BIOINFORMATICS

Vol. 19 Suppl. 1 2003, pages i180–i182
DOI: 10.1093/bioinformatics/btg1023

GENIA corpus—a semantically annotated corpus for bio-textmining

J.-D. Kim[1,*], T. Ohta[2], Y. Tateisi[1] and J. Tsujii[1,2]

[1]CREST, Japan Science and Technology Corporation, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan and [2]Department of Computer Science, University of Tokyo, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: Natural language processing (NLP) methods are regarded as being useful to raise the potential of text mining from biological literature. The lack of an extensively annotated corpus of this literature, however, causes a major bottleneck for applying NLP techniques. GENIA corpus is being developed to provide reference materials to let NLP techniques work for bio-textmining.

Results: GENIA corpus version 3.0 consisting of 2000 MEDLINE abstracts has been released with more than 400 000 words and almost 100 000 annotations for biological terms.

Availability: GENIA corpus is freely available at http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA.

Keywords: Text Mining, Information Extraction, Corpus, Natural Language Processing, Computational Molecular Biology

INTRODUCTION

Text mining from biological literature is emerging as one of the main issues in bioinformatics research, and NLP methods are regarded as being useful to raise the potential of text mining from this literature. While the techniques are relatively domain-portable, reference materials, e.g. corpora, are not. The lack of an extensively annotated corpus can thus be seen as a major bottleneck for applying NLP techniques to bioinformatics.

GENIA corpus, a semantically annotated corpus of biological literature, is being compiled and annotated in the scope of GENIA project. It is aiming at providing high quality reference materials to let NLP techniques work for bioinformatics and at providing the gold standard for the evaluation of text mining systems.

Recently, we released GENIA corpus version 3.0. It consists of 2000 abstracts taken from MEDLINE database, and contains more than 400 000 words and almost 100 000 annotations that have been hand-coded for biological terms.

This paper is intended to provide a general introduction to GENIA corpus. Brief statistics on the annotations made in the corpus are also given.

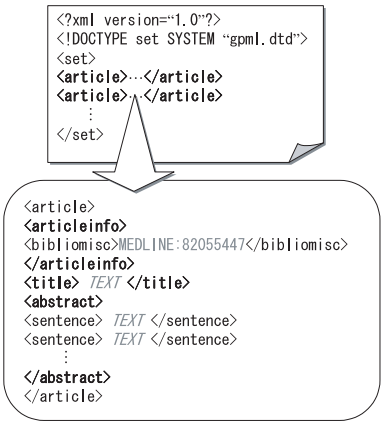

Fig. 1. Configuration of GENIA corpus.

GENIA CORPUS

GENIA corpus is a collection of articles extracted from MEDLINE database. Since we wanted our annotation work to converge on biological reactions concerning transcription factors in human blood cells, we selected articles with the MeSH terms, *human*, *blood cell* and *transcription factor*.

In GENIA corpus, the articles are encoded in an XML-based mark-up scheme† where each article contains its MEDLINE ID, title and abstract in that order and all the texts in the abstracts are segmented into sentences, resulting in the configuration illustrated in Figure 1.

The main value of the GENIA corpus comes from its annotation: all the abstracts and their titles have

*To whom correspondence should be addressed.

† For more detailed description, see Kim *et al.* (2000).

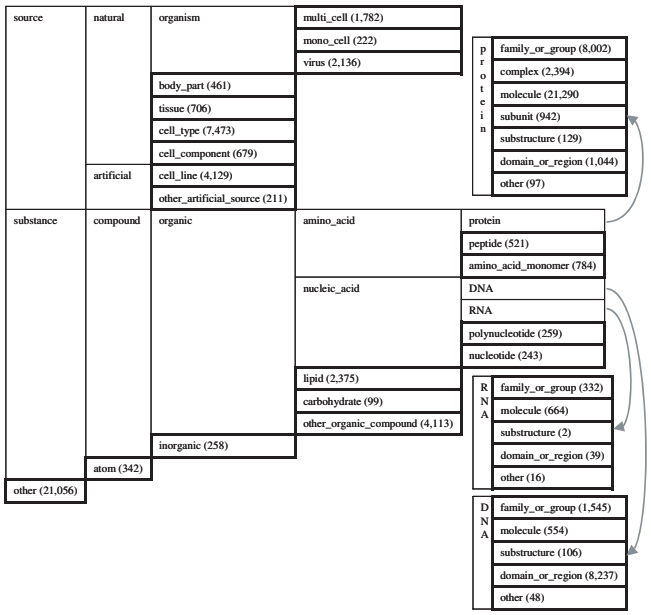

Fig. 2. GENIA ontology and statistics from GENIA corpus.

been marked-up by two domain experts for biologically meaningful terms, and these terms have been semantically annotated with descriptors from the GENIA ontology.

GENIA ONTOLOGY

GENIA ontology is a taxonomy of, currently, 47 biologically relevant nominal categories[‡]. Figure 2 shows the hierarchy of GENIA ontology, where the leftmost column lists the top three concepts, *biological source*, *biological substance* and *other*. Here, the *other* is not actually a biological concept but is prepared for the terms that are regarded as biological concepts but are not identified with any other concepts in the ontology. More classified concepts are listed on the right to the respective super concepts. The concepts in bold boxes are the terminal concepts and they form the actual tag set for semantic annotation.

GENIA ontology is encoded in DAML+OIL[§], an XML-based ontology language and is provided together with GENIA corpus.

LINGUISTIC ANNOTATION

In the GENIA corpus, biological terms are annotated. Semantically, they are defined as the terms identifiable with any terminal concepts in GENIA ontology. Yet syntactically, they are not simply defined. Often, terms, especially biological entity names, are compared to names

[‡] For detailed description on the categories, see Ohta *et al.* (2002).
[§] http://www.daml.org/

```
〈term〉 := 〈qualifier〉* 〈head noun〉
〈qualifier〉 := 〈adjective〉|〈noun modifier〉
```

Fig. 3. Syntactic definition of term.

```
<cons sem="G#other_name"><cons sem="G#DNA_domain_or_region">IL-2 gene</cons>
transcription</cons> in <cons sem="G#cell_type">T cells</cons>
```

Fig. 4. Annotating terms.

of named entities, like names of humans, organizations, etc., but from the linguistic perspective, they are quite different. First of all, names of named entities are mostly proper nouns, which means they won't be attributed by any specifiers or qualifiers[¶], while terms including biological entity names are mostly general nouns and they can appear in text with a variety of specifiers or qualifiers. So a policy on inclusion or exclusion of specifiers and qualifiers in terms is required.

In our annotation scheme, specifiers are not included in terms and the inclusion of qualifiers is left to the experts judgment. It may seem arbitrary, but we are expecting we could collect statistics on experts' tendencies concerning the in/exclusion and based on the statistics we could set a more elaborate scheme for future annotations or revision. Based on our annotation scheme, the term is syntactically defined as in Figure 3.

Mostly, terms can be annotated by simply inserting mark-ups around them as exemplified in Figure 4, where three terms of *IL-2 gene*, *IL-2 gene transcription* and *T cells* appear in the text and are marked-up by being enclosed by **cons** tags. Note that *IL-2 gene* is recursively annotated inside *IL-2 gene transcription*.

However, when terms appear in coordinated clauses involving ellipsis, the annotation is not simply achieved, since we cannot find all the terms fully spelled at the surface of the text. For example, the text, '*CD2 and CD 25 receptors*' refers to two terms, *CD2 receptors* and *CD25 receptors*, but *CD2 receptors* doesn't appear in the text. In such case, by revealing the shared part (*receptors*) and the coordinated parts (*CD2* and *CD25*), and then constructing a linguistic parse on them, we can annotate all the terms at a higher level than the surface, which is shown in Figure 5. Figure 6 shows another example of (extremely) coordinated clause and the annotation on it[‖].

As you may perceive from the above examples, according to our annotation scheme, even terms spelled in full

[¶] Specifiers include ordinals, cardinals and determiners. Qualifiers include adjectives and noun modifiers. We follow Allen (1995)'s definition of specifiers and qualifiers.
[‖] In this example, lines are aligned and indented to make it easy to read.

```
<cons sem="(AND G#protein_molecule G#protein_molecule)"><cons>CD2</cons> and
<cons>CD25</cons> <cons>receptors</cons></cons>
```

Fig. 5. Annotating terms involving ellipsis in coordinated clauses—simple case.

```
<cons sem="(AND (AND G#other_name G#other_name) (AND G#other_name G#other_name))">
        <cons sem="(AND G#cell_type G#cell_type)">
                <cons>B</cons>
                and
                <cons>T</cons>
                <cons>lymphocyte</cons>
        </cons>
        <cons>activation</cons>
        and
        <cons>mitogenesis</cons>
</cons>
```

Fig. 6. Annotating terms involving ellipsis in coordinated clauses—extreme case.

```
B and <cons sem="G#other_name"><cons sem="G#cell_type">T lymphocyte </cons>
activation</cons> and mitogenesis
```

Fig. 7. Result of leaving annotations for terms on surface.

in text, like *CD25 receptors* in Figure 5 or *T lymphocyte* and *T lymphocyte activation* in Figure 6, can be broken up into pieces. It may seem strange, but the underlying idea is that because, e.g. the *receptors* in the example of Figure 5 is not only the head noun of *CD25 receptors* but also the head noun of *CD2 receptors*, it should be treated differently from the case when it appears in non-coordinated clauses.

We, however, see that there may be high demand for simple annotations that reveal just the terms appearing at the surface of text. Thus, we provide a software tool that removes all the higher level annotations from GENIA corpus, leaving only the surface level terms. Through the tool, the annotation in Figure 6 can be changed into that in Figure 7.

STATISTICS

Table 1 lists the basic statistics on GENIA corpus and Table 2 shows the number of annotated objects in the corpus. The number of total annotations (**cons** elements) made in GENIA corpus is 96 582, among which 89 682 are made for surface level terms, 1583 are for higher level terms that are actually involving 3431 terms, and 5137 are made just for identifying the building blocks (**cons** elements without **sem** value) for higher level annotations. The total number of recovered terms is 93 293. The detailed number of annotated terms are shown in Figure 2.

Table 1. Basic statistics on GENIA corpus

	number	average	
Abstract	2 000		
Sentence	18 545	9.27s/a	
Word	436 967	218.48w/a	23.56w/s

Table 2. statistics on annotated objects

	No of cons	No of terms
Simple	89 862	89 862
Complex	1 583	3 431
No mean	5 137	0
Total	96 582	93 293

On the other hand, from the surface version of GENIA corpus obtained through the software tool mentioned above, 91 569 terms are identified, which are roughly 2% more than the surface terms identified from original GENIA corpus but roughly 2% less than the total terms annotated in GENIA corpus.

CONCLUSION

GENIA corpus 3.0 has been released with linguistically rich annotations including sentence boundaries, term boundaries, term classifications, semi-structured coordinated clauses, recovered ellipsis in terms, etc. We hope to encourage many researchers to make use of GENIA corpus for their research, and expect much feedback from them that would be the most valuable source for further improvement of the corpus.

ACKNOWLEDGEMENT

The research is partially supported by Genome Information Science Project (MEXT, Japan) and Information Mobility Project (CREST, JST, Japan).

REFERENCES

Kim,J.-D., Ohta,T., Tateisi,Y. and Tsujii,J. (2001) XML-based linguistic annotation of corpus. In *Proceedings of the first NLP and XML Workshop*. pp. 44–53.

Ohta,T., Tateisi,Y. and Kim,J.-D.C. (2002) The GENIA Corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the Human Language Technology Conference*. to be appeared.

Allen,J. (1995) *Natural Language Understanding*. Benjamin Cummings, Redwood City, pp. 25–28.