Linguistic and Biological Annotations of Biological Interaction Events

Tomoko Ohta[1], Yuka Tateisi[1], Jin-Dong Kim[1], Akane Yakushiji[2], Jun-ichi Tsujii[1,2,3]

[1] Japan Science and Technology Agency
[2] Department of Computer Science, University of Tokyo
[3] School of Informatics, University of Manchester
{okap,yucca,jdkim,akane,tsujii}@is.s.u-tokyo.ac.jp

Abstract

This paper discusses an augmentation of a corpus of research abstracts in biomedical domain (the GENIA corpus) with two kinds of annotations: tree annotation and event annotation. The tree annotation identifies the linguistic structure that encodes the relations among entities. The event annotation reveals the semantic structure of the biological interaction events encoded in the text. With these annotation we aim to provide a link between the clue and the target of biological event information extraction.

1. Background

Information extraction in biomedical field has become a widely researched application of natural language technologies. Convincing results of named entity extraction have been reported (e.g. Zhou et al. 2004, Kou et al. 2005) and now research focus is shifting to extraction of verbal information such as relations, interactions, and other events between entities such as proteins and genes.

Traditionally, events and relations are extracted using patterns on surface text around a certain sets of verbs (Sekimizu et al. 1998, Ono et al. 2001, Blaschke et al. 2002). However, in natural language text, the same relation can take various syntactic forms. For example, the event that a substance A activates another substance B can be expressed in phrases like *A activates B, B is activated by A, A is an activator of B, activation of B by A,* etc., so that building patterns to cover all the possible syntactic variations is a time-consuming work.

Recently, to overcome this problem, more strategic and systematic analysis using deeper NLP techniques has been suggested. One of the promising strategies is using deep parsers which can abstract the syntactic variation of relations between verbs and their arguments (predicate-argument structure) in the text, and constructing extraction rules on the abstracted structures (Temkin et al. 2003, Daraselia et al. 2004, Kim et al. 2004, Ahmed et al. 2005, Yakushiji et al. 2005). By identifying the predicate-argument structure, the variety of expressions in the example above is normalized into a triplet like (*verb, logical-subject, logical-object*)=(*activate, A, B*). The relation between *A* and *B* can be extracted using the normalized structure, so that the number of extraction rules or patterns necessary for extraction can be reduced.

For this approach to be successful, a robust parser with high accuracy is necessary. Also, a mechanism for building patterns for information extraction efficiently is required. As recent advances in NLP technology depend on machine-learning techniques, annotated corpora from which system can acquire rules including grammar and lexicon for parsing and patterns for information extraction are indispensable resources. We are annotating syntactic tree structure and interaction events on biomedical research abstracts. The tree structure can capture the syntactic relation of verbs and their arguments, and the interaction event annotation explicitly marks the target of information extraction. In other words, the two annotations can reveal the linguistic and biological aspects of the events described in research abstracts. With the annotation of the both aspects on a same set of text, the corpus can be a useful resource for integrated systems such as information extraction using deep linguistic analysis. Similar effort of corpus construction is being done at the University of Pennsylvania in the different subdomain of biomedicine (Kulick et al. 2004).

2. Overview of the Corpus

The corpus discussed in this paper is an augmentation of the GENIA corpus (Kim et al., 2003). The base text of the corpus consists of research articles indexed in the MEDLINE database[1] concerning transcription factors in human blood cells. While a MEDLINE record has various meta-data such as author names and publication date, only the title, abstract text, and identification number (PMID) are retained[2] and encoded in XML. All the text in the title and abstract is segmented into sentences and annotations are done on these sentences. The purpose of the annotation is two-fold: to make the biomedical knowledge encoded in the text transparent and to reveal the syntactic structure behind the text. Eventually, our objective is to establish the mapping between the knowledge pieces and the linguistic structures. In the public version of the GENIA corpus version 3.02p, 1999 abstracts are annotated with technical terms (including named entities) and parts of speech. The two annotations can serve as the target and the linguistic clue of named entity extraction.

The two new annotations, tree annotation and event annotation, are augmentations of the GENIA corpus concerning the verbal aspect of information extraction. The two annotations are done independently of each other. We plan to merge the two corpora at a later stage. The major reason for separate annotation is that we can use (variations of) existing annotation schemes so that the systems built on existing corpora can be easily applied to the new corpus. In addition, merging can highlight the

[1] http://www.nlm.nih.gov/pubs/factsheets/MEDLINE.html

[2] In XML terms, the base of the corpus is the *ArticleTitle*, *AbstractText* and *PMID* elements, and their ancestors to preserve the path to those elements in the original MEDLINE records. All the original attributes of the retained elements are removed because the purpose of retaining the elements is for preserving the path information.

problems of applying existing schemes. For example, the crossing of element boundaries are reported in merging part-of-speech and term corpora (Tateisi et al, 2004) and tree and entity corpora (Bies et al, 2005), and led to the

<sentence id= “S2”><cl cat=“S”><cl cat=“PP”>**In** <cl cat=“NP”>**the present paper**</cl></cl>**,** <cl cat=“NP” role=”SBJ” id="i55"><cl cat=“NP”>**the binding**</cl><cl cat=“PP”>**of**<cl cat=“NP”>**a [125I]-labeled aldosterone derivative**</cl></cl><cl cat=“PP”>**to** <cl cat=“NP”><cl cat=“NP”>**plasma membrane rich fractions**</cl><cl cat=“PP”>**of HML**</cl></cl></cl></cl><cl cat=“VP”>**was** <cl cat=“VP”>**studied** <cl cat=“NP” null="NONE" ref="i55"/></cl></cl>.</cl></sentence>

(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. A sentence with tree-annotation in XML and PTB formats.

revision of criteria on tokenization and constituent boundary. We expect this sort of problems that lead to improvement of the annotation scheme arise in integration.

3. Tree Annotation

The tree annotation (a.k.a GENIA Treebank, GTB) reveals the syntactic structure of text. Overall, the annotation has been performed following the Penn Treebank II (PTB) annotation scheme (Beis et al, 1995) widely used in natural language processing community. We have made some modification, mainly simplification, in order for non-biologists to consistently annotate the structure without deep knowledge in the domain.

The modifications we made to the PTB annotation scheme mainly involves the treatment of noun phrases and function tags. The modifications that involve noun phrases are:

- Labels NX and NAC are not used. The phrases that are labeled as NX (the head of a complex noun phrase) or NAC (a prenominal modifier that is not a constituent) are labeled as NP.
- The internal structure of a noun phrase may be left unstructured unless coordination is involved.

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. On the other hand, the internal structure of the noun phrases is 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 exception is the cases that involve coordination where it is necessary to explicitly mark up the coordinated constituents.

Some function tags in PTB are not used in the current version, although we plan to use them in the future. Currently, only the function tags that identify the case elements (e.g. SBJ for the surface subject and LGS for the logical subject of passive sentences), displaced constituents (e.g., TPC for topicalized elements) and temporal relation (TMP) of adverbial phrases. This is, again, in order to simplify the annotation process. Those function tags unused in GTB are in semantic nature (MNR: manner, etc) which are not supposed to be easy for non-biologist to decide.

There is one modification with which we have richer information than the original Penn Treebank scheme. In GTB, coordination is always explicitly marked. Establishment of coordination structure is crucial to the construction of semantic structure but coordination structure is often syntactically ambiguous. The explicit marking helps training of machine-learning-based parsers and construction of heuristic rules to resolve syntactic ambiguity.

In the XML encoding, a constituent (clause) is delimited into a <cl> element whose *cat* attribute represents its syntactic category. A null constituent is marked as a childless element. Other function tags are encoded as attributes. Figure 1 shows an example of annotated sentence in XML, and the corresponding PTB notation. The *cat* attribute indicates the syntactic category of the constituent: the value “S” means sentence, “NP” noun phrase, “PP” prepositional phrase, and “VP” verb phrase. The *role* attribute is for grammatical roles (cases), and the value “SBJ” means that the element serves as the (surface) subject of the sentence. A null element, that is, the trace of the object of *studied* moved by passivization, is denoted by “<cl cat=“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 the PTB notation.

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 annotators.

The annotation of GTB is done by annotators with linguistic knowledge but without expert level of biological knowledge. Such annotators can determine the syntactic structure quite consistently, as a small inter-annotator agreement test on 10 abstracts showed the agreement rate between two such annotators was 94.5 % (Tateisi et al, 2005).

4. Interaction Event Annotation

Event annotation of GENIA is performed to identify biomedical events mentioned in natural language text. We first performed preliminary annotation for 500 MEDLINE abstracts according to the event annotation scheme developed by the Caderige project (Alphonse et al., 2004). Then, based on the experience and statistics from the preliminary annotation, we redefined the scheme.

In our current scheme, a biological *event* is defined as a temporal occurrence that happens to one or more

biological entities. Especially, a number of events which cause some specific change on genes or gene products (proteins) are defined in the GENIA event ontology (section 4.1), becoming the target of annotation. An event is associated with its *type*, *theme*s and *cause*s. The type of an event is defined as a class in the GENIA ontology. A *theme* is an object undergoing change during the event and a *cause* is an object causing the change.

Annotation of the events has been being done over the GENIA term annotation. Usually, biological entities marked up in the term corpus may become a *theme* or a *cause* of an event in the event corpus, but sometimes an event can be a *theme* or a *cause* of another event.

4.1. Event type and Ontology

In natural language text, events in the same event class may be referred to by different expressions. For example, the word *induces* in the sentence *Lipopolysaccharide induces phosphorylation of MAD3* …(PMID:8505309)[3] and the word *enhancement* in *Enhancement of human immunodeficiency virus 1 replication in monocytes by 1,25-dihydroxycholecalciferol. (PMID1650477)* refer to events that belong to the same class (*positive regulation*). To capture the variations of expressions referring to the same class, a controlled vocabulary of event descriptor is required. For this purpose, we defined a taxonomy of the classes of biological events (the GENIA event ontology).

- ● Artificial Process
- ▾ Biological Process (0008150)
  - ▾ Cellular Process (0009987)
    - ● Cell Communication (0007154)
    - ● Cell Differentiation (0030154)
  - ▾ Physiological process (0007582)
    - ● Localization (0051179)
    - ▾ Metabolism (0008152)
      - ▸ DNA metabolism (0006259)
      - ● Gene expression
      - ▸ Protein metabolism (0019538)
      - ▸ RNA metabolism (0016070)
      - ● Transcription (0006350)
  - ▾ Regulation (0050789)
    - ● Negative regulation(0048519)
    - ● Positive regulation (0048518)
  - ▾ Viral life cycle (0016032)
    - ● Initiation of viral infection (0019059)
    - ● Viral genome expression (0019080)
- ● Correlation
- ▾ Molecular function (0003674)
  - ● Binding (0005488)
  - ● Catalysis (0003824)

Figure 2. Upper level of the GENIA event ontology. The ● preceding a node indicates that the node is a leaf; ▾ indicates that it is an intermediate nodes whose children are shown in the figure; ▸ sign indicates that there are children not shown in the figure.

There are 37 concepts in the taxonomy organized in four subontologies: *artificial process*, *biological process*, *correlation*, and *molecular process*. Most of the concepts of *biological process* and the *molecular process* subontologies are taken from the subontologies of the same name in the Gene Ontology (GO) (Ashburner et al. 2000). Such concepts in these subontologies are associated with the accession number of the corresponding concepts in GO, and the parent-child relationships among the concepts are preserved. The other two subontologies (artificial process and correlation) are for the concepts of events not covered in GO but relevant to our annotation.

```
<sentence id="S2"><term id="T5">Mice</term> transgenic for the <term id="T6"><term id="T7">human T cell leukemia virus</term> (<term id="T8">HTLV-I</term>) <term id="T9">Tax</term> gene</term> develop <term id="T10">fibroblastic tumors</term> that express <term id="T11">NF-kappa B-inducible early genes</term>.</sentence>
<event id="E1">
  <type class="Cell_differentiation"/>
  <theme idref="T10"/>
  <clue>Mice transgenic for the human T cell leukemia virus (HTLV-I) Tax gene <clueType>develop</clueType> fibroblastic tumors that express NF-kappa B-inducible early genes.</clue>
</event>
<event id="E2">
  <type class="Gene_expression"/>
  <theme idref="T11"/>
  <cause idref="T10"/>
  <clue>Mice transgenic for the human T cell leukemia virus (HTLV-I) Tax gene develop fibroblastic tumors <linkCause>that</linkCause> <clueType>express</clueType> NF-kappa B-inducible early genes.
  </clue>
</event>
```

Figure 3. A sentence with Event annotation. The attributes of *term* elements except *id* are not shown for legibility.

Figure 2 shows the upper level of the ontology (up to depth 4; the maximum depth of the tree is 6). The number in the parentheses is the GO accession numbers corresponding to the node. The ontology is encoded in the Web Ontology Language (OWL) recommended by the World Wide Web Consortium (Bechhofer et al., 2004).

4.2. The Annotation Scheme

In the XML encoding, a sentence may be followed by one or more event elements each of which encodes an event mentioned in the sentence. The event element encodes the type, themes and causes of an identified event (Figure 3). For the type of an event, a descriptor from the GENIA event ontology may be specified. For the themes and causes of an event, the IDs of pre-annotated terms or events may be referenced. The *clue* element has been prepared in the event element to reveal the text parts which participate in mentioning the event. Inside the clue element, the text spans which are responsible for mentioning the type of the event, linking the event type to the themes and linking the event type to the causes are

[3] The number given in the parenthesis is the PMID of the source abstract.

marked-up as the *clueType*, *linkTheme* and *linkCause* elements respectively.

For example, in the sentence shown in Figure 3, the event E1 identifies the event of tumor development which has been classified as a Cell_differentiation event of which the theme is the *fibroblastic tumors*. The text span *develop* is determined to give a clue for the event classification.

The event E2 identifies an event of Gene_expression whose theme is *NF-kappa B-inducible early genes* and cause is *Mice*. The text span *express* has been determined to give a clue for determining the event class and *that* to link the cause and the event.

5. Current Status of the Corpus

The tree annotation for the 500-abstract subset of the GENIA corpus, both in XML and in PTB format, has been made publicly available since June 2005 on our web site[4] as GENIA Treebank Beta Version (GTB-Beta). Recently we have enhanced the volume to 1500.

As to the event annotation, initial annotation for the same 500-abstract set as GTB-Beta is completed. We will make the set publicly available after error corrections.

6. Concluding Remarks

We have annotated the linguistic (tree structure) and biological (interaction) aspects of verbal information in biological domain, on the GENIA corpus. In tree annotation, we basically followed the Penn Treebank scheme widely used in the natural language processing community. In event annotation, we have defined a new scheme based on the one used by the Caderige project. A subset of 500 abstracts of the GENIA corpus is annotated for both tree and event.

So far, the two annotations are done independently, but future works include the integration or merging two annotations into one. Another work in the future is annotation of deeper predicate-argument information such as one produced by HPSG parsers or one annotated in Propbank (Kingsbury et al, 2002).

7. References

Ahmed S.T., Chidambaram D., Davulcu H., Baral C. (2005). IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text. In *Proceedings of Biolink 2005*. pp. 54-61.

Alphonse E., Aubin S., Bessieres P., Bisson G., Hamon T., Lagarrigue S., Nazarenko A., Manine A., Nedellec C., Vetah M., Poibeau T., Weissenbacher D. (2004). Event-based Information Extraction for the biomedical domain: the Caderige project. In *Proceedings of JNLPBA 2004*. pp. 43-49.

Ashburner M, Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M.,and Sherlock G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*. 25(1). pp. 25-29.

Bechhofer S., van Harmelen F., Hendler J., Horrocks I., McGuinness D.L., Patel-Schneider, P.F., Stein L.A., 2004. OWL Web Ontology Language Reference. http://www.w3.org/TR/2004/REC-owl-ref-20040210/.

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

Bies A., Kulick, S., Mandel M. (2005). Parallel Entity and Treebank Annotation, In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*. pp.21-28.

Blaschke C., Valencia A. (2002). The Frame-Based Module of the SUISEKI Information Extraction System. *IEEE Intelligent Systems.* 17(2). pp. 14-20.

Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*. 20(5). pp. 604-611.

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

Kim J-J., Park J-C. (2004). Bioie: retargetable information extraction and ontological annotation of biological interactions from the literature. *Journal of Bioinformatics and Computational Biology*. 2(3). pp. 551-568.

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

Kou Z, Cohen W.W., Murphy R.F. (2005). High-recall protein entity recognition using a dictionary. *Bioinformatics.* 21(suppl. 1). pp. i266-i273.

Kingsbury P., Palmer M., Marcus M. (2002). Adding Semantic Annotation to the Penn Treebank. In *Proceedings of HLT 2002*.

Ono T, Hishigaki H, Tanigami A, Takagi T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*. 17(2). pp. 155-161.

Sekimizu T., Park H.S., Tsujii J. (1998). Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts. *Genome Informatics*. pp.62-71.

Tateisi Y., Tsujii J. (2004). Part-of-Speech Annotation of Biology Research Abstracts. In *Proceedings of LREC 2004, IV*. pp. 1267-1270.

Tateisi Y., Yakushiji A., Ohta T. Tsujii, J. (2005). Syntax Annotation for the GENIA corpus. In *Proceedings of IJCNLP 2005, Companion volume.* pp. 222-227.

Temkin J.M., Gilder M.R. (2003). Extraction of protein interaction information from unstructured text using a context-free grammar.*Bioinformatics*. 19(16). pp. 2046-2053.

Yakushiji A., Miyao Y., Tateisi Y., Tsujii J. (2005). Biomedical Information Extraction with Predicate-Argument Structure Patterns. In *Proceedings of SMBM 2005.* pp. 60-69.

Zhou G-D. (2004). Recognizing Names in Biomedical Texts using Hidden Markov Model and SVM plus Sigmoid. In *Proceedings of JNLPBA 2004*. pp.1-7.

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