GENIA-GR: a Grammatical Relation Corpus for Parser Evaluation in the Biomedical Domain

Yuka Tateisi[1], Yusuke Miyao[2], Kenji Sagae[2], Jun'ichi Tsujii[2,3]

[1]Department of Informatics, Kogakuin University
1-24-2 Nishi-shinjuku, Shinjuku-ku, Tokyo, 163-8677, Japan
[2]Graduate School of Information Science and Technology, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
[3]School of Computer Science, University of Manchester/National Centre for Text Mining
131 Princess Street, Manchester, M1 7DN, UK
E-mail: yucca@cc.kogakuin.ac.jp, yusuke@is.s.u-tokyo.ac.jp, sagae@is.s.u-tokyo.ac.jp, tsujii@is.s.u-tokyo.ac.jp

Abstract

We report the construction of a corpus for parser evaluation in the biomedical domain. A 50-abstract subset (492 sentences) of the GENIA corpus (Kim et al., 2003) is annotated with labeled head-dependent relations using the grammatical relations (GR) evaluation scheme (Carroll et al., 1998) ,which has been used for parser evaluation in the newswire domain.

1. Introduction

With the sharp increase in the amount of published information available in biology and related fields, biomedical text mining has become an important research tool. Although shallow text processing techniques based on pattern matching have been employed with some success to biomedical text, there is now a growing interest in applying more advanced natural language processing techniques, such as full syntactic parsing, to identify more complex relationships between concepts in the biomedical literature. Some of these relationships are hierarchical in nature, and some involve phrases that are widely separated in the text, making them difficult to recognize using only surface patterns.

Among the general full-sentence syntactic parsing approaches that have been applied to biomedical text recently, many can be broadly categorized as either constituent parsers that identify the phrase structure of input sentences, or dependency parsers that identify direct relationships between words in input sentences. Because different parsers use different styles of syntactic representation, it is difficult to compare the accuracy of different syntactic analysis approaches. In the parsing community, the use of a fixed set of dependency-based grammatical relations (Carroll et al., 1998) has been proposed as a way of comparing parsers that use different formalisms. While this evaluation scheme is not as widely used as PARSEVAL, it has recently gained some traction as a more framework-independent alternative, and has been used in the evaluation of parsers that work with various types of syntactic representations, from the one in the widely used Penn Treebank, to others based on lexicalized grammar formalisms, such as CCG and HPSG (Preiss, 2003; Kaplan et al 2004; Clark and Curran, 2007; Miyao et al., 2007).

One obstacle in applying the same type of formalism-independent dependency-based evaluation scheme to parsers for biomedical text is the lack of a manually annotated corpus to serve as a gold-standard. Previous work on evaluation of parsers in the biomedical domain (Pyysalo et al., 2007; Clegg and Shepherd, 2007) used corpora automatically converted to the Stanford dependency scheme (de Marneffe et al., 2006) from gold-standard phrase structure trees in the Penn Treebank (PTB) (Marcus et al., 1993) format. Although the Stanford dependency (SD) format was originally intended to be used by NLP applications, and not in parser evaluation, a program that converts structures in PTB format into SD structures is freely available, and is distributed with the Stanford parser [1], making SD convenient for comparison among PTB parsers. However, this automatic conversion produces a significant number of errors, and the resulting SD data cannot be considered a gold-standard corpus. Because these errors are undocumented, the suitability of the resulting corpus for parser evaluation is questionable (although the conversion may still be well suited for use in applications). Furthermore, the lack of a complete detailed description of the different dependency types in the SD scheme makes it unlikely that accurate and reliable conversion from other representation formats (such as CCG- or HPSG-based formats) can be achieved.

To address these concerns, we propose the use of Carroll et al.'s GR evaluation scheme for parsers in the biomedical domain. The GR scheme was designed specifically for parser evaluation, and it has been used for that purpose by different NLP research groups. In addition, manually annotated gold-standard corpora exist in other domains[2] (Briscoe and Carroll, 2006). We have started the creation of a gold-standard corpus for parser evaluation in the biomedical domain, by annotating sentences from the GENIA corpus (Kim et al., 2003) using Carroll et al.'s GR scheme. In this paper, we describe the main issues in the creation of the evaluation corpus, including our application of the GR scheme to a domain that is significantly different than other domains

[1] http://nlp.stanford.edu/software/lex-parser.shtml

[2] http://www.informatics.sussex.ac.uk/research/groups/nlp/carroll/greval.html

that this scheme has been applied to previously.

2. The Grammatical Relations scheme

The Grammatical Relation scheme (GR) was proposed aiming at a framework-independent metric for parsing accuracy (Carroll et al., 1998). A set of 700 sentences extracted from Section 23 of the Penn Treebank, the same set as the PARC 700 Dependency Bank (King et al.,2004), was manually annotated and made publicly available in addition to an older set of 500 sentences from the SUSANNE corpus. Carroll and Briscoe (2006) used 140 of the 700 sentences as the development set, and the remaining 560 were used as the gold standard data for evaluation. The same 560-sentence set is also used as the gold standard by Clark and Curran (2007) and Miyao et al. (2007). Hereafter, we refer to the gold standard set as PTB-GR.

```
They market cable-TV on the very grazing opportunities
CNN seeks to discourage.
(ncsubj market They _)
(iobj market on)
(dobj market cable-TV)
(dobj on opportunities)
(det opportunities the)
(ncmod _ opportunities grazing)
(cmod _ opportunities seeks)
(ncsubj seeks CNN _)
(ncsubj discourage CNN _)
(dobj discourage opportunities)
(xcomp to seeks discourage)
(ncmod _ opportunities very)
```

Figure 1: GR example in the PARC 700 set

Figure 1 shows an example of GR annotation. GR represents labeled syntactic dependencies between words. For example, `ncsubj` means a non-clausal subject, `dobj`, and `ncmod` expresses a non-clausal modifier. Most relations are binary, while a few relation types have additional slots that represent subtypes of the relations. For example, `(xcomp to seeks discourage)` means that *discourage* is a to-infinitival complement of *seeks*. Refer to (Briscoe, 2006) for the definition of these relation types.

GR annotations are almost purely syntactic, and therefore lack the means to evaluate the true potential of deep linguistic parsers that compute relationships based on semantics. However, it should be noted that GR represents non-local dependencies such as control/raising and movement. In this example, `(ncsubj discourage CNN _)` indicates a control relation, `(dobj discourage opportunities)` expresses a moved object of *discourage*, and `(cmod _ opportunities seeks)` indicates a relation between a relative clause and its antecedent. Since these relations are not explicitly represented by PTB parsers, this scheme may serve as a starting point in the identification of the added benefits of deep parsing and the discussion of problems in framework-independent evaluation. On the other hand, identifying most of the relationships in the GR scheme in the output of shallow phrase structure parsers requires matching of tree patterns, which makes it challenging to evaluate those parsers.

3. The GENIA-GR corpus

3.1 Our annotation scheme

In the PTB-GR corpus, a sentence form is accompanied by `id` (sentence identifier), `date` (date of last modification), and `validators` (names of annotators), `structure` (phrase structure), and `rasp` (the dependency structure) elements (in XML terms). We have adopted the GR scheme with the change that our scheme also includes a `named_entities` element that lists the named entities in a sentence where internal structure is not annotated. On the other hand, our corpus does not include the phrase structure annotation, for which we plan to integrate the annotation of the GENIA treebank (Ohta et al. 2006) in the future.

The corpus is encoded in XML. Elements regarding document structure are inherited from the GENIA corpus. The sentence element has three attributes `id` (sentence identifier), `date` (date of last modification), and `validators` (names of annotators), and includes the elements `named_entities`, `sentence_form`, and `rasp`. The `named_entities` element consists of one or more term elements, which is the entity annotated in the GENIA term corpus with `id` (term identifier inherited from the term corpus), `sem` (semantic class inherited from the term corpus), and `span` (beginning and ending position of the term). Only multi-word terms, not all of the terms annotated in the term corpus, are listed in the `named_entities` element. The `sentence_form` consists of the tokenized (as in the GENIA treebank) form of the sentence. The *rasp* element consists of the list of elements in the form of `<XXX>`*`head dependent`*`</XXX>`, where *XXX* is one of the types in the GR annotations, with an optional attribute that shows the subtype of the GR annotations. For example, `(ncsubj market They _)` is encoded into `<ncsubj initial="_"> market They </ncsubj>` and `(xcomp to seeks discourage)` is encoded into `<xcomp subtype="to">seeks discourage</xcomp>`. For multi-word terms declared in the `named_entities` element, only their `id`s are referred to in the dependency structure described in the `rasp` element.

3.2 The base corpus

From the GENIA corpus, we chose 50 abstracts (492 sentences) that satisfy the following conditions: 1) the abstract is indexed with MeSH[3] term *NF-kappa B*, 2) the full text is available freely from PubMed Central[4], and 3)

[3] Medical Subject Headings (MeSH) is the U.S. National Library of Medicine's controlled vocabulary thesaurus, used for indexing PubMed database.

[4] PubMed Central (PMC) is the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences

the abstract is outside the set used for training parsers by Hara et al (2007). The abstracts satisfying the condition 1) are annotated with event structure (Kim et al., 2008) and used for the construction of NFκB pathway by Oda et al. (2007). This means that the target information to be extracted with syntactic parsing is available for these abstracts. The condition 2) is for enhancing the corpus to full-text set. The condition 3) is set because the parser described in (Hara et al., 2007) is one of the parsers to be evaluated with this corpus.

The corpus is not large, but it is comparable in size to PTB-GR (560 sentences). We believe it is sufficient for evaluation purposes.

3.3 The method of annotation

The abstracts are tokenized in the same ways as the corresponding sentences in the GENIA Treebank. The 492 sentences were tokenized into 13943 tokens.

The annotation was accomplished by manually cleaning up the output of the RASP system (Briscoe et al., 2006) by the first author, who has a natural language processing background and has been involved in the GENIA corpus especially in linguistic annotations (Part-of-speech and Treebank). The PTB-GR corpus and the GENIA term corpus, but not the GENIA treebank, were used for reference. The annotator inserts comments to questionable structures in the form of free comments in `note` elements.

```
sentence( id(99384376.6)
named_entities(
   term(id(T23) span(12:13) (NF-kappaB element))
   term(id(T25) span(17:19) (long terminal repeat))
)
sentence_form(Our transient expression data revealed
   that dopamine stimulated transcription through the
   NF-kappaB element present in the long terminal
   repeat .)
rasp(
   (ccomp that:6 revealed:5 stimulated:8)
   (ncsubj revealed:5 data:4 _)
   (det data:4 Our:1)
   (ncmod _ data:4 transient:2)
   (ncmod _ data:4 expression:3)
   (ncsubj stimulated:8 dopamine:7 _)
   (dobj stimulated:8 transcription:9)
   (ncmod _ stimulated:8 through:10)
   (cmod _ *T23* present:14)
   (dobj through:10 *T23*)
   (det *T23* the:11)
   (iobj present:14 in:15)
   (dobj in:15 *T25*)
   (det *T25* the:16)
))
```

Figure 2: Example of annotation.

4. Annotation Results

4.1 The annotated corpus

An example of the annotation is shown in Figure 2. For the sake of readability, the annotation is presented in the form similar to the PTB-GR corpus, and some of the attributes are omitted. *T23* and *T25* respectively denote the reference to the term *NF-kappaB element* and *long term repeat*, declared as terms in the `named-entities` element. The numbers after the colons following the tokens in the relations denote the position of the token in the sentence.

In the 492 sentences, 10029 relations[5] are annotated in total. The distribution of relation types are shown in Table 1. The `arg` type, underspecified type encoding the relation between a verb and its argument (subject or complement), is used for embedded mathematical expression in GENIA-GR (See section 4.2).

Relation Type	GENIA-GR		PTB-GR		
	N_{GENIA}	F_{GENIA}	N_{PTB}	F_{PTB}	F_{GENIA}/F_{PTB}
arg	24	0.05			
aux	293	0.60	401	0.72	0.83
ccomp	176	0.36	290	0.52	0.69
cmod	125	0.25	165	0.29	0.86
conj	975	1.98	591	1.06	1.88
csubj	6	0.01	3	0.01	2.28
det	990	2.01	1115	1.99	1.01
dobj	2181	4.43	1762	3.15	1.41
iobj	962	1.96	545	0.97	2.01
ncmod	2163	4.40	3548	6.34	0.69
ncsubj	946	1.92	1351	2.41	0.80
obj2	1	0.00	20	0.04	0.06
passive	330	0.67	228	0.41	1.65
pcomp	5	0.01	23	0.04	0.25
pmod	6	0.01	13	0.02	0.53
ta	254	0.52	286	0.51	1.01
xcomp	294	0.60	380	0.68	0.88
xmod	294	0.60	178	0.32	1.88
xsubj	4	0.01	7	0.01	0.65
Total	10029		10906		

Table 1: Distribution of relation types. N_{GENIA} is the number of relation types in total 492 sentences in GENIA-GR, and N_{PTB} is the number of relation types in the total 560 sentences in PTB-GR. F_{GENIA} and F_{PTB} are the frequency of the type per sentence in GENIA-GR and PTB-GR, respectively.

The relative frequencies of relation types show the characteristics of MEDLINE abstracts compared to newswire text. For example, a higher frequency of `conj` (conjunction) and `xmod` (non-saturated clause modifying a head) suggests the more complex structure of the sentences in the abstracts, where more coordination and phrasal modifiers are found.

Higher frequency of `iobj` (prepositional complements of verbs, adjectives, and verbal nouns) shows that there are

journal literature from over 200 journals.

[5] As the corpus is in the clean-up process, the number may be slightly different in the final version.

more prepositional phrases in the abstracts, together with the fact that `dobj` (nominal direct object of a verb or a preposition) is relatively less frequent compared to `iobj`, which indicates the lower frequency of direct objects of verbs (otherwise the frequency of `dobj` would be higher than that of `iobj`). This suggests that more verbs are nominalized in the abstracts, with their arguments taking the form of prepositional phrases. However, it was not easy to distinguish between the prepositional arguments and modifiers, and thus the distinction between `iobj` (PP arguments) and `ncmod` (non-phrasal modifiers including PPs).

The following section discusses the problems we encountered in the annotation process and their (partial) solutions.

4.2 Problematic structures in biomedical text

We have not added any new relation types to the original GR scheme, but we have applied our own interpretation to the original guidelines in some cases specific to biomedical text and for those cases that have not been covered in the GR documentation.

The cases include the following:

4.2.1 Appositive without comma

Subtype `ta` (text adjunct) of `ncmod` relation, undocumented in (Briscoe 2006), is used for apposition without commas like *nuclear factor NF kappa B*, where *nuclear factor* and *NF kappa B* are appositives, as illustrated in Figure 3, where *NF kappa B* denotes the reference to the phrase *NF kappa B* declared as a term.

```
(ncmod factor nuclear _)
(ncmod factor *NF kappa B* ta)
```

Figure 3: The annotation of *nuclear factor NF kappa B.*

4.2.2 Gene sequence

A gene sequence may be denoted by the numbers denoting the starting and ending positions, connected by *to* (*e.g. 296 to 302*). We could not find the similar constructions in the PTB-GR, so that we decided to annotate the beginning position (*296*)as the head and *to* plus the ending position (*to 302*) is dependent on the head as a prepositional modifier, as illustrated in Figure 4.

```
(ncmod ta residues 296)
(ncmod _ 296 to)
(dobj to 302)
```

Figure 4: The annotation of *residues 296 to 302*

4.2.3 X Y-ing (Y-ed) Z

A frequent construction is a triplet *X Y-ing (Y-ed) Z* (NP-participle-NP, e.g., *NF kappa B binding domain*), where X is an argument of Y and together they modify Z. In PTB-GR, a similar construction is not frequent, mainly because in such cases X and Y are hyphenated and make one token like X-Ying Z. One example in PTB-GR is *New York-based pharmaceutical industry research firm* where *-based* is dependent as a non-clause modifier (`ncmod`) on *New York* and *New York* is dependent as `ncmod` on *firm*. We follow this interpretation as shown in Figure 5 for compatibility for the time being (see also section 5).

```
NF kappa B binding domain
 (ncmod _ *NF kappa B* binding)
(ncmod _ domain *NF kappa B*)
```

Figure 5. Annotation of *NF kappa B binding domain* in the same way as the annotation of *New York-based pharmaceutical industry research firm* in PTB-GR

4.2.4 Mathematical formula embedded in text:

There were two formulae embedded in the text in the 50 abstracts. In an embedded formula, we decided to annotate equality operators and the like (=, >, etc.) as verbs, and their arguments as dependent on them as `arg`. In complex cases such as*2 x NFKappaB &gt; or = SlVmac239 approximately deltaNFkappaB approximately deltaSpl234 approximately substNFkappaB approximately substSpl2 approximately substSp23* (where *&gt;* is the replacement of the sign > and *approximately* is the replacement of the sign ≈) all signs except the first are treated as if modifying the argument right in front of it , as shown in Figure 6.

```
(arg or *2 x NFkappaB*)
(conj or &gt;)
(conj or =)
(arg or SlVmac239)
(xmod SlVmac239 approximately _)
(arg approximately deltaNFkappaB:26 _)
```

Figure 6: The annotation of *2 x NFKappaB &gt; or = SlVmac239 approximately deltaNFkappaB*

4.2.5 References in the text:

```
(ta comma J.Virol.72 and)
(ta colon J.Virol.72 5852-5861)
(ta comma J.Virol.72 1998)
(conj and M.Rothe)
(conj and L.Chene)
(conj and M.Nugeyre:42
(conj and F.Barre-Sinoussi)
(conj and N.Israel:47)
```

Figure 7: The annotation of *M.Rothe , L.Chene , M.Nugeyre , F.Barre-Sinoussi , and N.Israel , J.Virol.72 : 5852-5861 , 1998*

References to other literatures are characteristic elements of scientific text. The style varies, but some journals require a rich information like in an example *We have previously demonstrated that ...(M.Rothe , L.Chene , M.Nugeyre , F.Barre-Sinoussi , and N.Israel , J.Virol.72 : 5852-5861 , 1998*). Two abstracts included such a style of references. In such cases, the list of authors can be considered as an appositive to *We,* or the entire bibliographic information enclosed in the parentheses can be considered as an appositive to the whole sentence (or, when the reference appears in the middle of the sentence, a phrase preceding the reference. The tentative solution is to annotate them in a way that, in the parentheses, the head is the journal name, and the authors, volumes, pages, and year is treated as adjuncts. The whole structure is treated as dependent on the main verb of the sentence, if it appears at the end, or the head of the preceding phrase, if it appears in the middle of the sentence, as illustrated in Figure 7.

4.3 Inconsistencies with the existing GENIA annotations

In the process of annotation, inconsistencies in the existing GENIA annotation, especially concerning the boundaries, were found.

Sometimes the unit for the head-dependent structure violates the token boundary assigned by the POS annotation. This typically occurs around a hyphen-bound word, treated as one token in the GENIA POS corpus following the Penn Treebank POS annotation scheme. For example, take *renal cell carcinoma-derived gangliosides.* In the phrase, *derived,* a part of a token in the GENIA POS corpus, takes *renal cell carcinoma* as an argument and modifies *gangliosides.* The Gold560 corpus has entries in a similar form (eg. *New York-based pharmaceutical industry research firm*), where tokenization is changed from the original PennTreebank corpus.

Fortunately, most of the cases occurred inside the terms, thus left unannotated. Still, there are cases like *N- and C-terminal* where a token had to be split.

There are also cases in which the dependency annotations have to violate the term boundary. For example, in the phrase *more mature cell lines,* it is clear that *more* modifies *mature* while *mature cell lines* is annotated as a term. In such cases, the term is declared, with a note that it is unused, but not used in the rasp section of the annotation.

4.4 Other difficult constructions

In the checking process, several constructions with more frequent note annotations indicating that the annotation is questionable are found. These are the distinction between PP-complement and modifier, coordination, and PP-attachment. Coordination and PP-attachment involves the deep understanding of the context, and are being cleared with the biologist in the GENIA project on a case-by-case basis. Complement-modifier distinction and some cases of coordination were confusing to the annotator, partly due to the GR annotation scheme.

4.4.1 Complement vs Modifier

Nominalized verbs appear frequently in a biomedical text. Their complements are in the form of PPs, and are thus difficult to distinguish from modifiers by using the form as a clue. As to how to determine whether a particular PP is not clearly documented in (Briscoe, 2006), the PTB-GR corpus was used for reference to similar verbs. In some cases, especially in cases involving 'biomedical' verbs like *phosphorylate* and *transcribe*, the predicate-argument relations annotated in the GENIA event corpus were referenced. However, this means that the annotation criteria is more semantic compared to PTB-GR, where, for example, the semantic subject denoted by the *by*-phrase in passive constructions are annotated as ncmod, while the semantic subject denoted by *of*-phrase accompanying action-related nouns (e.g. *a big buyer of the high-risk, high-yield issues*) are annotated as iobj

4.4.2 Coordination

The case of coordination involving ellipsis, like *I kappa B alpha and RelA (p65) -mediated induction of the c-rel gene* and *the phosphorylation and rapid proteolytic degradation of I kappa B alpha,* seems to be frequent in the abstracts of scientific papers.

a) (conj and CD4(+))
(conj and CD8(+))
(ncmod *T lymphocytes* and _)
b) (ncmod ellip CD4(+) _)
(ncmod *T lymphocytes* CD8(+) _)
(conj and ellip)
(conj and *T lymphocytes*)
c) (ncmod *T lymphocytes* CD4(+) _)
(ncmod *T lymphocytes* CD8(+) _)
(conj and CD4(+))
(conj and CD8(+))

Figure 8: Coordination

In addition to the difficulty of determining the structure without deep domain knowledge, there is a problem that arises from the descriptive power of the GR scheme. The GR scheme does not have a mechanism that fully represents the distributive reading of coordination. For example, *CD4(+) and CD8(+) T lymphocytes* can be read in two ways: 1) the T lymphocytes that are CD4-positive and CD8-positive at the same time 2) the T lymphocytes that are CD4-positive and the T lymphocytes that are CD8-positive. The first interpretation can be straightforwardly encoded into Figure 8-a. However, the second interpretation is hard to encode in the GR scheme. In PTB-GR, phrases like *the London and Tokyo markets* (meaning *the London market and the Tokyo market*) are encoded into a structure like Figure8-a.

The GR scheme does have a mechanism to encode ellipsis, by placing *ellip* in the place of ellipsis, so that the second interpretation can encode a structure as in Figure 8-b. However, the problem is that there is no way to anchor the ellipsis to indicate what is omitted.

Alternatively, the second interpretation can be encoded into Figure 8-c, but here the entities that are coordinated is quite different from the ones that are predicted from the widely used phrase structures. In the actual annotation, the structures like Figure 8-b and 8-c coexist in the corpus, although annotation is done by only one annotator.

5. Discussions

Although in the abstracts sentences are longer, and, as a consequence, tend to have more complex structure, the declaration of terms has alleviated the difficulty by removing the dependencies inside those terms, which are often NPs whose internal structures are not relevant. This is usually consistent to the policy of other corpora like BioInfer (Pyysalo et al., 2007) where pre-modifier attachment in noun phrases without internal coordination is not resolved.

The concept of 'terms' in the GENIA annotation is broader than named entities and the terms sometimes include post-noun modifiers and, even in some cases, coordination. This is the major source of the inconsistency of boundaries as described in section 4.3. The particular example in the section needs the re-annotation on the side of the term corpus, but there is no simple solution such as excluding adjectives from the terms, as there are in examples like *simian immunodeficiency virus* that has very similar structure but regarded as a name (of a virus). Similarly, as there are terms like *signal transducers and activator of transcription,* which are generally recognized as a name of a protein, even post-noun modifiers or coordination cannot be blindly separated from terms.

As a separate process from the annotation of the structure outside the terms, we have compiled a list of GENIA terms appearing in the 50 abstracts, and have annotated the dependency structure inside them where possible. The result of this annotation is not included in the current annotation, but once a policy of breaking up terms into smaller components is established, they can be included in the corpus. The list of annotated terms is also expected to work as a 'gazetteer' when we expand the annotation to a larger set in order to avoid the problem of inconsistency, as structure annotation is especially more difficult in the domain-specific terms for prospective annotators who are basically linguists.

As for the other inconsistency with the existing GENIA, namely, the tokenization problem, it would be the POS corpus that should be corrected. The tokenization policy of the GENIA POS corpus follows that of the PTB corpus for compatibility and for ease of combining PTB and GENIA in training POS taggers. However, it has been pointed out that the PTB tokenization policy is not suitable, and finer-grained tokenization is necessary for biomedical texts (Yamamoto et al., 2004, Kulick et al., 2004, Tomanek et al., 2007). For the precise dependency annotation, changing tokenization policy is necessary. Using xml scheme dual annotation of PTB-based and finer-grained tokenization could be possible in the GENIA POS corpus, which should be done.

The source of other difficulties is that the dependency annotated with the scheme is not a shallow syntactic structure, and is also not a fully semantic structure. As a result, information that can be annotated cannot be rich enough to express some kind of structure that can be expressed in other formalisms. For example, the PTB scheme has the mechanism to co-index the NULL element into a word or a phrase that is supposed to be in the place of NULL. A similar mechanism may be incorporated into the GR scheme to fully annotate the ellipsis, and that would be convenient for annotating the coordination such as the one mentioned in section 4.4.2.

In the two cases of X Y-ing Z structure mentioned in the section 4.2.3, the current annotation scheme loses the information that X is an argument of Y. In addition, the current scheme loses the distinction between *NF kappa B binding domain,* where both *NF kappa B,* and *domain* are arguments of *binding*, and *NF kappa B binding activity* where *activity* is not an argument, but an appositive (i.e., *NF kappa B binding = activity*). These can be distinguished if xmod instead of ncmod is used as in the following examples illustrated in Figure 9.

NF kappa B binding domain (dobj binding *NF kappa B*) (xmod _ domain binding) (ncsubj binding domain _)
NF kappa B binding activity (dobj binding *NF kappa B*) (xmod ta activity binding)

Figure 8. *NF kappa B binding domain* vs *NF kappa B binding activity*

On the other hand, due to the specialized nature of the text, it will be difficult to correctly annotate the precise structure without deep domain knowledge, and also, an annotation that is semantically too precise may not be suitable for comparison between parsers with various levels of depth in analysis. A tradeoff between the richness of the information and convenience for evaluation must be established, with informative documentation thereof. A possible solution may be using the underspecification of relation types in evaluation, but whether underspecification involving complex structures (e.g, unifying the three annotations in Figure 8) can be achieved in a straightforward way is yet to be discussed.

Another kind of problem involving the descriptive power of the annotation scheme is that there are parts of scientific text that has different syntax from ordinary English, like references and mathematical expressions. In the abstracts, the cases are few, but if the corpus should be enhanced to annotate full text of scientific papers, more references or expressions may have to be annotated. As for the reference, one solution is to declare them like we did for named entities and not analyze the inside, at least at the parsing stage. However, it would not be appropriate to do so for mathematical expressions, because sometimes an expression behaves like a (embedded) sentence. For example, in a mathematics textbook, it is usual for one to encounter sentences like *We can find integers a and b such that $ma-mb=1$* and $T_n<T_{n+1}$ *If $n>0$.* The use and syntax of embedded mathematics (and also chemical) expression would be worth investigating seriously with more examples.

6. Conclusions and future works

We have created a prototype version of a dependency corpus from the subset of GENIA. Currently, the corpus is in the cleaning up process. At the same time, the annotation scheme is being refined with regards to the difficult cases, including the ones reported in the paper.

After the cleanup with the refined scheme, we plan to publish the set with the manual enhanced with the 'local' guidelines, from the GENIA web site [6]. Annotation experiments involving more than one annotator, and enhancing the annotation to full texts from abstracts are in plan.

Also, we are interested in the comparison of our dependency corpus with the annotation created by manually correcting automatically created, publicly-available dependency corpora derived from the GENIA corpus, e.g., DepGENIA (Rinaldi et al., 2005). This would help checking if the current annotation is truly formalism-independent.

7. Acknowledgements

This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan), and Genome Network Project (MEXT, Japan).

The authors are grateful to Tomoko Ohta for her suggestions and comments as a biologist.

8. References

Briscoe, E. (2006). An introduction to tag sequence grammars and the RASP system parser, Technical Report (UCAM-CL-TR-662), Cambridge University Computer Laboratory.

Briscoe, E., Carroll, J. (2006). Evaluating the Accuracy of an Unlexicalized Statistical Parser on the PARC DepBank. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Sydney, Australia, pp. 41-48.

Briscoe, E., Carroll, J., Watson, R. (2006). The Second Release of the RASP System. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, Sydney, Australia, pp. 77-80.

Carroll, J., Briscoe, E., and Sanfilippo, A. (1998). Parser evaluation: a survey and a new proposal. In *Proceedings of the LREC 1998*, Granada, Spain, pp. 447-454.

Clark S., Curran, J. (2007). Formalism-Independent Parser Evaluation with CCG and DepBank. In *Proceedings of ACL 2007*, Prague, Czech Republic, pp.248-255.

Clegg, A.B., Shepherd, A.J. (2007). Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics,* 8:24.

Hara, T., Miyao, Y., Tsujii, J. (2007). Evaluating Impact of Re-training a Lexical Disambiguation Model on Domain Adaptation of an HPSG Parser. In *Proceedings of IWPT 2007*, Prague, Czech Republic.

Kaplan, R., Riezler, S., King, T., Maxwell, J., Vasserman, A., Crouch., R (2004). Speed and accuracy in shallow and deep stochastic parsing. In *Proceedings of HLT/NACL 2004*. Boston, Massachusetts, USA, pp. 97-104.

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

Kim, J-D., Ohta, T., Tsujii J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10.

King, T.H., Crouch, R., Riezler, S., Dalrymple, M., Kaplan, R.M. (2003). The PARC 700 Dependency Bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora*, held at EACL03, Budapest

Kulick S., Bies A., Liberman M., Mandel M., McDonald R., Palmer M., Schein A., Ungar L. (2004). Integrated Annotation for Biomedical Information Extraction.In *Proceedings of Biolink 2004*. pp. 61-68

Marcus M.P., Santorini B., Marcinkiewicz M.A. (1993). Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, Vol.19, No.2, pp. 313-330.

de Marneffe, M-C., MacCartney, B., Manning, C.D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*, Genoa, Italy, 2006,..

Miyao, Y., Sagae. K., Tsujii, J (2007). Towards Framework-Independent Evaluation of Deep Linguistic Parsers. In *Proceedings of Grammar Engineering across Frameworks,* Stanford, California, USA, pp. 238-258.

Oda, K., Kim, J-D., Ohta, T., Tateisi, Y., Tsujii, J. New challenges for text mining: Mapping between text and manually curated pathways, In *Proceedings of LBM 2007*, Singapore, Singapore.

Ohta, T., Tateisi, Y. Kim, J-D., Yakushiji A., and Jun-ichi Tsujii, J. (2006) Linguistic and Biological Annotations of Biological Interaction Events. In *Proceedings of LREC 2006*, Genoa, Italy, 2006, pp. 1402-1405.

Preiss, J. (2003). Using grammatical relations to compare parsers. In *Proceedings of EACL 03*, Budapest, Hungary, pp. 291–296.

Pyysalo, S., Ginter, F., Laippala, V., Haverinen, K., Heimonen, J., Salakoski, T. (2007). On the unification of syntactic annotations under the Stanford dependency scheme: A case study on BioInfer and GENIA. In *Proceedings of BioNLP Workshop at ACL 2007*, Prague, Czech Republic, pp. 25-32.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics* 8:50.

Rinaldi, F., Schneider, G., Kaljurand K., Hess, M., Andronis, C., Persidis, A., Konstanti, O. (2005). Relation Mining over a Corpus of Scientific Literature. In *Proceedings of AIME 2005*, Aberdeen, Scotland, pp. 550–559.

Tomanek, K., Wermter, J., Hahn, U. (2007). A reappraisal of sentence and token splitting for life science documents., *Medinfo* 12(Pt1), pp. 524-528.

Yamamoto, K., Kudo, T., Konagaya, A., Matsumoto, Y. (2004). Use of morphological analysis in protein name recognition, *Journal of Biomedical Informatics,* Vol 37, Issue 6, pp. 471-482.

[6] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/