Finding Anchor Verbs for Biomedical IE Using Predicate-Argument Structures

Akane YAKUSHIJI† **Yuka TATEISI**‡† **Yusuke MIYAO**† **Jun'ichi TSUJII**†‡
†Department of Computer Science, University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033 JAPAN
‡CREST, JST (Japan Science and Technology Agency)
Honcho 4-1-8, Kawaguchi-shi, Saitama 332-0012 JAPAN
{akane,yucca,yusuke,tsujii}@is.s.u-tokyo.ac.jp

Abstract

For biomedical information extraction, most systems use syntactic patterns on verbs (*anchor verbs*) and their arguments. Anchor verbs can be selected by focusing on their arguments. We propose to use predicate-argument structures (PASs), which are outputs of a full parser, to obtain verbs and their arguments. In this paper, we evaluated PAS method by comparing it to a method using part of speech (POSs) pattern matching. POS patterns produced larger results with incorrect arguments, and the results will cause adverse effects on a phase selecting appropriate verbs.

1 Introduction

Research in molecular-biology field is discovering enormous amount of new facts, and thus there is an increasing need for information extraction (IE) technology to support database building and to find novel knowledge in online journals.

To implement IE systems, we need to construct *extraction rules*, i.e., rules to extract desired information from processed resource. One subtask of the construction is defining a set of *anchor verbs*, which express realization of desired information in natural language text.

In this paper, we propose a novel method of finding anchor verbs: extracting anchor verbs from predicate-argument structures (PASs) obtained by full parsing. We here discuss only finding anchor verbs, although our final purpose is construction of extraction rules. Most anchor verbs take topical nouns, i.e., nouns describing target entities for IE, as their arguments. Thus verbs which take topical nouns can be candidates for anchor verbs. Our method collects anchor verb candidates by choosing PASs whose arguments are topical nouns. Then, semantically inappropriate verbs are filtered out. We leave this filtering phase as a future work, and discuss the acquisition of candidates. We have also investigated difference in verbs and their arguments extracted by naive POS patterns and PAS method.

When anchor verbs are found based on whether their arguments are topical nouns, like in (Hatzivassiloglou and Weng, 2002), it is important to obtain correct arguments. Thus, in this paper, we set our goal to obtain anchor verb candidates and their correct arguments.

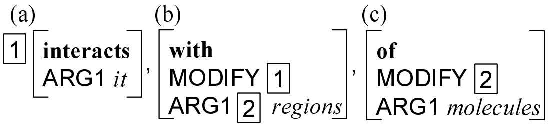
2 Background

There are some works on acquiring extraction rules automatically. Sudo et al. (2003) acquired subtrees derived from dependency trees as extraction rules for IE in general domains. One problem of their system is that dependency trees cannot treat non-local dependencies, and thus rules acquired from the constructions are partial. Hatzivassiloglou and Weng (2002) used frequency of collocation of verbs and topical nouns and verb occurrence rates in several domains to obtain anchor verbs for biological interaction. They used only POSs and word positions to detect relations between verbs and topical nouns. Their performance was 87.5% precision and 82.4% recall. One of the reasons of errors they reported is failures to detect verb-noun relations.

To avoid these problems, we decided to use PASs obtained by full parsing to get precise relations between verbs and their arguments. The obtained precise relations will improve precision. In addition, PASs obtained by full parsing can treat non-local dependencies, thus recall will also be improved.

The sentence below is an example which supports advantage of full parsing. A gerund "**activating**" takes a non-local semantic subject "*IL-4*". In full parsing based on Head-Driven Phrase Structure Grammar (HPSG) (Sag and Wasow, 1999), the subject of the whole sentence and the semantic subject of "**activating**" are shared, and thus we can extract the subject of "**activating**".

> *IL-4* may mediate its biological effects by **activating** *a tyrosine-phosphorylated DNA binding protein.*


(a)
$\boxed{1}$ interacts ARG1 it
(b)
with MODIFY $\boxed{1}$ ARG1 $\boxed{2}$ regions
(c)
of MODIFY $\boxed{2}$ ARG1 molecules

It **interacts with** *non-polymorphic regions* **of** *major histocompatibility complex class II molecules.*

Figure 1: PAS examples

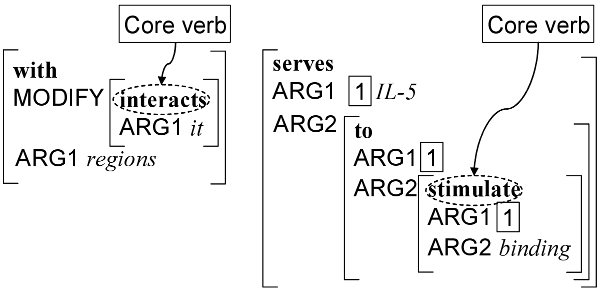


Figure 2: Core verbs of PASs

3 Anchor Verb Finding by PASs

By using PASs, we extract candidates for anchor verbs from a sentence in the following steps:

1. Obtain all PASs of a sentence by a full parser. The PASs correspond not only to verbal phrases but also other phrases such as prepositional phrases.
2. Select PASs which take one or more topical nouns as arguments.
3. From the selected PASs in Step 2, select PASs which include one or more verbs.
4. Extract a *core verb*, which is the innermost verbal predicate, from each of the chosen PASs.

In Step 1, we use a probabilistic HPSG parser developed by Miyao et al. (2003), (2004). PASs obtained by the parser are illustrated in Figure 1.[1] **Bold words** are predicates. Arguments of the predicates are described in ARGn (n $= 1, 2, \ldots$). MODIFY denotes the modified PAS. Numbers in squares denote shared structures. Examples of core verbs are illustrated in Figure 2. We regard all arguments in a PAS are arguments of the core verb.

Extraction of candidates for anchor verbs from the sentence in Figure 1 is as follows. Here, "*regions*" and "*molecules*" are topical nouns.

In Step 1, we obtain all the PASs, (a), (b) and (c), in Figure 1.

Next, in Step 2, we check each argument of (a), (b) and (c). (a) is discarded because it does not have a topical noun argument.[2] (b) is selected because ARG1 "*regions*" is a topical noun. Similarly, (c) is selected because of ARG1 "*molecules*".

And then, in Step 3, we check each POS of a predicate included in (b) and (c). (b) is selected because it has the verb "**interacts**" in $\boxed{1}$ which shares the structure with (a). (c) is discarded because it includes no verbs.

Finally, in Step 4, we extract a core verb from (b). (b) includes $\boxed{1}$ as MODIFY, and the predicate of $\boxed{1}$ is the verb, "**interacts**". So we extract it.

4 Experiments

We investigated the verbs and their arguments extracted by PAS method and POS pattern matching, which is less expressive in analyzing sentence structures but would be more robust.

For topical nouns and POSs, we used the GENIA corpus (Kim et al., 2003), a corpus of annotated abstracts taken from National Library of Medicine's MEDLINE database. We defined topical nouns as the names tagged as protein, peptide, amino acid, DNA, RNA, or nucleic acid. We chose PASs which take one or more topical nouns as an argument or arguments, and substrings matched by POS patterns which include topical nouns. All names tagged in the corpus were replaced by their head nouns in order to reduce complexity of sentences and thus reduce the task of the parser and the POS pattern matcher.

4.1 Implementation of PAS method

We implemented PAS method on LiLFeS, a unification-based programming system for typed feature structures (Makino et al., 1998; Miyao et al., 2000).

The selection in Step 2 described in Section 3 is realized by matching PASs with nine PAS templates. Four of the templates are illustrated in Figure 3.

4.2 POS Pattern Method

We constructed a POS pattern matcher with a partial verb chunking function according to (Hatzivassiloglou and Weng, 2002). Because the original matcher has problems in recall (its verb group detector has low coverage) and precision (it does not consider other words to detect relations between verb groups and topical nouns), we implemented

[1] Here, named entities are regarded as chunked, and thus internal structures of noun phrases are not illustrated.

[2] (a) may be selected if the anaphora ("*it*") is resolved. But we regard anaphora resolving is too hard task as a subprocess of finding anchor verbs.

$$\left[\begin{array}{l}\textbf{*any*}\\ \text{ARG1 } \mathit{N1}\end{array}\right] \quad \mathit{N1} = \text{topical noun}$$

$$\left[\begin{array}{l}\textbf{*any*}\\ \text{MODIFY } \mathit{*any*}\\ \text{ARG1 } \mathit{N1}\end{array}\right] \quad \mathit{N1} = \text{topical noun}$$

$$\left[\begin{array}{l}\textbf{*any*}\\ \text{ARG1 } \mathit{N1}\\ \text{ARG2 } \mathit{N2}\end{array}\right] \quad \begin{array}{l}\mathit{N1} = \text{topical noun}\\ \text{or N2} = \text{topical noun}\end{array}$$

$$\left[\begin{array}{l}\textbf{*any*}\\ \text{MODIFY } \mathit{*any*}\\ \text{ARG1 } \mathit{N1}\\ \text{ARG2 } \mathit{N2}\end{array}\right] \quad \begin{array}{l}\mathit{N1} = \text{topical noun}\\ \text{or N2} = \text{topical noun}\end{array}$$

$$\vdots \qquad\qquad \vdots$$

Figure 3: PAS templates

$N\ \omega\ VG\ \omega\ N$
$N\ \omega\ VG$
$VG\ \omega\ N$
N: is a topical noun
VG: is a verb group which is accepted by a finite state machine described in (Hatzivassiloglou and Weng, 2002) **or one of** {**VB, VBD, VBG, VBN, VBP, VBZ**}
ω: is 0–4 tokens **which do not include** {**FW, NN, NNS, NNP, NNPS, PRP, VBG, WP, ***}
(Parts in **Bold letters** are added to the patterns of Hatzivassiloglou and Weng (2002).)

Figure 4: POS patterns

our POS pattern matcher as a modified version of one in (Hatzivassiloglou and Weng, 2002).

Figure 4 shows patterns in our experiment. The last verb of VG is extracted if all of Ns are topical nouns. Non-topical nouns are disregarded. Adding candidates for verb groups raises recall of obtained relations of verbs and their arguments. Restriction on intervening tokens to non-nouns raises the precision, although it decreases the recall.

4.3 Experiment 1

We extracted last verbs of POS patterns and core verbs of PASs with their arguments from 100 abstracts (976 sentences) of the GENIA corpus. We took up not the verbs only but tuples of the verbs and their arguments (VAs), in order to estimate effect of the arguments on semantical filtering.

Results

The numbers of VAs extracted from the 100 abstracts using POS patterns and PASs are shown in Table 1. (Total − VAs of verbs not extracted by the other method) are not the same, because more than one VA can be extracted on a verb in a sentence. POS patterns method extracted more VAs, although their correctness is not considered.

	POS patterns	PASs
Total	1127	766
VAs of verbs not extracted by the other	478	105

Table 1: Numbers of VAs extracted from the 100 abstracts

	Appropriate	Inappropriate	Total
Correct	43	12	55
Incorrect	20	23	43
Total	63	35	98

Table 2: Numbers of VAs extracted by POS patterns (in detail)

4.4 Experiment 2

For the first 10 abstracts (92 sentences), we manually investigated whether extracted VAs are syntactically or semantically correct. The investigation was based on two criteria: "appropriateness" based on whether the extracted verb can be used for an anchor verb and "correctness" based on whether the syntactical analysis is correct, i.e., whether the arguments were extracted correctly.

Based on human judgment, the verbs that represent interactions, events, and properties were selected as semantically appropriate for anchor verbs, and the others were treated as inappropriate. For example, "**identified**" in "We **identified** *ZEBRA protein*." is not appropriate and discarded.

We did not consider non-topical noun arguments for POS pattern method, whereas we considered them for PAS method. Thus decision on correctness is stricter for PAS method.

Results

The manual investigation results on extracted VAs from the 10 abstracts using POS patterns and PASs are shown in Table 2 and 3 respectively.

POS patterns extracted more (98) VAs than PASs (75), but many of the increment were from incorrect POS pattern matching. By POS patterns, 43 VAs (44%) were extracted based on incorrect analysis. On the other hand, by PASs, 20 VAs (27%) were extracted incorrectly. Thus the ratio of VAs extracted by syntactically correct analysis is larger on PAS method.

POS pattern method extracted 38 VAs of verbs not extracted by PAS method and 7 of them are correct. For PAS method, correspondent numbers are

	Appropriate	Inappropriate	Total
Correct	44	11	55
Incorrect	14	6	20
Total	58	17	75

Table 3: Numbers of VAs extracted by PASs (in detail)

11 and 4 respectively. Thus the increments tend to be caused by incorrect analysis, and the tendency is greater in POS pattern method.

Since not all of verbs that take topical nouns are appropriate for anchor verbs, automatic filtering is required. In the filtering phase that we leave as a future work, we can use semantical classes and frequencies of arguments of the verbs. The results with syntactically incorrect arguments will cause adverse effect on filtering because they express incorrect relationship between verbs and arguments. Since the numbers of extracted VAs after excluding the ones with incorrect arguments are the same (55) between PAS and POS pattern methods, it can be concluded that the precision of PAS method is higher. Although there are few (7) correct VAs which were extracted by POS pattern method but not by PAS method, we expect the number of such verbs can be reduced using a larger corpus.

Examples of appropriate VAs extracted by only one method are as follows: (A) is correct and (B) incorrect, extracted by only POS pattern method, and (C) is correct and (D) incorrect, extracted by only PAS method. **Bold words** are extracted verbs or predicates and *italic words* their extracted arguments.

(A) This delay is associated with down-regulation of many *erythroid cell-specific genes*, **including** *alpha- and beta-globin*, band 3, band 4.1, and . . . .

(B) . . . show that several elements in the . . . region of the *IL-2R alpha gene* **contribute** to *IL-1* responsiveness, ...

(C) The *CD4 coreceptor* interacts with non-polymorphic regions of . . . molecules on non-polymorphic cells and **contributes** to T cell activation.

(D) Whereas *activation* of the HIV-1 enhancer following T-cell stimulation is **mediated** largely through binding of the . . . factor NF-kappa B **to** two adjacent *kappa B sites* in . . . .

5 Conclusions

We have proposed a method of extracting anchor verbs as elements of extraction rules for IE by using PASs obtained by full parsing. To compare our method with more naive and robust methods, we have extracted verbs and their arguments using POS patterns and PASs. POS pattern method could obtain more candidate verbs for anchor verbs, but many of them were extracted with incorrect arguments by incorrect matching. A later filtering process benefits by precise relations between verbs and their arguments which PASs obtained. The shortcoming of PAS method is expected to be reduced by using a larger corpus, because verbs to extract will appear many times in many forms. One of the future works is to extend PAS method to handle events in nominalized forms.

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

Vasileios Hatzivassiloglou and Wubin Weng. 2002. Learning anchor verbs for biological interaction patterns from published text articles. *International Journal of Medical Informatics*, 67:19–32.

Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi, and Jun'ichi Tsujii. 2003. GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl. 1):i180–i182.

Takaki Makino, Minoru Yoshida, Kentaro Torisawa, and Jun-ichi Tsujii. 1998. LiLFeS — towards a practical HPSG parser. In *Proceedings of COLING-ACL'98*.

Yusuke Miyao, Takaki Makino, Kentaro Torisawa, and Jun-ichi Tsujii. 2000. The LiLFeS abstract machine and its evaluation with the LinGO grammar. *Natural Language Engineering*, 6(1):47–61.

Yusuke Miyao, Takashi Ninomiya, and Jun'ichi Tsujii. 2003. Probabilistic modeling of argument structures including non-local dependencies. In *Proceedings of RANLP 2003*, pages 285–291.

Yusuke Miyao, Takashi Ninomiya, and Jun'ichi Tsujii. 2004. Corpus-oriented grammar development for acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank. In *Proceedings of IJCNLP-04*.

Ivan A. Sag and Thomas Wasow. 1999. *Syntactic Theory*. CSLI publications.

Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. 2003. An improved extraction pattern representation model for automatic IE pattern acquisition. In *Proceedings of ACL 2003*, pages 224–231.