verb as an adjective or a participle, miscellaneous errors such as omission of a subtype of label (such as '-PRD' or '-SBJ) or the position of <PRN> tags

with regards to ‘,’ for the inserted phrase, or the errors which look like just ‘careless’. Such disagreements and mistakes are at least partially eliminated when reliable taggers and parsers are available for preprocessing

5 Discussion

The result of the inter-annotator agreement test indicates that the writing style rather than the contents of the research abstracts is the source of the difficulty in tree annotation. Contrary to the expectation that the lack of domain knowledge causes a problem in annotation on attachments of modifiers, the number of cases where annotation of modifier attachment needs domain knowledge is small. This indicates that linguists can annotate most of syntactic structure without an expert level of domain knowledge.

A major source of difficulty is coordination, especially the ones involving ellipsis. Coordination is reported to be difficult phenomena in annotation of different levels in the GENIA corpus (Tateisi and Tsujii, 2004), (Kim et al., 2003). In addition to the fact that this is the major source of inter-annotator agreement, the annotator often commented the coordinated structure as ‘unsure’. The problem of coordination can be divided into two with different nature: one is that the annotation policy is still not well-established for the coordination involving ellipsis, and the other is an ambiguity when the coordinated phrase has modifiers.

Syntax annotation of coordination with ellipsis is difficult in general but the more so in annotation of abstracts than in the case of general texts, because in abstracts authors tend to pack information in limited number of words. The PTB guideline dedicates a long section for this phenomena and allows alternatives in annotation, but there are still cases which are not well-covered by the scheme. For example, in addition to the disagreement, the phrase illustrated in Figure 2a and Figure 2b shows another problem of the annotation scheme. Both annotators fail to indicate that it is ‘mediated’ that was to be after ‘IL-1’ because there is no mechanism of coindexing a null element with a part of a token.

This problem of ellipsis can frequently occur in research abstracts, and it can be argued that the tokenization criteria must be changed for texts in biomedical domain (Yamamoto and Satou, 2004) so that such fragment as ‘IL-18’ and ‘mediated’ in ‘IL-18-ediated’ should be regarede as separate tokens. The Pennsylvania biology corpus (Kulick et al., 2004) partially solves this problem by separating a token where two or more subtokens are connected with hyphens, but in the cases where a shared part of the word is not separated by a hyphen (e.g. ‘metric’ of ‘stereo- and isometric alleles’) the word including the part is left uncut. The current GTB follows the GENIA corpus that it retains the tokenization criteria of the original Penn Treebank, but this must be reconsidered in future.

For analysis of coordination with ellipsis, if the information on full forms is available, one strategy would be to leave the inside structure of coordination unannotated in the treebank corpus (and in the phase of text analysis the structure is not established in the phase of parsing but with a different mechanism) and later merge it with the coordination structure annotation. The GENIA term corpus annotates the full form of a technical term whose part is omitted in the surface as an attribute of the ‘<cons>’ element indicating a technical term (Kim et al., 2003). In the above-mentioned Pennsylvania corpus, a similar mechanism (‘chaining’) is used for recovering the full form of named entities. However, in both corpora, no such information is available outside the terms/entities.

The cases where scope of modification in coordinated phrases is problematic are few but they are more difficult in abstracts than in general texts because the resolution of ambiguity needs domain knowledge. If term/entity annotation is already done, that information can help resolve this type of ambiguity, but again the problem is that outside the terms/entities such information is not available. It would be practical to have the structure flat but specially marked when the tree annotators are unsure and have a domain expert resolve the ambiguity, as the sentences that needs such intervention seems few. Some cases of ambiguity in modifier attachment (which do not involve coordination) can be solved with similar process.

We believe that other type of disagreements can be solved with supplementing criteria for linguistic phenomena not well-covered by the scheme, and annotator training. Automatic preprocessing by POS taggers and parsers can also help increase the consistent annotation.

6 Conclusion

A subset of the GENIA corpus is annotated for syntactic (tree) structure. Inter-annotator agreement test indicated that the annotation can be done stably by linguists without much knowledge in biology, provided that proper guideline is established for linguistic phenomena particular to scientific research abstracts. We have made the 500-abstract corpus in both XML and PTB formats and made it publicly available as "the GENIA Treebank beta version" (GTB-beta). We are in further cleaning up process of the 500-abstract set, and at the same time, initial annotation of the remaining abstracts is being done, so that the full GENIA set of 2000 abstracts will be annotated with tree structure.

For parsers to be useful for information extraction, they have to establish a map between syntactic structure and more semantic predicate-argument structure, and between the linguistic predicate-argument structures to the factual relation to be extracted. Annotation of various information on a same set of text can help establish these maps. For the factual relations, we are annotating relations between proteins and genes in cooperation with a group of biologists. For predicate-argument annotation, we are investigating the use of the parse results of the Enju parser.

Acknowledgments

The authors are grateful to annotators and colleagues that helped the construction of the corpus. This work is partially supported by Grant-in-Aid for Scientific Research on Priority Area C "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Brants,T.(2000). TnT: a statistical part-of-speech tagger, *Proceedings of the sixth conference on Applied natural language processing*, pp.224-231, Morgan Kaufmann Publishers Inc.

Beis.A., Ferguson,M., Katz,K., and MacIntire,R.(1995). Bracketing Guidelines for Treebank II Style: Penn Treebank Project, University of Pennsylvania

Hasida, K. (2000). GDA: Annotated Document as Intelligent Content. *Proceedings of COLING'2000 Workshop on Semantic Annotation and Intelligent Content.*

Kazama,J., Miyao,Y., and Tsujii,J.(2001) A Maximum Entropy Tagger with Unsupervised Hidden Markov Models, *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, pp. 333-340.

Kim,J-D, Ohta,T., Tateisi,Y. and Tsujii,J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics. 19(suppl. 1).* pp. i180-i182. Oxford University Press.

Koike,A., Niwa,Y., and Takagi,T. (2004) Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics, Advanced Access published on October 27, 2004;* doi:10.1093/bioinformatics/bti084.Oxford University Press.

Kulick,S., Bies,A., Liberman,M., Mandel,M., McDonald,R., Palmer,M., Schein,A., Ungar,L., Winters,S. and White,P. (2004) Integrated Annotation for Biomedical Information Extraction. *BioLINK 2004: Linking Biological Literature, Ontologies, and Databases,* pp. 61-68.Association for Computational Linguistics.

Miyao,Y. and Tsujii,J. (2004a). Deep Linguistic Analysis for the Accurate Identification of Predicate-Argument Relations**.** *Proceedings of COLING 2004.* pp. 1392-1397.

Tateisi,Y. and Tsujii,J. (2004). Part-of-Speech Annotation of Biology Research Abstracts. *Proceedings of the 4th International Conference on Language Resource and Evaluation (LREC2004).* IV. pp. 1267-1270, European Language Resources Association.

Wermter, J. and Hahn, U. (2004). An annotated German-language medical text corpus. *GMDS 2004 meeting,* http://www.egms.de/en/meetings/gmds2004/04gmds168.shtml.

Yamamoto,K., and Satou,K (2004). Low-level Text Processing for Life Science, *Proceedings of the SIG meeting on Natural Language Processing, Information Processing Society of Japan, IPSJ-SIGNL-159* (In Japanese).

Yang,XF., Zhou,GD., Su,J., and Tan.,CL (2004). Improving Noun Phrase Coreference Resolution by Matching Strings. *Proceedings of 1st International Joint Conference on Natural Language Processing (IJCNLP'2004),* pp226-233.