Effect of Utilizing Terminology on Extraction of Protein-Protein Interaction Information from Biomedical Literature

Junko Hosaka
RIKEN
Genomic Sciences Center
Suehiro-cho 1-7-22
Tsurumi-ku, Yokohama,
Kanagawa, Japan
jhosaka@gsc.riken.go.jp

Judice L.Y. Koh
Institute for Infocomm Research
21 Heng Mui Keng Terrace,
Singapore 119613

judice@i2r.a-star.edu.sg

Akihiko Konagaya
RIKEN
Genomic Sciences Center
Suehiro-cho 1-7-22
Tsurumi-ku, Yokohama,
Kanagawa, Japan
konagaya@gsc.riken.go.jp

Abstract

As the amount of on-line scientific literature in the biomedical domain increases, automatic processing has become a promising approach for accelerating research. We are applying syntactic parsing trained on the general domain to identify protein-protein interactions. One of the main difficulties obstructing the use of language processing is the prevalence of specialized terminology. Accordingly, we have created a specialized dictionary by compiling on-line glossaries, and have applied it for information extraction. We conducted preliminary experiments on one hundred sentences, and compared the extraction performance when (a) using only a general dictionary and (b) using this plus our specialized dictionary. Contrary to our expectation, using only the general dictionary resulted in better performance (recall 93.0%, precision 91.0%) than with the terminology-based approach (recall 92.9%, precision 89.6%).

1 Introduction

With the increasing amount of on-line literature in the biomedical domain, research can be greatly accelerated by extracting information automatically from text resources. Approaches to automatic extraction have used co-occurrence (Jenssen, 2001), full parsing (Yakushiji, 2001), manually built templates (Blaschke, 2001), and a natural language system developed for a neighboring domain, with modifications e.g. regarding semantic categories (Friedman, 2001).

In order to extract information such as protein-protein interactions from scientific text, it is insufficient to check only co-occurrences. Constructing a satisfactory set of rules for full parser is quite complex and the processing requires a tremendous amount of calculation.

One of the main difficulties in using language processing in the biomedical domain is the prevalence of specialized terminology, including protein names. It is impossible to obtain a complete list of protein names in the current rapidly developing circumstances: notations vary, and new names are steadily coined. To bypass these problems, we start with words expressing interactions, and then seek the elements which are actually interacting, based on the syntactic structure. These elements may be the proteins which interest us. We are using the Apple Pie Parser ver.5.9[1], a syntactic parser trained on the Penn Tree Bank (PTB) (recall 77.45%, precision 75.58%).

2 Data Preparation

We restricted test sentences to syntactically well-formed ones, so that we could examine the adequacy of our syntactically-based extraction rules. We assumed that a general-purpose dictionary (GPD) obtained from the PTB would be insufficient for handling biomedical literature. Therefore, we combined on-line glossaries to construct our own terminology dictionary, which we call the Medical Library Dictionary (MLD).

[1] http://www.cs.nyu.edu/cs/projects/proteus/app/

2.1 Test Sentences

We received from a biologist a list of words denoting interactions and 1000 abstracts retrieved from Medline using the PubMed[2]. These abstracts are related to Interleukin-6, a secreted protein whose main function is to mediate inflammatory response in the body. Medline is the bibliographic database of the National Library of Medicine (NLM) in the United States. PubMed is an NLM service which provides access to Medline and additional life science journals.

Out of the word list, we focused on "activate", as this can effectively express the interaction of two elements. We first ran the syntactic parser on the sentences containing the string "activat*[3]", then picked only sentences that contain the verbal "activat*". There were approximately 1000 such sentences. Second, we consulted the sentences annotated by two professional annotators. They marked phrases containing verbal "activat*" and the corresponding agents and recipients. They also evaluated the parsing results related to the phrases. We then selected 100 sentences randomly from the sentences to which both annotators gave the same marking and same evaluation.

To determine the reliability of the annotators' judgment and the difficulty of the task, we calculated the KAPPA coefficient of their responses, and found it to be 0.54 (Hosaka and Umetsu, 2002). This degree of agreement can be interpreted as "moderate" (Carletta, 1997).

2.2 The Medical Library Dictionary

We assumed that biological, chemical, and medical terminology is used in our domain. Therefore, the MLD was compiled from four glossaries in these areas: Biochemical Glossary[4] (BG), Cancernet Dictionary[5] (CD), Medical Chemistry Dictionary[6] (MCD) and Life Science Dictionary[7] (LSD). In addition to the MLD, we used the Medical Subject Headings (MeSH[8]). MeSH is a controlled vocabulary created by the NLM. We used the C chapter (Diseases). The dictionary size is given in Table 1. The number of terms for MeSH represents unique terms, and includes synonyms as well as chemical names:

Dictionary	Source glossary	Number of terms
MLD	BG	723
	CD	2,414
	MCD	122
	LSD	32,405
MeSH	MeSH	300,263

Table 1. Size of terminology dictionaries

The MLD contained 32,698 unique terms and the GPD 88,707 words. We then removed MLD terms which already were listed in the GPD. This removal resulted in a reduced MLD consisting of 25,772 terms (uniMLD). In addition, there were 401 duplicated terms found in both the MeSH and the MLD. In this case, we retained the words in the MLD, so that the number of MeSH terms decreased to 300,263 (uniMeSH). For the experiment, we used the combination of uniMLD and uniMeSH (MLD-M). When we used both GPD and MLD-M, we called this combination MLD+. Table 2 summarizes the dictionary sizes:

Dictionary			Number of terms
MLD+	GPD		88,707
	MLD-M	uniMLD	25,772
		uniMeSH	119,599

Table 2. Size of dictionaries used for experiment

Among the four glossaries, only the LSD had part of speech (POS), since it was a bilingual resource. The MeSH had only nouns. In the other three glossaries, the POS has not been defined. Our parser included out-of-vocabulary handling. We supposed, however, that appropriate POS would raise the performance. Therefore, we assigned POS to these entries semi-automatically.

3 Extraction Rules

We manually defined extraction rules for active and passive sentences. We converted the parsing output into XML format, and then applied the rules. The following example illustrates the procedure. The parser can print the parsing results in several ways, with or without POS. Our extraction rules do specify POS; however, for simplicity, we suppress them in the example below.

[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
[3] "*" indicates any string.
[4] http://www.fhsu.edu/chemistry/twiese/glossary/biochemglossary.htm
[5] http://www.cancer.gov/dictionary/
[6] http://www.chem.qmw.ac.uk/iupac/medchem/
[7] http://lsd.pharm.kyoto-u.ac.jp/index.html
[8] http://www.nlm.nih.gov/mesh/meshhome.html

Input sentence:

```
We find that ACK-2 can be
activated by cell adhesion in a
Cdc42-dependent manner.
```

Syntactic structure in XML:

<S><NPL[9]>We</NPL><VP>find
<SBAR>that<SS[10]>
<NPL>**ACK-2**</NPL>
<VP>**can**<VP>**be**<VP>**activated**
<PP>by<NPL>**cell adhesion**</NPL></PP>
<PP>in
<NPL>a Cdc42-dependent manner</NPL></PP>
</VP></VP></VP></SS></SBAR></VP>.</S>

Extraction steps:

- Find a VP "activat*" as a starting word.
- Extract the highest VP containing "activat*" up to the point where a PP headed by "by" is encountered. → "can be activated"
- Find the nearest NP/NPL to the left of the "activat*" phrase.
- Extract the highest NP/NPL. → "ACK-2"

4 Preliminary Evaluation

We applied our extraction rules to two sets consisting of the parsing outputs from 100 sentences: parsing with the GPD and with the MLD+.

To measure the extraction performance, we prepared a gold standard: a biologist marked phrases containing verbal "activat*" and its corresponding interacting entities. We regarded system extractions as correct if they contained the marked phrases.

The matrix shown in Table 3 defines three combinations of gold standard and system extraction results, A, B, and C:

	Gold Standard	System
A	extracted	extracted
B	extracted	not extracted
C	not extracted	extracted

Table 3. Evaluation matrix

[9] NPL is a specific category for the parser, representing the lowest NP.
[10] SS is a specific category for the parser, representing an S which is not the top S.

We measured our system's recall and precision rates shown in Table 4 as follows:

Recall: A/ (A+B)
Precision: A/ (A+C)

	Recall %		Precision %	
	GPD	MLD+	GPD	MLD+
VP	98.9	97.9	94.9	93.9
Agent	83.3	86.4	80.6	88.4
Recipient	96.6	94.2	87.6	86.2
All	93.0	92.9	91.0	89.6

Table 4. Extraction performance

We found that it is most difficult to extract an Agent. For this task only, use of our MLD+ improved the system's performance. For other phrases, however, the system performed slightly better when the GDP alone was used.

5 Effect of Specialized Terminology

Our 100 sentences contained about 2,500 words. From the MLD-M, 236 terms (uniMeSH 48, uniMLD 188) were identified. That is, specialized terms contributed about 9 percent of all words. If we consider that the uniMLD is about one-third the size of the GPD, as shown in Table 2, the actual hit rate for terms turned out to be rather low.

As shown in Table 4, use of a terminology dictionary does not always raise the extraction performance. We analyzed sentences from which the information was correctly extracted when only the GPD was used but erroneously extracted when the MLD+ was used. There were six sentences with nine such cases. We found the following three reasons for negative effects:

1. A POS was incorrectly assigned for the context (three cases)
2. A term was correctly identified, but a multi-word building failed (two cases)
3. A POS was correctly assigned, but a phrase building failed (four cases)

Some examples follow. In these, the categories were taken from the PTB[11]. On the left is the parsing result with the GPD only, and on the right is that with the MLD+:

[11] NNPX is a specific category of the Apple Pie Parser, representing NNP or NNPS.

In the presence of Tax, both Cdk4 and Cdk6 were activated.

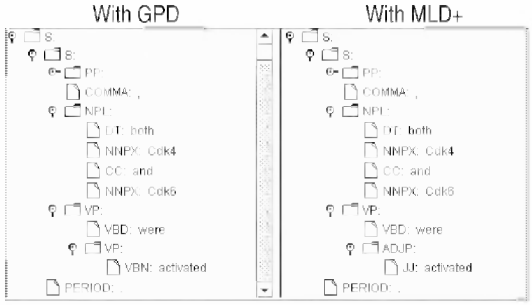

Figure 1. Failure in POS assignment

In, the string "activated", which should be a verb, was assigned falsely as an adjective. In the LSD, "activated" is listed as both POS. This suggests that "activated" is more often used as an adjective in this context in the general domain.

We recently found that ... PI3K was activated in vitro by direct tyrosine phosphorylation.

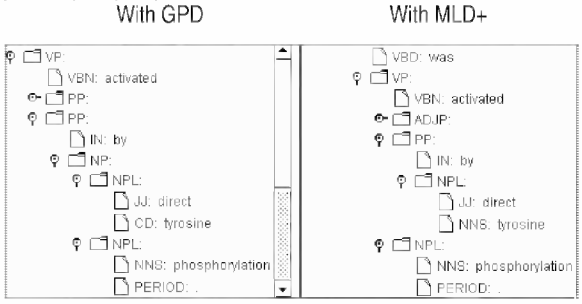

Figure 2. Failure in multi-word building

InFigure 2, the POS of "tyrosine" was correctly assigned. However, the system failed to build a multi-word-term with "phosphorylation".

Further, the appearance of ... suggested that CNF1 activated the Cdc42... .

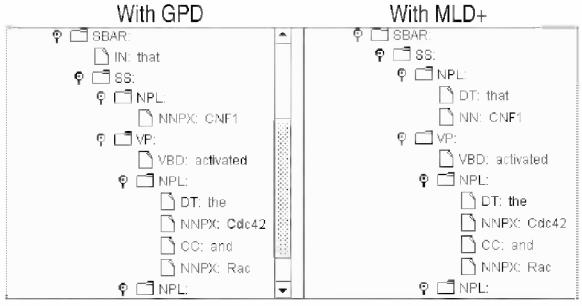

Figure 3. Failure in phrase construction

In Figure 3, "CNF1" got the right POS. However, the preceding "that" is falsely assigned as a determiner. Nouns may often be used with determiners in the general domain.

6 Discussion and Conclusion

In this experiment, information extraction with a general dictionary resulted in slightly better performance than that with specialized dictionary.

Even if a POS is correctly assigned, parsing can fail if the parser is trained on a different domain. To retrain a parser, an annotated corpus is needed, though a construction of such a corpus will be time consuming. In the meantime, we believe the best way is to represent domain-specific structures manually through rules. We observed cases where a term was correctly recognized but the system failed to identify a multi-word-term. To cope with this problem, we will further integrate terminology dictionaries, such as the Unified Medical Language System[12].

We conducted this experiment with a small set of syntactically well-formed sentences. To examine the validity of the result, we are planning further tests with more sentences.

Acknowledgement

We thank Dr. I. Kurochkin for his biomedical advice and Dr. M. Seligman for reading the draft.

References

Blaschke, Christian and Valencia, Alfonso. 2001. The potential use of SUISEKI as a protein interaction discovery tool. *Genome Informatics*, 12: 123-134.

Carletta, Jean, et al. 1997. The reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, 23(1): 13-31.

Friedman, Carol, et.al. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Proc. of ISMB*, 17(Suppl.1): S74-S82.

Hosaka, Junko and Umetsu, Ryo. 2002. Toward the extraction of protein-protein interaction information from immunology literature. *Proc. of IPSJ-SIG-NL*,150: 15-20.

Jenssen, Tor-Kristian, et al. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28: 21-28.

Yakushiji, Akane, et al. 2001. Event extraction from biomedical papers using a full parser. *Proc. of PSB*, 6: 408-419.

[12] http://www.nlm.nih.gov/research/umls/