Part-of-Speech Annotation of Biology Research Abstracts

Yuka Tateisi[*,†], **Jun-ichi Tsujii**[†,*]

[*]CREST, Japan Science and Technology Agency
[†]University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033 Japan
{yucca,tsujii}@is.s.u-tokyo.ac.jp

Abstract

A part-of-speech (POS) tagged corpus was built on research abstracts in biomedical domain with the Penn Treebank scheme. As consistent annotation was difficult without domain-specific knowledge we made use of the existing term annotation of the GENIA corpus. A list of frequent terms annotated in the GENIA corpus was compiled and the POS of each constituent of those terms were determined with assistance from domain specialists. The POS of the terms in the list are pre-assigned, then a tagger assigns POS to remaining words preserving the pre-assigned POS, whose results are corrected by human annotators. We also modified the PTB scheme slightly. An inter-annotator agreement tested on new 50 abstracts was 98.5%. A POS tagger trained with the annotated abstracts was tested against a gold-standard set made from the interannotator agreement. The untrained tagger had the accuracy of 83.0%. Trained with 2000 annotated abstracts the accuracy rose to 98.2%. The 2000 annotated abstracts are publicly available.

1. Introduction

Research on automatic information extraction from literature in the biomedical domain using natural language processing is rapidly growing. Annotated corpora are crucial resources for the research and there have been several corpora available where protein names e.g., (Olsson et al., 2002), substance names and other technical terms e.g., (Kim et al., 2003) or alias and coreference resolution e.g., (Medstract Project, 2002) are annotated. An ongoing work of integrated annotation is described in (Kulick et al., 2004).

Here we describe a part-of-speech (POS) tagged corpus on the MEDLINE abstracts. Parts of speech are the most basic, but yet useful, information for text processing. We annotate the POS information to the raw texts of the GENIA corpus (Kim et al., 2003). By annotating POS to the same text set in which technical terms are annotated, we expect the corpus to be a useful resource for building term recognition systems, and also adapting existing POS taggers and other applications to biomedical domain. In this paper, we describe the corpus, the problems encountered during the annotation process, and experimental results on inter-annotator agreement.

2. Outline of the corpus

The GENIA corpus is an annotated corpus which contains 2000 MEDLINE abstracts that were collected using the search terms *human*, *transcription factors*, and *blood cells*. Technical term information, i.e., the names of substances, sources (biological locations where the substances are found), and other technical terms relevant to the descriptions of biological events, are marked up with their semantic class in XML language. A part of the corpus is shown in Figure 1. This corpus is called 'the term corpus' hereafter to distinguish from its base text on which the POS information is annotated, which will be called 'the raw corpus'.

We assign POS to each word in the text according its syntactic role. This principle is applied even to the words that are part of multi-word to terms. That is, each component of a multi-word term is assigned a POS according to the syntactic role of the word, not according to the role of the term as in e.g., CLAWS (Wynne, 1996) scheme. The decision is because a corpus annotated with both term information and the POS of every word can be useful for training and evaluating the term extraction systems that uses the POS and other syntactic feature of words.

Our annotation scheme for POS corpus is based on that of Penn Treebank corpus (Santorini, 1990) widely used in constructing general-purpose statistics-based NLP-systems. Using the same scheme used in such systems would enable us to use those systems to help construct the corpus on one hand. On the other hand, the corpus can be used for evaluation of applicability of such systems to texts in biomedical domain.

3. Nature of biomedical abstracts

A preliminary experiment showed several problems to annotation originating form the nature of biomedical research abstracts. An inter-annotator agreement rate on 50 abstracts taken from the GENIA raw corpus and the PTB scheme, between two master-course students in linguistics, was 86.7% in kappa-score (Carletta, 1996). The problems mainly result from the characteristics of the base text. Unlike everyday English text, the research abstracts in molecular biology domain include 1) (non-proper) names and abbreviations that begin with capital letters, 2) chemical and numeric expressions that includes non-alphanumeric characters such as commas, parentheses and hyphens, 3) participles of unfamiliar verbs that describe domain-specific events, and 4) fragments of words.

Especially, names and abbreviations that begin with capital letters (e.g., *NFAT*, *CD4*, *RelB*) make the distinction between proper and common nouns problematic. The

This work was partially supported by Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

<abstract>
<sentence><cons lex="IL-2-mediated_T_cell_proliferation" sem="G#other_name"><cons lex="IL-2" sem="G#protein_molecule">IL-2</cons>-mediated <cons lex="T_cell" sem="G#cell_type">T cell</cons> proliferation</cons> is a critical early event ...</sentence>
...

Figure 1: The GENIA term corpus

agreement rate rose to 96.4 % when we ignored the distinction between NNP and NN tags, and NNPS and NNS tags. As many technical terms, including names of substances, and abbreviations that begin with capital letters, annotators cannot rely on initial capital letters to decide whether an unfamiliar word is a proper noun or not without domain knowledge. To make the matter more complex, abbreviation of adjectival expression is often found. For example, *CD4-* is an abbreviation of *CD4-negative* and appears exclusively in prenominal positions to describe a state of cells. This means that we cannot simply regard abbreviation as nouns.

It often happens that non-alphanumeric characters appear in what appears to be a token. Chemical and numeric expressions may include commas and parentheses (e.g., *1,25(OH)2D3*, *beta-(1,3)-glucan*, *t(3;3)(q21;q26)*) that usual tokenizers separate from the rest, while an abbreviation of a term in parentheses often follows the full term. Thus, annotators often cannot decide whether to correct the tokenization around parentheses and other punctuations. Another major problem occurs when names with hyphens (e.g., *c-Rel*, *beta-globin*) are in prenominal positions because the PTB guideline says that "hyphenated modifiers should be tagged as adjectives (JJ)" and annotators must know whether a particular expression is a name that should be tagged as NN or two words connected with hyphens that should be tagged as JJ.

The distinction between adjectives and (present or past) participles is also difficult because it involves context and semantic judgment. Thehe same type of inconsistency is found even in the PTB corpus, but it is more problematic in biomedical texts because of unfamiliar verbs.

In abstracts, sometimes only a part of words are written in place of a full word (e.g. *up- and downregulate* instead of *upregulate and downregulate*, *transcription factor(s)* instead of *transcription factor or transcription factors*). We could not find this type of expression in PTB corpus and the tags for fragments of words like prefixes and suffixes were out of the scope of the PTB tag set.

These problems indicate that even POS tagging that is regarded as the most basic task of syntactic processing requires the knowledge of domain specialists.

4. Annotation process

From the preliminary results, we decided to modify the scheme and make use of the term corpus as a resource in order to retain consistency[1].

The use of NNP and NNPS tags is limited so that only the names of the months, the names of authors of the papers, journals, research institutes, and initials of patients and other people who contributed to experiments described in the paper, as they are clear from the context. All other nouns are tagged as common nouns, even when a person's name appears as a part of other names (e.g., *Cushing's syndrome*, *Southern blotting*). The decision is because while the distinction is costly for consistency, from the viewpoint of syntactic processing such as parsers the need for the distinction is rather small. Tokenization is changed so that the parentheses and other punctuations in a name are not separated. We still retain the rule that two or more words that are connected with hyphens (e.g. *CD4-negative*) as one token. Prefixes and postfixes are tagged based on their syntactic role. For example, the token *up-* in *up- and downregulation* is assigned RP tag because it originate from a particle (*regulate up*), and NNS tag is assigned to the token *s* in *factor(s)*.

To make the annotation task easier for non-biologists, we made use of the existing annotation of the GENIA corpus. Most of the POS of the components of technical terms and other domain-specific expressions can be determined independently from the context. Thus if the POS of these terms are correctly pre-assigned, and then a POS tagger can determines the POS of the remaining words respecting the pre-assigned POS, the errors caused by those terms can be reduced.

We compiled a list of frequent terms annotated in the GENIA corpus and assigned a POS to each word of each term (e.g. NF/NN kappa/NN B/NN) with assistance from researchers in biochemistry and immunology. Size of the list was rather small, about 600 entries. Some common patterns in chemical expressions and abbreviations are compiled into regular expressions (e.g. IL-[0-9]/NN, CD[0-9]*-/JJ) to annotate the common pattern. Common expressions involving -ing and -ed forms of verbs (e.g. signaling/NN pathway/NN), and the suffixes that determine the class of names and therefore the POS (e.g. -toxin/NN) are also added.

In the actual annotation process, the text was first tokenized using Penn tokenizer. Then a perl script (post-tokenizer) is run on the result of the tokenizer to correct tokenization errors mostly around chemical expressions like *2,3,7,8-tetrachrolodibenzodioxin*, and then another script (pre-tagger) assigns POS to the components of technical terms. A modified version of JunK POS tagger (Kazama et al., 2001), which is reported to have 96.84% accuracy on the PTB Wall Street Journal corpus, is used to determine the

[1]As far as we knew at the start of the project there were no specialist dictionaries in which all the POS of the constituents of technical terms were provided. For example, the specialist lexicon in UMLS (National Library of Medicine, 2003) has the POS of terms as a whole but not those of individual constituents

```
<abstract>
<sentence><cons lex="IL-2-mediated_T_cell_proliferation" sem="G#other_name"><cons lex="IL-2"
sem="G#protein_molecule"><w c="*">IL-2</w></cons><w c="JJ">-mediated</w> <cons lex="T_cell"
sem="G#cell_type"><w c="NN">T</w> <w c="NN">cell</w></cons> <w c="NN">proliferation</w></cons>
<w c="VBZ">is</w> <w c="DT">a</w> <w c="JJ">critical</w> <w c="JJ">early</w> <w c="NN">event</w>
...</sentence>
...
```

Figure 2: POS information merged with the GENIA term corpus

POS of remaining words. The tagger uses the pre-annotated POS as constraints to that of remaining tokens, i.e., it preserves any preexisting assignment and assigns only the POS consistent one to other tokens. The pre-assigned POS are specially marked, so that in the human-correction phase, annotator can see which words are assigned POS based on the term list. Finally errors in tokenization and POS in the output of the tagger are corrected by a human annotator and another annotator checks the result. A guideline of annotation supplementary to the PTB manual was compiled with examples of problematic cases dependent on context.

Three master-course students in linguistics participated as annotators. They have periodically met to discuss the problems and settle the disagreement, and the tagging guideline and the term lists are enriched with the problematic examples encountered. So far, we have annotated 2000 abstracts made publicly available.

5. Experimental results

Inter-annotator agreement was tested between an annotator that has participated in creating the 2000-abstract corpus and another annotator who has not participated in the project so far. A new set of 50 abstracts taken from MEDLINE with same search terms as that of the GENIA term corpus(*human*, *transcription factors*, *blood cells*) were annotated. The annotators were given the pre-tagged texts (the output of the tagger with special marks on pre-assigned POS) and corrected them independently of each other. Unlike the actual annotation process the results of human annotators were not re-checked. The two results are aligned and null tokens are inserted to the place where tokenization disagreement occurred. That is, if the string 'abc' us tagged as 'a/A b/B c/C' by one annotator and 'abc/D' by the other, the latter result is adjusted to 'abc/D / / / ' by inserting null tokens with null POS tags.

In the actual result, there were two disagreements in tokenization where three tokens are inserted in total. No common characteristics were found in the two disagreements. The adjusted number of tokens was 11179, of which 11025 were agreed. The simple agreement rate was 98.6% and the kappa-score was 98.5%.

Overall it can be concluded that with our current process POS tagging can be done consistently without much domain-specific knowledge. However, the technical terms are still the largest problem, causing 46 disagreements. The most frequent disagreements (19) were on the abbreviation of plural expressions where inconsistency between NN and NNS occurred. The NN-JJ disagreement between slash- or hyphen-bound names in prenominal position is the next frequent (13). Another frequent disagreement was between NN and FW (10). Two was NN-CD disagreement on the positions on a gene (*14q13.2* and *11q13*). The other was NN-JJ disagreement on the word *paracrine*.

Another frequent type of disagreement (26) was between CC and DT on *both* in *both ... and* construction (24) and *either* in *either ... or* construction (2). It was found out that this came from the misunderstanding of the guideline by the new annotator and this is expected to be removed easily. However, annotators may make mistakes of this kind because the coordinated phrases, hence the distances between the coordinators, tend to be long in research abstracts. Indeed, there were other disagreements that make us suspect that the style of research abstracts, where sentences tend to be long and complex, is another source of difficulty in POS annotation. There were 14 disagreements that involved the tensed forms of verbs (9 VBD-VBN, 2 VBP-VB, 1 VBP-JJ, 1 VBP-NN, and 1 VBZ-NN). These cases require full understanding of syntactic structure of the sentence and are more difficult than others where POS can be determined by local contexts.

There were 15 NN-VBG disagreements of which 12 were on the word *binding*. There were 6 JJ-VBG disagreements and 4 JJ-VBN disagreements where no particular words are frequently disagreed. The remaining disagreements have no common characteristics.

In another experiment, three versions of the JunK tagger with and without the preprocessor (six in total) were tested. The first one was the original version of the JunK tagger. The second was trained on 670-abstract subset of the POS-annotated corpus corresponding to the GENIA corpus Ver. 1.1. The third was trained on full 2000-abstract set of the corpus. The same 50 abstracts used for the inter-annotator agreement experiment was used in the experiment. The abstracts were tokenized by the Penn-tokenizer and the post-tokenizer was run on the output to fix the errors that can be corrected mechanically (unpretagged set). The pre-tagger assigned the POS to technical terms (pretagged set). The three versions of the JunK tagger were run on both unpretagged and pretagged set. The six results were compared against a gold-standard set made by re-checking the results of human annotators from the previous experiment.

The results are shown in Table 1 while the human annotators' accuracies against the gold standard were 99.6 and 98.7%. This result shows that the corpus has enough quality to be used for training POS taggers for adaptation to subdomain. The pre-tagger is effective when the training size is small but the advantage of pre-tagging is reduced with respect to the accuracy when the training set is large enough.

Size	Unpretagged	Pretagged
0	83.0	93.2
670	96.3	98.0
2000	98.2	98.2

Table 1: Accuracy of the JunK tagger trained with various sizes of training sets

However, according to annotators pre-tagging was helpful because the special marking of pre-tagged POS reduces the mental burden of annotators.

6. Merging with term corpus

The POS corpus is provided in three formats. One is a "PTB-like" format where there are one TOKEN/POS pair par line. Another is an XML format where tokens are represented in *w* elements and the POS is represented as the *c* attribute. Yet another is a "merged" format where the POS annotation is merged into the term corpus (Figure 2). In this version the DTD assumed that the *w* elements are inside the *cons* elements. However, sometimes a token was split by the <*cons*> tags, i.e., a technical term represented by a *cons* element is inside a token represented by a *w* element. For example, in Figure 2 the token *IL-2-mediated* because of <*cons*> tags around *IL-2*. In such cases, we made each fragment one *w* element. The last fragment of the split token is assigned the original POS assigned to the whole token and all others are assigned * as the value of the *c* attribute.

In the 2000-abstract corpus, 7652 such token were found. Out of these tokens, 6690 include the - symbol(s) (either as a hyphen or a minus sign), 1084 included slashes, and 544 include the both. Many of the tokens (about 4600) had construction like 'name-adjectival' such as *lipopolysaccharide-induced* and *AP-1-dependent*. Most of rest of tokens including the - symbol(s) and the tokens including slash(es) were two or more names connected with hyphens or slashes.

On the other hand, of tokens that includes neither slash nor the - symbol, the most frequent (86) were the tokens like *<cons><w c="*">CD4</w></cons><w c="JJ">+</w>* which is an abbreviation of *CD4-positive*. There are 51 cases where a part of a name is recognized as a separate term, e.g. *<cons><w c="*">Stat</w></cons><w c="NN">5a</w></cons>*. A few cases (24) included no non-alphabetic characters as in *<cons><cons><w c="JJ">homo-</w></cons> <w c="CC">and</w> <cons><w c="*">hetero</w></cons><cons"><w c="NNS">dimers</w></cons></cons>* that involve a coordination and ellipsis. Although it is natural to argue that the 'name-adjectival' cases and the names bound by hyphens and slashes should be regarded as multiple tokens, it is not clear whether tokens should be separated in cases like the latter three.

7. Conclusions

We have made a part-of-speech (POS) tagged corpus built on MEDLINE abstracts used as the base of the GENIA corpus. We made use of the existing term annotation of the GENIA corpus to annotate the POS to the constituents of technical terms. An inter-annotator agreement test showed that with the process the corpus is consistently annotated by non-biologists. However, the investigation of disagreement indicated that technical terms are still problematic and more intelligent preprocessor using exhaustive dictionary might be necessary.

We also tested a POS tagger trained with the annotated corpus. The results showed that when trained on 2000 abstracts the accuracy was almost a human level, thus showing that the corpus is useful for training a POS tagger to adapt to the subdomain.

The 2000 annotated abstracts are publicly available from our website (`http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA`). Another 1500 abstracts are being annotated and will also be publicly available.

As the training experiment is indicating that we have made a large enough quantity, after finishing the current abstracts we plan to improve the quality rather than further increasing the volume. Especially, the tokenization of slash- and hyphen-bound expressions should be reinvestigated because it is a source of disagreement and merging with the term corpus indicate that biologists regard a part of these expressions as one unit.

8. References

Carletta, J. C., 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Kazama, J., Y. Miyao, and J. Tsujii, 2001. A maximum entropy tagger with unsupervised hidden markov models. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*.

Kim, J-D., T. Ohta, Y. Tateisi, and J. Tsujii, 2003. Genia corpus - a semantically annotated corpus for biotextmining. *Bioinformatics*, 19:180i–182i.

Kulick, S., A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer, E. Pancoast, A. Schein, L. Ungar, P. White, and S. Winters, 2004. Integrated annotation for biomedical information extraction. To be presented in HLT/NAACL2004 (BioLink2004).

Medstract Project, 2002. Initial annotated corpora. http://medstract.org/gold-standards.html.

National Library of Medicine, 2003. UMLS knowledge resources documentation. http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML.

Olsson, F., G. Eriksson, K. Franzen, L. Asker, and P Liden, 2002. Notions of correctness when evaluating protein name taggers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*.

Santorini, B., 1990. Part-of-speech tagging guidelines for the Penn Treebank project. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Wynne, M., 1996. A post-editor's guide to CLAWS7 tagging. http://www.hcu.ox.ac.uk/BNC/what/claws7.html.