

# Annotation of Predicate-argument Structure on Molecular Biology Text

Yuka Tateisi Tomoko Ohta Jun-ichi Tsujii

CREST, JST

University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

{yucca, okap, tsujii}@is.s.u-tokyo.ac.jp

## Abstract

Annotated corpora are essential resources for natural language processing. This paper describes our approach for building a corpus annotated with predicate-argument structure on research abstracts in molecular biology domain. Observation of the records in a database of cell signaling events and corresponding research abstracts showed that extracting predicate-argument structure is a useful intermediate step for extracting reaction events but analysis of verb phrases is not sufficient because reactions and relations are often expressed in nominal phrases. Based on the observation, we try to annotate the predicate-argument structure in verbs including their nominalized forms in order to help develop extraction systems based on predicate-argument analysis.

## 1 Introduction

Corpora in which various linguistic information is annotated explicitly in structured way are essential resources for natural language processing (NLP) research. Part-of-speech corpora and treebanks have been developed and contributed to development of NLP systems as training and testing materials. For information extraction purpose, deeper knowledge than syntactic features, such as predicate-argument structure, is necessary, and the corpora carrying such information are also being developed. For example,

PropBank corpus (Kingsbury et al., 2002) adds predicate-argument structure to Penn Treebank as relation between nodes on the trees using predefined argument frames of verbs.

In recent years, information extraction systems in biomedical field using natural language technologies have been constructed. Convincing results of named entity extraction have been reported and now research focus is shifting to extraction of interactions and other events and relations between proteins and genes.

Traditionally such events and relations are extracted using patterns on surface text around a certain sets of verbs using systems such as POS taggers, regular expression matchers, and shallow parsers. Recently, due to limitation of the scope of verbs and expressions that pattern-based approach can handle, more strategic and systematic analysis using deeper NLP techniques are suggested. For example, a system involving full parsers is reported, which claims high performance (Temkin and Gilder, 2003). However, evaluation of results is difficult because there is no publicly available corpus based on research literatures in the domain and annotated with deeper structures.

In natural language sentences, an event or relation is expressed as a verb, and the participants involved are expressed as the arguments of the verb. Thus, predicate-argument structure is a useful intermediate structure for extraction for event or relation. This paper discusses the nature of predicate-argument structures which must be extracted for event information extraction in molecular biology domain, and a scheme for annotating such structures.

Nature. 1997 Mar 20;386(6622):296-9.  
**Activation of the transcription factor MEF2C by the MAP kinase p38 in inflammation.**  
 For cells of the innate immune system to mount a host defence response to infection, they must recognize products of microbial pathogens such as lipopolysaccharide (LPS), the endotoxin secreted by Gram-negative bacteria. These cellular responses require intracellular signalling pathways, such as the four MAP kinase (MAPK) pathways. ♡In mammalian cells the MAPK p38 is thought to play an important role in the regulation of cellular responses during infection through its effects on the expression of proinflammatory molecules. One means of understanding the role of p38 in these responses is to identify proteins with functions regulated by p38-catalysed phosphorylation. Here we demonstrate a link between the p38 pathway and a member of the myocyte-enhancer factor 2 (MEF2) group of transcription factors. ♣We found that in monocytic cells, LPS increases the transactivation activity of MEF2C through p38-catalysed phosphorylation. ♠One consequence of MEF2C activation is increased c-jun gene transcription. Our results show that p38 may influence host defence and inflammation by maintaining the balance of c-Jun protein consumed during infection.

A	Cell_Signaling : LPS $\mapsto$ MAPKp38 From_molecule LPS To_molecule p38-MAPK Effect activation Reference Han_1997 Role bacteria infection
B	Cell_Signaling : MAPKp38 $\mapsto$ MEF2 From_molecule p38-MAPK To_molecule MEF2 Effect activation Interaction phosphorylation Reference Han_1997
C	Gene_Expression : MEF2 $\mapsto$ c-Jun From_molecule MEF2 To_molecule c-Jun Effect activation Reference Han_1997 Tissue monocyte Tissue_System

Figure 1: An abstract of a research paper and related CSNDB records

## 2 How events are expressed in biomedical literature

The information that domain experts need to extract does not always appear straightforwardly in sentences. For example, in Cell Signal Network Database (CSNDB) (Takai-Igarashi and Kaminuma, 1999), a database of cell signaling events developed at National Institute of Health Science of Japan, the three records on the right-hand side of Figure 1 are extracted from the paper whose abstract is shown on the left-hand side. In the abstract, the title is relevant to Record B, Sentence ♣ is relevant to Records A and B, and Sentence ♠ is relevant to Relation C.<sup>1</sup> To extract the Records A and B from the sentence ♣ not only analysis of syntax and semantic structures but inference using the background knowledge on molecular biology (Figure 2).

The sentence ♣ (Figure 2-a) can be analyzed into the predicate-argument structure in Figure 2-b. However, the relations expressed in the extracted records are not expressed as relations between the

VERB and ARGs in the text. There are two problems. One is that aliases of the molecule names are used, and the other is that the relation is actually expressed in noun phrases in the form of nominalized verbs and their modifiers.

It is true that the word LPS, the name of a molecule in Record A appears as the subject of *that*-clause of the sentence, but the other two names, namely MAPKp38 (that appears in both records) and MEF2 (that appears in Record B), do not even appear in the sentence. The information that p38 means MAPKp38 does not appear explicitly in this abstract. Anaphora resolution involving the sentence ♡ is necessary to find that the word p38 in sentence ♣ actually means *MAPKp38*. On the other hand, one must resort to the domain knowledge that MEF2C is a member of a protein family MEF2 to find the name in the relation B.

Even when the relevant names of molecules were found, there is still a problem: in sentence ♣, the verb *increase* takes LPS2 as the subject (ARG1), but p38 and MEF2C do not appear as its object (ARG2). The word *MEF2C* appears as a part of ARG2, which is a noun phrase headed by *activity*. In the noun phrase the head-noun *activity* takes an

<sup>1</sup>The correspondences of sentences and extracted records are not explicitly stated in CSNDB records. Symbols in the abstract in the figure are added by us for explanation purpose.

argument (subject) as a form of *of*-phrase. On the other hand, the word *p38* is a part of the modifier (MOD), which is a prepositional phrase headed by *through*. It is not even in the argument position of the preposition *through*, but a part of the modifier of the argument (p38-catalyzed).

Construction like the sentence ♣ is not uncommon in abstracts. Especially, certain class of verbs that take reactions in the form of nominalized verbs as its objects (e.g. *induce*, *inhibit*) frequently appears where extraction of the reaction stated as the object of the verb is necessary. For example, in sentence *Inhibition of protein phosphatase 2A induces serine/threonine phosphorylation, subcellular redistribution, and functional inhibition of STAT3.*, STAT3 is inhibited as the result of phosphatase 2A inhibition but in *Treatment of human resting T cells with phorbol esters strongly induced the expression of IL-2R alpha and the activation of NF.kappa B.*, NF.kappa B is activated as the result of treatment of T cells. Thus, in information extraction from molecular biology research abstracts, it is not sufficient to extract the predicate-argument structure only from verb phrases, but extraction from other phrase involving nominalized verbs is crucial.

The observation above shows that a complicated process involving deep analyses of text is necessary for extracting reaction information. Yet, we believe that extraction of predicate-argument structure is an important intermediate step because the extracted predicate-argument structure can be translated into the information to be stored in a database with the help of domain knowledge in more systematic way than traditional methods that try to extract the information directly using surface pattern in sentences. For example, if the predicate-argument structure of not only verb phrases but other phrases involving nominalized verbs can be successfully extracted, one can draw a picture like Figure 2-c with the help of the knowledge about phosphorylation, in order to extract the information that LPS activates MAPKp38 and the resulting substance, the activated MAPKp38, catalyze the phosphorylation of MEF2C to increase its activity. Then, with the help of an ontology of entities, relations stated in Records A and B (Figure 2-d) can be extracted. As construction of ontologies of entities in molecular biology domain is rapidly advancing (e.g., (The Gene Ontology Con-

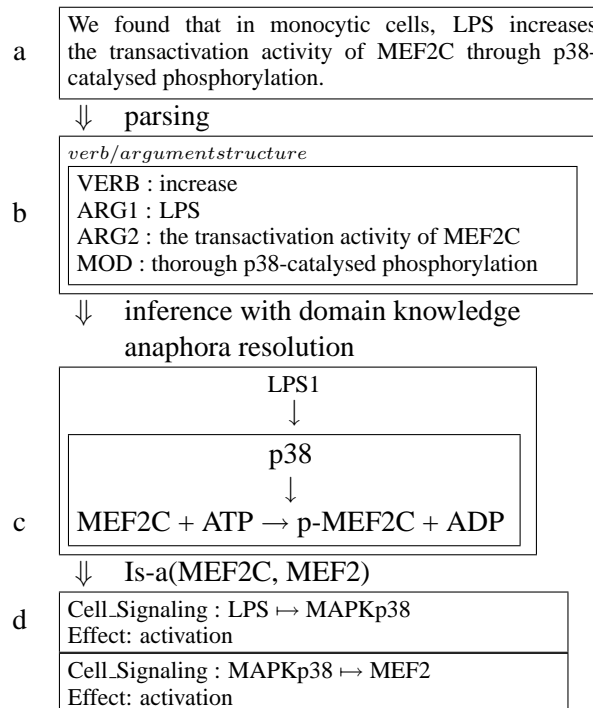


Figure 2: Deep analysis for extraction

sortium, 2004)), the construction of the last step will be relatively easy.

### 3 Annotation for Deep Structure

The previous section showed that the sentence in biomedical texts have quite a complicated structure and shallower analysis like pattern matching or (syntactic) parsing is not likely to be able to extract information in systematic way. Instead, we believe that first extracting predicate-argument structure and then extracting the necessary information using inference with domain knowledge is more feasible.

An obstacle for this approach is the lack of publicly available resources, e.g., lexicon of domain-specific verbs and corpus annotated with predicate-argument structure. Available corpora in this domain are mostly annotated with information relevant to names. For example in GENIA corpus (GENIA, 2003) technical terms including names of substances are annotated with their classification; in Medstruct corpus (Medstruct, 2001) abbreviations and aliases are annotated. As far as we know, no corpus annotated with verbal and structural information (i.e. treebanks or proposition banks) in this domain is

publicly available, but each research group is using their own small-scale corpus focused on events and relations of their interest. More common and general corpus is desired for evaluation of systems, and preferably it should be large enough to be used for training purpose.

#### 4 Annotation Strategies

We develop a corpus based on research abstracts taken from MEDLINE database, annotated with predicate-argument structures. We annotate for every verb in every sentence in a text. As is shown in Figure 1, not all the sentences are related to the information extracted from the text. However, what is information to be extracted depends on an application or a user. Also, identifying the sentence or sentences most relevant to the desired information from the entire document can be another research topic. Thus, it is useful if all the sentences are annotated. Considering the use of a parser which must analyze the structure of entire sentences, every verb in sentences should be annotated. Even if such syntactic reason can be ignored, limiting annotation to verbs directly relevant to domain-specific information such as *activate* or *phosphorylate* limits the usefulness of the corpus because of the verbs like *show* and *suggest* which shows the confidence of the information stated in the sentence.

Unlike PropBank annotation (Kingsbury et al. , 2002), we annotate the nominalized verbs including gerunds since the information in such nominal phrases is crucial as shown in the previous section. In addition, we do not assume that a corpus annotated with syntactic structures or a lexicon with predefined frame patterns is available. The main reason is the lack of such resources. As stated before, commonly-available treebanks in this specific domain is yet to be constructed. The lexicon that provides the argument frame patterns for domain-specific usage of verbs is also lacking. Even SPECIALIS lexicon of Unified Medical Language System (UMLS), the most known resource in biomedical domain developed at National Library of Medicine (National Library of Medicine, 2003), does not have semantic roles for arguments of verbs. We expect that such resource can be constructed in parallel to annotation process. As for the tree

structure, in addition to unavailability we do not regard the pre-annotation is essential. One reason for this is that full syntactic structure is not necessary for defining the relationship between the predicate and the arguments. The syntactic information that is essential is the boundaries of the predicate and each of arguments. Their internal structures need not be fully analyzed unless the arguments have verbal nature and have to be recursively analyzed into predicates and arguments. Another reason is that preferred background of annotators for this task is molecular biology rather than linguistics because of the semantic nature of the task. cannot expect the annotators to be experts in linguistics. Considering the situation, the annotation is expected to be two-staged: first domain-experts annotate the structure in rather informal, intuitive way, and then linguists translate the structure into more formal one. It is even possible that the annotation of predicates and arguments given by domain experts helps linguists to annotate the full syntactic structure of sentences by resolving possible ambiguity.

Based on the strategies we started to annotate MEDLINE abstracts using a preliminary scheme, a rather intuitive one. For each sentence, a record is created separately including relation ID, verb, subject, object, and other conditions (Figure 3). In each record, the value of ID is an ID of the relation, unique through a document, *verb* field the predicate in base verbal form, *subject* its logical subject, *object* its logical object, and *other* other arguments. At this stage we leave the classification of arguments as a rough one, and refine later with the argument frame defined through the process of annotation. Each field of records is linked with text using starting and ending positions.

The records and the positional information can be easily encoded in XML format. For example, an annotation is Figure 4 can be encoded into XML format as in Figure 5 where the sentence in the text is encoded in the <sentence> element and the annotated predicate-argument structure is encoded in the <structure> element. The correspondence between the <sentence> and the <structure> elements are expressed by the attributes *sid* and *ref-s*. The <e> element corresponds to the record of the table in Figure 4 where the value of the *sem* attribute corresponds to the value in the cells of the table. The *eid*

We found that in monocytic cells, LPS increases the transactivation activity of MEF2C through p38-catalysed phosphorylation.

ID	verb	subject	object	other1	other2
#1	find	we	#2		
#2	increase	LPS	#3	through #4	in mono-tonic cells
#3	transactivate	MEF2C			
#4	phosphorylate	MEF2C			
#5	catalyze	p38	#4		

Figure 3: Annotation on the sentence (Figure 2-a) in table form: the positional information is omitted in this figure.

HLA-DO was shown to block HLA-DM function.

ID	verb	subject	object	other1	other2
#1	show		#1		
#2	block	HLA-DO	HLA-DM function		

Figure 4: A simpler example of annotation in table form

attribute is used for referring to structure expressed in another record. The components of <e> elements and the corresponding segments of the sentence is linked by the *cid* and *ref-c* attributes.

```
<sentence sid=100><c cid=1>HLA-DO</c> <c cid=2>was shown</c> <c cid=3>to <c cid=4>block</c> <c cid=5>HLA-DM function</c></c>.</sentence>

<structures ref-s=100><e eid=1><verb sem=gshowh ref-c=2 /><obj sem=g#2h ref-c=3 /></e><e eid=2><verb sem=gblockh ref-c=4 /><subj sem=gHLA-DOh ref-c=1 /><obj sem=gHLA-DM functionh ref-c=5 /></e></structures>
```

Figure 5: Annotation of the sentence in Figure 4 in XML format

## 5 Conclusions

We have demonstrated the difficulty of extracting information on biological reactions and the need for corpora annotated with predicate-argument structure, including annotations on nominalized forms. We set annotation strategies based on the observations on database records from CSNDB and corresponding MEDLINE abstract. Currently we are in the stage of preliminary annotation trying to refine the annotation scheme.

## Acknowledgements

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.

## References

- Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding Semantic Annotation to the Penn Tree-Bank. *Proceedings of the Human Language Technology Conference*.
- The Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32: D258-D261.
- GENIA Project. 2003. GENIA Corpus Ver 3.02. <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>
- Medstrat Project. 2001. Initial Annotation Corpora. <http://medstrat.org/gold-standards.html>
- National Library of Medicine. 2003. UMLS Documentation 2003AC. <http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML>
- Takako Takai-Igarashi and Tsuguchika Kaminuma. 1999. A Pathway Finding System for the Cell Signaling Networks Database. *In Silico Biology*, 1:129-146.
- Joshua M. Temkin and Mark R. Gilder. 2003. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, 19(16):2046-2053.