An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing

Tomoko Ohta
Yoshimasa Tsuruoka^{*†}
Jumpei Takeuchi
Jin-Dong Kim

Yusuke Miyao
Akane Yakushiji[‡]
Kazuhiro Yoshida
Yuka Tateisi[§]

Takashi Ninomiya[¶]
Katsuya Masuda
Tadayoshi Hara
Jun'ichi Tsujii

Department of Computer Science, University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033 JAPAN

{okap, yusuke, ninomi, tsuruoka, akane, kmasuda, tj-jug,
kyoshida, harasan, jdkim, yucca, tsujii}@is.s.u-tokyo.ac.jp

Abstract

We present a practical HPSG parser for English, an intelligent search engine to retrieve MEDLINE abstracts that represent *biomedical events* and an efficient MEDLINE search tool helping users to find information about biomedical entities such as *genes*, *proteins*, and the *interactions* between them.

1 Introduction

Recently, biomedical researchers have been facing the vast repository of research papers, e.g. MEDLINE. These researchers are eager to search biomedical correlations such as protein-protein or gene-disease associations. The use of natural language processing technology is expected to reduce their burden, and various attempts of information extraction using NLP has been being made (Blaschke and Valencia, 2002; Hao et al., 2005; Chun et al., 2006). However, the framework of traditional information retrieval (IR) has difficulty with the accurate retrieval of such relational concepts. This is because relational concepts are essentially determined by semantic relations of words, and keyword-based IR techniques are insufficient to describe such relations precisely.

This paper proposes a practical HPSG parser for English, **Enju**, an intelligent search engine for the accurate retrieval of relational concepts from

	F-Score	
	GENIA treebank	Penn Treebank
HPSG-PTB	85.10%	87.16%
HPSG-GENIA	86.87%	86.81%

Table 1: Performance for Penn Treebank and the GENIA corpus

MEDLINE, **MEDIE**, and a GUI-based efficient MEDLINE search tool, **Info-PubMed**.

2 Enju: An English HPSG Parser

We developed an English HPSG parser, Enju¹ (Miyao and Tsujii, 2005; Hara et al., 2005; Ninomiya et al., 2005). Table 1 shows the performance. The F-score in the table was accuracy of the predicate-argument relations output by the parser. A predicate-argument relation is defined as a tuple $\langle \sigma, w_h, a, w_a \rangle$, where σ is the predicate type (e.g., adjective, intransitive verb), w_h is the head word of the predicate, a is the argument label (**MOD**, **ARG1**, ..., **ARG4**), and w_a is the head word of the argument. Precision/recall is the ratio of tuples correctly identified by the parser. The lexicon of the grammar was extracted from Sections 02-21 of Penn Treebank (39,832 sentences). In the table, 'HPSG-PTB' means that the statistical model was trained on Penn Treebank. 'HPSG-GENIA' means that the statistical model was trained on both Penn Treebank and GENIA treebank as described in (Hara et al., 2005). The GENIA treebank (Tateisi et al., 2005) consists of 500 abstracts (4,446 sentences) extracted from MEDLINE.

Figure 1 shows a part of the parse tree and fea-

^{*}Current Affiliation:

[†]School of Informatics, University of Manchester

[‡]Knowledge Research Center, Fujitsu Laboratories LTD.

[§]Faculty of Informatics, Kogakuin University

[¶]Information Technology Center, University of Tokyo

¹<http://www-tsujii.is.s.u-tokyo.ac.jp/enju/>

Figure 1: Snapshot of Enju

ture structure for the sentence “NASA officials vowed to land Discovery early Tuesday at one of three locations after weather conditions forced them to scrub Monday’s scheduled return.”

3 MEDIE: a search engine for MEDLINE

Figure 2 shows the top page of the MEDIE. MEDIE is an intelligent search engine for the accurate retrieval of relational concepts from MEDLINE² (Miyao et al., 2006). Prior to retrieval, all sentences are annotated with predicate argument structures and ontological identifiers by applying Enju and a term recognizer.

3.1 Automatically Annotated Corpus

First, we applied a POS analyzer and then Enju. The POS analyzer and HPSG parser are trained by using the GENIA corpus (Tsuruoka et al., 2005; Hara et al., 2005), which comprises around 2,000 MEDLINE abstracts annotated with POS and Penn Treebank style syntactic parse trees (Tateisi et al., 2005). The HPSG parser generates parse trees in a stand-off format that can be converted to XML by combining it with the original text.

We also annotated technical terms of genes and diseases in our developed corpus. Technical terms are annotated simply by exact matching of dictio-

nary entries and the terms separated by space, tab, period, comma, hat, colon, semi-colon, brackets, square brackets and slash in MEDLINE.

The entire dictionary was generated by applying the automatic generation method of name variations (Tsuruoka and Tsujii, 2004) to the GENA dictionary for the gene names (Koike and Takagi, 2004) and the UMLS (Unified Medical Language System) meta-thesaurus for the disease names (Lindberg et al., 1993). It was generated by applying the name-variation generation method, and we obtained 4,467,855 entries of a gene and disease dictionary.

3.2 Functions of MEDIE

MEDIE provides three types of search, **semantic search**, **keyword search**, **GCL search**. GCL search provides us the most fundamental and powerful functions in which users can specify the boolean relations, linear order relation and structural relations with variables. Trained users can enjoy all functions in MEDIE by the GCL search, but it is not easy for general users to write appropriate queries for the parsed corpus. The semantic search enables us to specify an event verb with its subject and object easily. MEDIE automatically generates the GCL query from the semantic query, and runs the GCL search. Figure 3 shows the output of semantic search for the query ‘What disease does dystrophin cause?’. This example will give us the most intuitive understandings of the proximal and structural retrieval with a richly annotated parsed corpus. MEDIE retrieves sentences which include event verbs of ‘cause’ and noun ‘dystrophin’ such that ‘dystrophin’ is the subject of the event verbs. The event verb and its subject and object are highlighted with designated colors. As seen in the figure, small sentences in relative clauses, passive forms or coordination are retrieved. As the objects of the event verbs are highlighted, we can easily see what disease dystrophin caused. As the target corpus is already annotated with diseases entities, MEDIE can efficiently retrieve the disease expressions.

4 Info-PubMed: a GUI-based MEDLINE search tool

Info-PubMed is a MEDLINE search tool with GUI, helping users to find information about biomedical entities such as *genes*, *proteins*, and

²<http://www-tsujii.is.s.u-tokyo.ac.jp/medie/>

Figure 2: Snapshot of MEDIE: top page⁴

Figure 3: Snapshot of MEDIE: 'What disease does dystrophin cause?'

the interactions between them³.

Info-PubMed provides information from MEDLINE on protein-protein interactions. Given the name of a *gene* or *protein*, it shows a list of the names of other *genes/proteins* which co-occur in sentences from MEDLINE, along with the frequency of co-occurrence.

Co-occurrence of two *proteins/genes* in the same sentence does not always imply that they interact. For more accurate extraction of sentences that indicate interactions, it is necessary to identify relations between the two substances. We adopted PASs derived by Enju and constructed extraction patterns on specific verbs and their arguments based on the derived PASs (Yakusiji, 2006).

Figure 4: Snapshot of Info-PubMed (1)

Figure 5: Snapshot of Info-PubMed (2)

Figure 6: Snapshot of Info-PubMed (3)

4.1 Functions of Info-PubMed

In the 'Gene Searcher' window, enter the name of a gene or protein that you are interested in. For example, if you are interested in Raf1, type "raf1" in the 'Gene Searcher' (Figure 4). You will see a list of genes whose description in our dictionary contains "raf1" (Figure 5). Then, drag

³<http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/>

one of the GeneBoxes from the ‘Gene Searcher’ to the ‘Interaction Viewer.’ You will see a list of genes/proteins which co-occur in the same sentences, along with co-occurrence frequency. The GeneBox in the leftmost column is the one you have moved to ‘Interaction Viewer.’ The GeneBoxes in the second column correspond to gene/proteins which co-occur in the same sentences, followed by the boxes in the third column, InteractionBoxes.

Drag an InteractionBox to ‘ContentViewer’ to see the content of the box (Figure 6). An InteractionBox is a set of SentenceBoxes. A SentenceBox corresponds to a sentence in MEDLINE in which the two gene/proteins co-occur. A SentenceBox indicates whether the co-occurrence in the sentence is direct evidence of interaction or not. If it is judged as direct evidence of interaction, it is indicated as Interaction. Otherwise, it is indicated as Co-occurrence.

5 Conclusion

We presented an English HPSG parser, **Enju**, a search engine for relational concepts from MEDLINE, **MEDIE**, and a GUI-based MEDLINE search tool, **Info-PubMed**.

MEDIE and Info-PubMed demonstrate how the results of deep parsing can be used for intelligent text mining and semantic information retrieval in the biomedical domain.

6 Acknowledgment

This work was partially supported by Grant-in-Aid for Scientific Research on Priority Areas “Systems Genomics” (MEXT, Japan) and Solution-Oriented Research for Science and Technology (JST, Japan).

References

C. Blaschke and A. Valencia. 2002. The frame-based module of the SUISEKI information extraction system. *IEEE Intelligent Systems*, 17(2):14–20.

Y. Hao, X. Zhu, M. Huang, and M. Li. 2005. Discovering patterns to extract protein-protein interactions from the literature: Part II. *Bioinformatics*, 21(15):3294–3300.

H.-W. Chun, Y. Tsuruoka, J.-D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. 2006. Extraction of gene-disease relations from MedLine using domain dictionaries and machine learning. In *Proc. PSB 2006*, pages 4–15.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Yusuke Miyao and Jun’ichi Tsujii. 2005. Probabilistic disambiguation models for wide-coverage HPSG parsing. In *Proc. of ACL’05*, pages 83–90.

Tadayoshi Hara, Yusuke Miyao, and Jun’ichi Tsujii. 2005. Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. In *Proc. of IJCNLP 2005*.

Takashi Ninomiya, Yoshimasa Tsuruoka, Yusuke Miyao, and Jun’ichi Tsujii. 2005. Efficacy of beam thresholding, unification filtering and hybrid parsing in probabilistic HPSG parsing. In *Proc. of IWPT 2005*, pages 103–114.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun’ichi Tsujii. 2005. Syntax Annotation for the GENIA corpus. In *Proc. of the IJCNLP 2005, Companion volume*, pp. 222–227.

Yusuke Miyao, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Ninomiya and Jun’ichi Tsujii. 2006. Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In *Proc. of ACL’06*, to appear.

Yoshimasa Tsuruoka, Yuka Tateisi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun’ichi Tsujii. 2005. Part-of-speech tagger for biomedical text. In *Proc. of the 10th Panhellenic Conference on Informatics*.

Y. Tsuruoka and J. Tsujii. 2004. Improving the performance of dictionary-based approaches in protein name recognition. *Journal of Biomedical Informatics*, 37(6):461–470.

Asako Koike and Toshihisa Takagi. 2004. Gene/protein/family name recognition in biomedical literature. In *Proc. of HLT-NAACL 2004 Workshop: Bioblink 2004*, pages 9–16.

D.A. Lindberg, B.L. Humphreys, and A.T. McCray. 1993. The unified medical language system. *Methods in Inf. Med.*, 32(4):281–291.

Akane Yakushiji. 2006. Relation Information Extraction Using Deep Syntactic Analysis. *Ph.D. Thesis*, University of Tokyo.