Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases

Yusuke Miyao[*] **Tomoko Ohta**[*] **Katsuya Masuda**[*] **Yoshimasa Tsuruoka**[†]
Kazuhiro Yoshida[*] **Takashi Ninomiya**[‡] **Jun'ichi Tsujii**[*†]
[*]Department of Computer Science, University of Tokyo
[†]School of Informatics, University of Manchester
[‡]Information Technology Center, University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033 JAPAN
{yusuke,okap,kmasuda,tsuruoka,kyoshida,ninomi,tsujii}@is.s.u-tokyo.ac.jp

Abstract

This paper introduces a novel framework for the accurate retrieval of relational concepts from huge texts. Prior to retrieval, all sentences are annotated with predicate argument structures and ontological identifiers by applying a deep parser and a term recognizer. During the run time, user requests are converted into queries of region algebra on these annotations. Structural matching with pre-computed semantic annotations establishes the accurate and efficient retrieval of relational concepts. This framework was applied to a text retrieval system for MEDLINE. Experiments on the retrieval of biomedical correlations revealed that the cost is sufficiently small for real-time applications and that the retrieval precision is significantly improved.

1 Introduction

Rapid expansion of text information has motivated the development of efficient methods of accessing information in huge texts. Furthermore, user demand has shifted toward the retrieval of more precise and complex information, including *relational concepts*. For example, biomedical researchers deal with a massive quantity of publications; MEDLINE contains approximately 15 million references to journal articles in life sciences, and its size is rapidly increasing, at a rate of more than 10% yearly (National Library of Medicine, 2005). Researchers would like to be able to search this huge textbase for biomedical correlations such as protein-protein or gene-disease associations (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 because relational concepts are essentially determined by semantic relations between words, and keyword-based IR techniques are insufficient to describe such relations precisely.

The present paper demonstrates a framework for the accurate real-time retrieval of relational concepts from huge texts. Prior to retrieval, we prepare a semantically annotated textbase by applying NLP tools including deep parsers and term recognizers. That is, all sentences are annotated in advance with semantic structures and are stored in a structured database. User requests are converted on the fly into patterns of these semantic annotations, and texts are retrieved by matching these patterns with the pre-computed semantic annotations. The accurate retrieval of relational concepts is attained because we can precisely describe relational concepts using semantic annotations. In addition, real-time retrieval is possible because semantic annotations are computed in advance.

This framework has been implemented for a text retrieval system for MEDLINE. We first apply a deep parser (Miyao and Tsujii, 2005) and a dictionary-based term recognizer (Tsuruoka and Tsujii, 2004) to MEDLINE and obtain annotations of predicate argument structures and ontological identifiers of genes, gene products, diseases, and events. We then provide a search engine for these annotated sentences. User requests are converted into queries of region algebra (Clarke et al., 1995) extended with variables (Masuda et al., 2006) on these annotations. A search engine for the extended region algebra efficiently finds sentences having semantic annotations that match the input queries. In this paper, we evaluate this system with respect to the retrieval of biomedical correlations

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 1017–1024,
Sydney, July 2006. ©2006 Association for Computational Linguistics

Symbol	CRP
Name	C-reactive protein, pentraxin-related
Species	Homo sapiens
Synonym	MGC88244, PTX1
Product	C-reactive protein precursor, C-reactive protein, pentraxin-related protein
External links	EntrezGene:1401, GDB:119071, ...

Table 1: An example GENA entry

and examine the effects of using predicate argument structures and ontological identifiers.

The need for the discovery of relational concepts has been investigated intensively in Information Extraction (IE). However, little research has targeted on-demand retrieval from huge texts. One difficulty is that IE techniques such as pattern matching and machine learning require heavier processing in order to be applied on the fly. Another difficulty is that target information must be formalized beforehand and each system is designed for a specific task. For instance, an IE system for protein-protein interactions is not useful for finding gene-disease associations. Apart from IE research, enrichment of texts with various annotations has been proposed and is becoming a new research area for information management (IBM, 2005; TEI, 2004). The present study basically examines this new direction in research. The significant contribution of the present paper, however, is to provide the first empirical results of this framework for a real task with a huge textbase.

2 Background: Resources and Tools for Semantic Annotations

The proposed system for the retrieval of relational concepts is a product of recent developments in NLP resources and tools. In this section, ontology databases, deep parsers, and search algorithms for structured data are introduced.

2.1 Ontology databases

Ontology databases are collections of words and phrases in specific domains. Such databases have been constructed extensively for the systematic management of domain knowledge by organizing textual expressions of ontological entities that are detached from actual sentences.

For example, GENA (Koike and Takagi, 2004) is a database of genes and gene products that is semi-automatically collected from well-known databases, including HUGO, OMIM, Genatlas, Locuslink, GDB, MGI, FlyBase, WormBase, CYGD, and SGD. Table 1 shows an example of a GENA entry. "Symbol" and "Name" denote short forms and nomenclatures of genes, respectively. "Species" represents the organism species in which this gene is observed. "Synonym" is a list of synonyms and name variations. "Product" gives a list of products of this gene, such as proteins coded by this gene. "External links" provides links to other databases, and helps to obtain detailed information from these databases. For biomedical terms other than genes/gene products, the Unified Medical Language System (UMLS) meta-thesaurus (Lindberg et al., 1993) is a large database that contains various names of biomedical and health-related concepts.

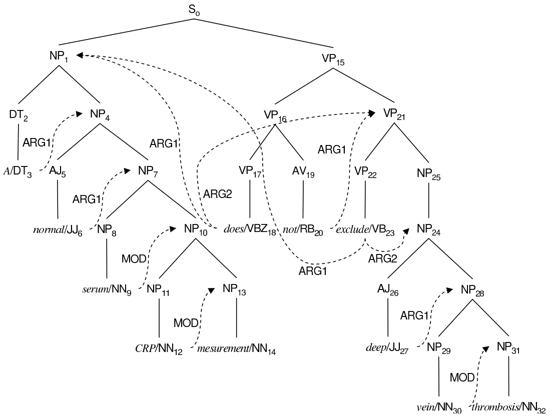

Figure 1: An output of HPSG parsing

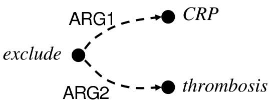

Figure 2: A predicate argument structure

Ontology databases provide mappings between textual expressions and entities in the real world. For example, Table 1 indicates that CRP, MGC88244, and PTX1 denote the same gene conceptually. Hence, these resources enable us to canonicalize variations of textual expressions of ontological entities.

2.2 Parsing technologies

Recently, state-of-the-art CFG parsers (Charniak and Johnson, 2005) can compute phrase structures of natural sentences at fairly high accuracy. These parsers have been used in various NLP tasks including IE and text mining. In addition, parsers that compute deeper analyses, such as predicate argument structures, have become available for the processing of real-world sentences (Miyao and Tsujii, 2005). Predicate argument structures are canonicalized representations of sentence meanings, and express the semantic relations of words explicitly. Figure 1 shows an output of an HPSG parser (Miyao and Tsujii, 2005) for the sentence "*A normal serum CRP measurement does not exclude deep vein thrombosis.*" The dotted lines express predicate argument relations. For example, the ARG1 arrow coming from "*exclude*" points to the noun phrase "*A normal serum CRP measurement*", which indicates that the subject of "*exclude*" is this noun phrase, while such relations are not explicitly represented by phrase structures.

Predicate argument structures are beneficial for our purpose because they can represent relational concepts in an abstract manner. For example, the relational concept of "*CRP excludes thrombosis*" can be represented as a predicate argument structure, as shown in Figure 2. This structure is universal in various syntactic expressions, such as passivization (e.g., "*thrombosis is excluded by CRP*") and relativization (e.g., "*thrombosis that CRP excludes*"). Hence, we can abstract surface variations of sentences and describe relational concepts in a canonicalized form.

2.3 Structural search algorithms

Search algorithms for structured texts have been studied extensively, and examples include XML databases with XPath (Clark and DeRose, 1999) and XQuery (Boag et al., 2005), and region algebra (Clarke et al., 1995). The present study focuses on region algebra extended with variables (Masuda et al., 2006) because it provides an efficient search algorithm for tags with cross boundaries. When we annotate texts with various levels of syntactic/semantic structures, cross boundaries are inherently nonnegligible. In fact, as described in Section 3, our system exploits annotations of predicate argument structures and ontological entities, which include substantial cross boundaries.

Region algebra is defined as a set of operators on *regions*, i.e., word sequences. Table 2 shows operators of the extended region algebra, where A and B denote regions, and results of operations are also regions. For example, "A & B" denotes a region that includes both A and B. Four containment operators, >, >>, <, and <<, represent ancestor/descendant relations in XML. For example, "A > B" indicates that A is an ancestor of B. In search algorithms for region algebra, the cost of retrieving the first answer is constant, and that of an exhaustive search is bounded by the lowest frequency of a word in a query (Clarke et al., 1995).

`[tag]`	Region covered with "`<tag>`"
A > B	A containing B
A >> B	A containing B (A is not nested)
A < B	A contained by B
A << B	A contained by B (B is not nested)
A - B	Starting with A and ending with B
A & B	A and B
A | B	A or B

Table 2: Operators of the extended region algebra

```
[sentence] >>
(([word arg1="$subject"] > exclude) &
 ([phrase id="$subject"] > CRP))
```

Figure 3: A query of the extended region algebra

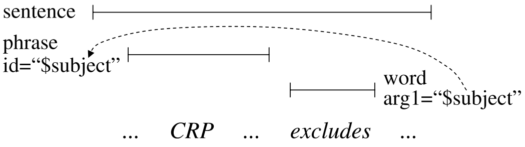

Figure 4: Matching with the query in Figure 3

Variables in the extended region algebra allow us to express shared structures and are necessary in order to describe predicate argument structures. For example, Figure 3 shows a formula in the extended region algebra that represents the predicate argument structure of "*CRP excludes something.*" This formula indicates that a sentence contains a region in which the word "*exclude*" exists, the first argument ("arg1") phrase of which includes the word "*CRP.*" A predicate argument relation is expressed by the variable, "$subject." Figure 4 shows a situation in which this formula is satisfied. Three horizontal bars describe regions covered by `<sentence>`, `<phrase>`, and `<word>` tags, respectively. The dotted line denotes the relation expressed by this variable. Given this formula as a query, a search engine can retrieve sentences having semantic annotations that satisfy this formula.

3 A Text Retrieval System for MEDLINE

While the above resources and tools have been developed independently, their collaboration opens up a new framework for the retrieval of relational concepts, as described below (Figure 5).

Off-line processing: Prior to retrieval, a deep parser is applied to compute predicate argument

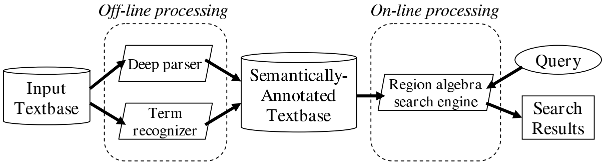

Figure 5: Framework of semantic retrieval

structures, and a term recognizer is applied to create mappings from textual expressions into identifiers in ontology databases. Semantic annotations are stored and indexed in a structured database for the extended region algebra.

On-line processing: User input is converted into queries of the extended region algebra. A search engine retrieves sentences having semantic annotations that match the queries.

This framework is applied to a text retrieval engine for MEDLINE. MEDLINE is an exhaustive database covering nearly 4,500 journals in the life sciences and includes the bibliographies of articles, about half of which have abstracts. Research on IE and text mining in biomedical science has focused mainly on MEDLINE. In the present paper, we target all articles indexed in MEDLINE at the end of 2004 (14,785,094 articles). The following sections explain in detail off-/on-line processing for the text retrieval system for MEDLINE.

3.1 Off-line processing: HPSG parsing and term recognition

We first parsed all sentences using an HPSG parser (Miyao and Tsujii, 2005) to obtain their predicate argument structures. Because our target is biomedical texts, we re-trained a parser (Hara et al., 2005) with the GENIA treebank (Tateisi et al., 2005), and also applied a bidirectional part-of-speech tagger (Tsuruoka and Tsujii, 2005) trained with the GENIA treebank as a preprocessor.

Because parsing speed is still unrealistic for parsing the entire MEDLINE on a single machine, we used two geographically separated computer clusters having 170 nodes (340 Xeon CPUs). These clusters are separately administered and not dedicated for use in the present study. In order to effectively use such an environment, GXP (Taura, 2004) was used to connect these clusters and distribute the load among them. Our processes were given the lowest priority so that our task would not disturb other users. We finished parsing the entire MEDLINE in nine days (Ninomiya et al., 2006).

# entries (genes)	517,773
# entries (gene products)	171,711
# entries (diseases)	148,602
# expanded entries	4,467,855

Table 3: Sizes of ontologies used for term recognition

Event type	Expressions
influence	effect, affect, role, response, . . .
regulation	mediate, regulate, regulation, . . .
activation	induce, activate, activation, . . .

Table 4: Event expression ontology

Next, we annotated technical terms, such as genes and diseases, to create mappings to ontological identifiers. A dictionary-based term recognition algorithm (Tsuruoka and Tsujii, 2004) was applied for this task. First, an expanded term list was created by generating name variations of terms in GENA and the UMLS meta-thesaurus[1]. Table 3 shows the size of the original database and the number of entries expanded by name variations. Terms in MEDLINE were then identified by the longest matching of entries in this expanded list with words/phrases in MEDLINE.

The necessity of ontologies is not limited to nominal expressions. Various verbs are used for expressing events. For example, activation events of proteins can be expressed by “activate,” “enhance,” and other event expressions. Although the numbers of verbs and their event types are much smaller than those of technical terms, verbal expressions are important for the description of relational concepts. Since ontologies of event expressions in this domain have not yet been constructed, we developed an ontology from scratch. We investigated 500 abstracts extracted from MEDLINE, and classified 167 frequent expressions, including verbs and their nominalized forms, into 18 event types. Table 4 shows a part of this ontology. These expressions in MEDLINE were automatically annotated with event types.

As a result, we obtained semantically annotated MEDLINE. Table 5 shows the size of the original MEDLINE and semantic annotations. Figure 6 shows semantic annotations for the sentence in Figure 1, where “-” indicates nodes of XML,[2]

[1] We collected disease names by specifying a query with the semantic type as “Disease or Syndrome.”

[2] Although this example is shown in XML, this textbase contains tags with cross boundaries because tags for predicate argument structures and technical terms may overlap.

# papers	14,785,094
# abstracts	7,291,857
# sentences	70,935,630
# words	1,462,626,934
# successfully parsed sentences	69,243,788
# predicate argument relations	1,510,233,701
# phrase tags	3,094,105,383
# terms (genes)	84,998,621
# terms (gene products)	27,471,488
# terms (diseases)	19,150,984
# terms (event expressions)	51,810,047
Size of the original MEDLINE	9.3 GByte
Size of the semantic annotations	292 GByte
Size of the index file for region algebra	954 GByte

Table 5: Sizes of the original and semantically annotated MEDLINE textbases

```
- <sentence sentence_id="e6e525">
 - <phrase id="0" cat="S" head="15" lex_head="18">
  - <phrase id="1" cat="NP" head="4" lex_head="14">
   - <phrase id="2" cat="DT" head="3" lex_head="3">
    - <word id="3" pos="DT" cat="DT" base="a" arg1="4">
     - A
   - <phrase id="4" cat="NP" head="7" lex_head="14">
    - <phrase id="5" cat="AJ" head="6" lex_head="6">
     - <word id="6" pos="JJ" cat="AJ" base="normal" arg1="7">
      - normal
    - <phrase id="7" cat="NP" head="10" lex_head="14">
     - <phrase id="8" cat="NP" head="9" lex_head="9">
      - <word id="9" pos="NN" cat="NP" base="serum" mod="10">
       - serum
     - <phrase id="10" cat="NP" head="13" lex_head="14">
      - <phrase id="11" cat="NP" head="12" lex_head="12">
       - <entity_name id="entity-1" type="gene"
          gene_id="GHS003134" gene_symbol="CRP"
          gene_name="C-reactive protein, pentraxin-related"
          species="Homo sapiens"
          db_site="EntrezGene:1401|GDB:119071|GenAtlas:CRP">
        - <word id="12" pos="NN" cat="NP" base="crp" mod="13">
         - CRP
      - <phrase id="13" cat="NP" head="14" lex_head="14">
       - <word id="14" pos="NN" cat="NP" base="measurement">
        - measurement
  - <phrase id="15" cat="VP" head="16" lex_head="18">
   - <phrase id="16" cat="VP" head="17" lex_head="18">
    - <phrase id="17" cat="VP" head="18" lex_head="18">
     - <word id="18" pos="VBZ" cat="VP" base="do"
        arg1="1" arg2="21">
      - does
    - <phrase id="19" cat="AV" head="20" lex_head="20">
     - <word id="20" pos="RB" cat="AV" base="not" arg1="21">
      - not
   - <phrase id="21" cat="VP" head="22" lex_head="23">
    - <phrase id="22" cat="VP" head="23" lex_head="23">
     - <word id="23" pos="VB" cat="VP" base="exclude"
        arg1="1" arg2="24">
      - exclude
...
```

Figure 6: A semantically annotated sentence

although the latter half of the sentence is omitted because of space limitations. Sentences are annotated with four tags,[3] "phrase," "word," "sentence," and "entity_name," and their attributes as given in Table 6. Predicate argument structures are annotated as attributes, "mod" and "arg$X$," which point to the IDs of the argument phrases. For example, in Figure 6, the `<word>` tag for "*exclude*" has the attributes `arg1="1"` and `arg2="24"`, which denote the IDs of the subject and object phrases, respectively.

Tag	Attributes
phrase	id, cat, head, lex_head
word	id, cat, pos, base, mod, arg$X$, rel_type
sentence	sentence_id
entity_name	id, type, gene_id/disease_id, gene_symbol, gene_name, species, db_site

Attribute	Description
id	unique identifier
cat	syntactic category
head	head daughter's ID
lex_head	lexical head's ID
pos	part-of-speech
base	base form of the word
mod	ID of modifying phrase
arg$X$	ID of the X-th argument of the word
rel_type	event type
sentence_id	sentence's ID
type	whether gene, gene_prod, or disease
gene_id	ID in GENA
disease_id	ID in the UMLS meta-thesaurus
gene_symbol	short form of the gene
gene_name	nomenclature of the gene
species	species that have this gene
db_site	links to external databases

Table 6: Tags (upper) and attributes (lower) for semantic annotations

3.2 On-line processing

The off-line processing described above results in much simpler on-line processing. User input is converted into queries of the extended region algebra, and the converted queries are entered into a search engine for the extended region algebra. The implementation of a search engine is described in detail in Masuda et al. (2006).

Basically, given subject x, object y, and verb v, the system creates the following query:

```
[sentence] >>
  ([word arg1="$subject" arg2="$object"
         base="v"] &
   ([phrase id="$subject"] > x) &
   ([phrase id="$object"] > y))
```

Ontological identifiers are substituted for x, y, and v, if possible. Nominal keywords, i.e., x and y, are replaced by `[entity_name gene_id="n"]` or `[entity_name disease_id="n"]`, where n is the ontological identifier of x or y. For verbal keywords, `base="v"` is replaced by `rel_type="r"`, where r is the event type of v.

4 Evaluation

Our system is evaluated with respect to speed and accuracy. Speed is indispensable for real-time interactive text retrieval systems, and accuracy is key for the motivation of semantic retrieval. That is, our motivation for employing semantic retrieval

[3] Additional tags exist for representing document structures such as "title" (details omitted).

Query No.	User input
1	*something* inhibit ERK2
2	*something* trigger diabetes
3	adiponectin increase *something*
4	TNF activate IL6
5	dystrophin cause *disease*
6	macrophage induce *something*
7	*something* suppress MAP phosphorylation
8	*something* enhance p53 (negative)

Table 7: Queries for experiments

```
[sentence] >>
  ([word rel_type="activation"] &
   [entity_name type="gene" gene_id="GHS019685"] &
   [entity_name type="gene" gene_id="GHS009426"])

[sentence] >>
  ([word arg1="$subject" arg2="$object"
         rel_type="activation"] &
   ([phrase id="$subject"] >
     [entity_name type="gene" gene_id="GHS019685"]) &
   ([phrase cat="np" id="$object"] >
     [entity_name type="gene" gene_id="GHS009426"]))
```

Figure 7: Queries of the extended region algebra for Query 4-3 (upper: keyword search, lower: semantic search)

was to provide a device for the accurate identification of relational concepts. In particular, high precision is desired in text retrieval from huge texts because users want to extract relevant information, rather than collect exhaustive information.

We have two parameters to vary: whether to use predicate argument structures and whether to use ontological identifiers. The effect of using predicate argument structures is evaluated by comparing "keyword search" with "semantic search." The former is a traditional style of IR, in which sentences are retrieved by matching words in a query with words in sentences. The latter is a new feature of the present system, in which sentences are retrieved by matching predicate argument relations in a query with those in a semantically annotated textbase. The effect of using ontological identifiers is assessed by changing queries of the extended region algebra. When we use the term ontology, nominal keywords in queries are replaced with ontological identifiers in GENA and the UMLS meta-thesaurus. When we use the event expression ontology, verbal keywords in queries are replaced with event types.

Table 7 is a list of queries used in the following experiments. Words in italics indicate a class of words: "*something*" indicates that any word can appear, and *disease* indicates that any disease expression can appear. These queries were selected by a biologist, and express typical relational concepts that a biologist may wish to find. Queries 1, 3, and 4 represent relations of genes/proteins, where ERK2, adiponectin, TNF, and IL6 are genes/proteins. Queries 2 and 5 describe relations concerning diseases, and Query 6 is a query that is not relevant to genes or diseases. Query 7 expresses a complex relation concerning a specific phenomena, i.e., phosphorylation, of MAP. Query 8 describes a relation concerning a gene, i.e., p53, while "(negative)" indicates that the target of retrieval is negative mentions. This is expressed by "not" modifying a predicate.

For example, Query 4 attempts to retrieve sentences that mention the protein-protein interaction "*TNF* activates *IL6*." This is converted into queries of the extended region algebra given in Figure 7. The upper query is for keyword search and only specifies the appearances of the three words. Note that the keywords are translated into the ontological identifiers, "activation," "GHS019685," and "GHS009426." The lower query is for semantic search. The variables in "arg1" and "arg2" indicate that "GHS019685" and "GHS009426" are the subject and object, respectively, of "activation".

Table 8 summarizes the results of the experiments. The postfixes of query numbers denote whether ontological identifiers are used. X-1 used no ontologies, and X-2 used only the term ontology. X-3 used both the term and event expression ontologies[4]. Comparison of X-1 and X-2 clarifies the effect of using the term ontology. Comparison of X-2 and X-3 shows the effect of the event expression ontology. The results for X-3 indicate the maximum performance of the current system. This table shows that the time required for the semantic search for the first answer, shown as "time (first)" in seconds, was reasonably short. Thus, the present framework is acceptable for real-time text retrieval. The numbers of answers increased when we used the ontologies, and this result indicates the efficacy of both ontologies for obtaining relational concepts written in various expressions.

Accuracy was measured by judgment by a biologist. At most 100 sentences were retrieved for each query, and the results of keyword search and semantic search were merged and shuffled. A biologist judged the shuffled sentences (1,839 sentences in total) without knowing whether the sen-

[4] Query 5-1 is not tested because "*disease*" requires the term ontology, and Query 6-2 is not tested because "macrophage" is not assigned an ontological identifier.

Query	Keyword search				Semantic search			
No.	# ans.	time (first/all)	precision	n-precision	# ans.	time (first/all)	precision	relative recall
1-1	252	0.00/ 1.5	74/100 (74%)	74/100 (74%)	143	0.01/ 2.5	96/100 (96%)	51/74 (69%)
1-2	348	0.00/ 1.9	61/100 (61%)	61/100 (61%)	174	0.01/ 3.1	89/100 (89%)	42/61 (69%)
1-3	884	0.00/ 3.2	50/100 (50%)	50/100 (50%)	292	0.01/ 5.3	91/100 (91%)	21/50 (42%)
2-1	125	0.00/ 1.8	45/100 (45%)	9/ 27 (33%)	27	0.02/ 2.9	23/ 27 (85%)	17/45 (38%)
2-2	113	0.00/ 2.9	40/100 (40%)	10/ 26 (38%)	26	0.06/ 4.0	22/ 26 (85%)	19/40 (48%)
2-3	6529	0.00/ 12.1	42/100 (42%)	42/100 (42%)	662	0.01/1527.4	76/100 (76%)	8/42 (19%)
3-1	287	0.00/ 1.5	20/100 (20%)	4/ 30 (13%)	30	0.05/ 2.4	23/ 30 (80%)	6/20 (30%)
3-2	309	0.01/ 2.1	21/100 (21%)	4/ 32 (13%)	32	0.10/ 3.5	26/ 32 (81%)	6/21 (29%)
3-3	338	0.01/ 2.2	24/100 (24%)	8/ 39 (21%)	39	0.05/ 3.6	32/ 39 (82%)	8/24 (33%)
4-1	4	0.26/ 1.5	0/ 4 (0%)	0/ 0 (—)	0	2.44/ 2.4	0/ 0 (—)	0/ 0 (—)
4-2	195	0.01/ 2.5	9/100 (9%)	1/ 6 (17%)	6	0.09/ 4.1	5/ 6 (83%)	2/ 9 (22%)
4-3	2063	0.00/ 7.5	5/100 (5%)	5/ 94 (5%)	94	0.02/ 10.5	89/ 94 (95%)	2/ 5 (40%)
5-2	287	0.08/ 6.3	73/100 (73%)	73/100 (73%)	116	0.05/ 14.7	97/100 (97%)	37/73 (51%)
5-3	602	0.01/ 15.9	50/100 (50%)	50/100 (50%)	122	0.05/ 14.2	96/100 (96%)	23/50 (46%)
6-1	10698	0.00/ 42.8	14/100 (14%)	14/100 (14%)	1559	0.01/3014.5	65/100 (65%)	10/14 (71%)
6-3	42106	0.00/3379.5	11/100 (11%)	11/100 (11%)	2776	0.01/5100.1	61/100 (61%)	5/11 (45%)
7	87	0.04/ 2.7	34/ 87 (39%)	7/ 15 (47%)	15	0.05/ 4.2	10/ 15 (67%)	10/34 (29%)
8	1812	0.01/ 7.6	19/100 (19%)	17/ 84 (20%)	84	0.20/ 29.2	73/ 84 (87%)	7/19 (37%)

Table 8: Number of retrieved sentences, retrieval time, and accuracy

tence was retrieved by keyword search or semantic search. Without considering which words actually matched the query, a sentence is judged to be correct when any part of the sentence expresses all of the relations described by the query. The modality of sentences was not distinguished, except in the case of Query 8. These evaluation criteria may be disadvantageous for the semantic search because its ability to exactly recognize the participants of relational concepts is not evaluated. Table 8 shows the precision attained by keyword/semantic search and n-precision, which denotes the precision of the keyword search, in which the same number, n, of outputs is taken as the semantic search. The table also gives the *relative recall* of the semantic search, which represents the ratio of sentences that are correctly output by the semantic search among those correctly output by the keyword search. This does not necessarily represent the true recall because sentences not output by keyword search are excluded. However, this is sufficient for the comparison of keyword search and semantic search.

The results show that the semantic search exhibited impressive improvements in precision. The precision was over 80% for most queries and was nearly 100% for Queries 4 and 5. This indicates that predicate argument structures are effective for representing relational concepts precisely, especially for relations in which two entities are involved. Relative recall was approximately 30–50%, except for Query 2. In the following, we will investigate the reasons for the residual errors.

Disregarding of noun phrase structures	45
Term recognition errors	33
Parsing errors	11
Other reasons	8
Incorrect human judgment	7
Nominal expressions	41
Phrasal verb expressions	26
Inference required	24
Coreference resolution required	19
Parsing errors	16
Other reasons	15
Incorrect human judgment	10

Table 9: Error analysis (upper: 104 false positives, lower: 151 false negatives)

Table 9 shows the classifications of the errors of semantic retrieval. The major reason for false positives was that our queries ignore internal structures of noun phrases. The system therefore retrieved noun phrases that do not directly mention target entities. For example, "*the increased mortality in patients with diabetes was caused by . . .*" does not indicate the trigger of diabetes. Another reason was term recognition errors. For example, the system falsely retrieved sentences containing "*p40*," which is sometimes, but not necessarily used as a synonym for "*ERK2*." Machine learning-based term disambiguation will alleviate these errors. False negatives were caused mainly by nominal expressions such as "*the inhibition of ERK2*." This is because the present system does not convert user input into queries on nominal expressions. Another major reason, phrasal verb expressions such as "*lead to*," is also a shortage of our current strategy of query creation. Because semantic annotations already include linguistic structures of these expressions, the present system can be improved further by creating queries on such expressions.

5 Conclusion

We demonstrated a text retrieval system for MEDLINE that exploits pre-computed semantic annotations[5]. Experimental results revealed that the proposed system is sufficiently efficient for real-time text retrieval and that the precision of retrieval was remarkably high. Analysis of residual errors showed that the handling of noun phrase structures and the improvement of term recognition will increase retrieval accuracy. Although the present paper focused on MEDLINE, the NLP tools used in this system are domain/task independent. This framework will thus be applicable to other domains such as patent documents.

The present framework does not conflict with conventional IR/IE techniques, and integration with these techniques is expected to improve the accuracy and usability of the proposed system. For example, query expansion and relevancy feedback can be integrated in a straightforward way in order to improve accuracy. Document ranking is useful for the readability of retrieved results. IE systems can be applied off-line, in the manner of the deep parser in our system, for annotating sentences with target information of IE. Such annotations will enable us to retrieve higher-level concepts, such as relationships among relational concepts.

Acknowledgment

This work was partially supported by Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" (MEXT, Japan), Genome Network Project (NIG, 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.

S. Boag, D. Chamberlin, M. F. Fernández, D. Florescu, J. Robie, and J. Siméon. 2005. XQuery 1.0: An XML query language.

E. Charniak and M. Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proc. ACL 2005*.

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.

J. Clark and S. DeRose. 1999. XML Path Language (XPath) version 1.0.

C. L. A. Clarke, G. V. Cormack, and F. J. Burkowski. 1995. An algebra for structured text search and a framework for its implementation. *The Computer Journal*, 38(1):43–56.

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.

T. Hara, Y. Miyao, and J. Tsujii. 2005. Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. In *Proc. IJCNLP 2005*.

IBM, 2005. *Unstructed Information Management Architecture (UIMA) SDK User's Guide and Reference*.

A. Koike and T. Takagi. 2004. Gene/protein/family name recognition in biomedical literature. In *Proc. Biolink 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.

K. Masuda, T. Ninomiya, Y. Miyao, T. Ohta, and J. Tsujii. 2006. Nested region algebra extended with variables. In Preparation.

Y. Miyao and J. Tsujii. 2005. Probabilistic disambiguation models for wide-coverage HPSG parsing. In *Proc. 43rd ACL*, pages 83–90.

National Library of Medicine. 2005. Fact Sheet MEDLINE. Available at `http://www.nlm.nih.gov/pubs/factsheets/medline.html`.

T. Ninomiya, Y. Tsuruoka, Y. Miyao, K. Taura, and J. Tsujii. 2006. Fast and scalable HPSG parsing. *Traitement automatique des langues (TAL)*, 46(2).

Y. Tateisi, A. Yakushiji, T. Ohta, and J. Tsujii. 2005. Syntax annotation for the GENIA corpus. In *Proc. IJCNLP 2005, Companion volume*, pages 222–227.

K. Taura. 2004. GXP : An interactive shell for the grid environment. In *Proc. IWIA2004*, pages 59–67.

TEI Consortium, 2004. *Text Encoding Initiative*.

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.

Y. Tsuruoka and J. Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proc. HLT/EMNLP 2005*, pages 467–474.

[5] A web-based demo of our system is available on-line at: `http://www-tsujii.is.s.u-tokyo.ac.jp/medie/`