

A Methodology for Terminology-based Knowledge Acquisition and Integration

Hideki Mima^{1†}, Sophia Ananiadou², Goran Nenadic² and Junichi Tsujii¹

¹Dept. of Information Science, University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan
{mima, tsujii}@is.s.u-tokyo.ac.jp

²Computer Science, University of Salford
Newton Building, Salford M5 4WT, UK
{S.Ananiadou, G.Nenadic}@salford.ac.uk

Abstract

In this paper we propose an integrated knowledge management system in which terminology-based knowledge acquisition, knowledge integration, and XML-based knowledge retrieval are combined using tag information and ontology management tools. The main objective of the system is to facilitate knowledge acquisition through query answering against XML-based documents in the domain of molecular biology. Our system integrates automatic term recognition, term variation management, context-based automatic term clustering, ontology-based inference, and intelligent tag information retrieval. Tag-based retrieval is implemented through interval operations, which prove to be a powerful means for textual mining and knowledge acquisition. The aim is to provide efficient access to heterogeneous biological textual data and databases, enabling users to integrate a wide range of textual and non-textual resources effortlessly.

Introduction

With the recent increasing importance of electronic communication and data sharing over the Internet, there exist an increasingly growing number of publicly accessible knowledge sources, both in the form of documents and factual databases. These knowledge sources (KSs) are intrinsically heterogeneous and dynamic. They are *heterogeneous* since they are autonomously developed and maintained by independent organizations for different purposes. They are *dynamic* since constantly new information is being revised, added and removed. Such an heterogeneous and dynamic nature of KSs imposes challenges on systems that help users to locate and integrate knowledge relevant to their needs.

Knowledge, encoded in textual documents, is organised around sets of specialised (technical) *terms* (e.g. names of proteins, genes, acids). Therefore, knowledge acquisition relies heavily on the recognition of terms. However, the main problems that make term recognition difficult are the lack of clear naming conventions and

terminology variation (cf. Jacquemin and Tzoukermann (1999)), especially in the domain of molecular biology. Therefore, we need a scheme to integrate terminology management as a key prerequisite for knowledge acquisition and integration.

However, automatic term extraction is not the ultimate goal itself, since the large number of new terms calls for a systematic way to access and retrieve the knowledge represented through them. Therefore, the extracted terms need to be placed in an appropriate framework by discovering relations between them, and by establishing the links between the terms and different factual databases.

In order to solve the problem, several approaches have been proposed. MeSH Term in MEDLINE (2002) and Gene Ontology (2002) provide a top-down controlled ontology framework, which aims to describe and constrain the terminology in the domain of molecular biology. On the other hand, automatic term acquisition approaches have been developed in order to address a dynamic and corpus-driven knowledge acquisition methodology (Mima et al., 1999; 2001a).

[†] Current affiliation: Dept. of Engineering, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan

Different approaches to linking relevant resources have also been suggested. The Semantic Web framework (Berners-Lee (1998)) aims to link relevant Web resources in bottom-up manner using the Resource Description Framework (RDF) (Bricklet and Guha, 2000) and an ontology. However, although the Semantic Web framework is powerful to express content of resources to be semantically retrieved, some manual description is expected using the RDF/ontology. Since no solution to the well-known difficulties in manual ontology development, such as the ontology conflicts/mismatches (Visser et al., 1997) is provided, an automated ontology management is required for the efficient and consistent knowledge acquisition and integration. TAMBIS (Baker et al., 1998) tried to provide a filter from biological information services by building a homogenising layer on top of the different sources using the classical mediator/wrapper architecture. It intended to provide source transparency using a mapping from terms placed in a conceptual knowledge base of molecular biology onto terms in external sources.

In this paper we introduce TIMS, an integrated knowledge management system in the domain of molecular biology, where terminology-based knowledge acquisition (KA), knowledge integration (KI), and XML-based knowledge retrieval are combined using tag information and ontology management tools. The management of knowledge resources, similarly to the Semantic Web, is based on XML, RDF, and ontology-based inference. However, our aim is to facilitate the KA and KI tasks not only by using manually defined resource descriptions, but also by exploiting NLP techniques such as automatic term recognition (ATR) and automatic term clustering (ATC), which are used for automatic and systematic ontology population.

The paper is organised as follows: in section 1 we present the overall TIMS architecture and briefly describe the components incorporated in the system, while section 2 gives the details of the proposed method for KA and KI. In the last section we present results, evaluation and discussion.

1 TIMS – system architecture

XML-based Tag Information Management System (TIMS) is a core machinery for managing XML tag information obtained from sub functional components. Its main aim is to facilitate an efficient mechanism for KA and KI through a query answering system for XML-based documents in the domain of molecular biology, by using a tag information database.

Figure 1 shows the system architecture of TIMS. It integrates the following modules via XML-based data exchange: JTAG — an annotation tool, ATRACT — an automatic term recognition and clustering workbench, and the LiLFes abstract machine, which we briefly describe in this section. ATRACT and LiLFes play a central role in the knowledge acquisition process, which includes term recognition, ontology population, and ontology-based

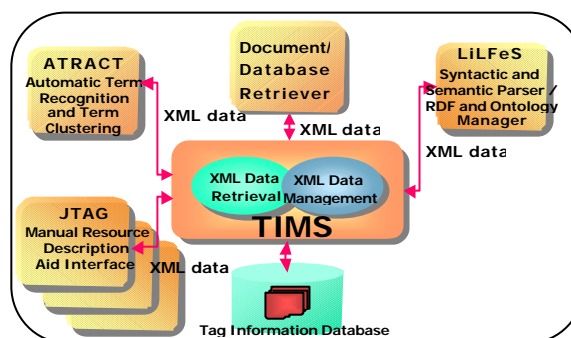


Figure 1: System architecture of TIMS

inference. In addition to these modules, TIMS implements an XML-data manager and a TIQL query processor (see Section 2).

1.1 JTAG

JTAG is an XML-based manual annotation and resource description aid tool. Its purpose is to support manual annotation (e.g. semantic tagging), adjusting term recognition results, developing RDF logic, etc. In addition, ontology information described in XML can also be developed and modified using the tool. All the annotations can be managed via a GUI.

1.2 ATRACT

In the domain of molecular biology, there is an increasing amount of new terms that represent newly created concepts. Since existing term

dictionaries cannot cover the needs of specialists, automatic term extraction tools are important for consistent term discovery. ATRACT (Mima et al., 2001a) is a terminology management workbench that integrates ATR and ATC. Its main aim is to help biologists to gather and manage terminology in the domain. The module retrieves and classifies terms on the fly and sends the results as XML tag information to TIMS.

The ATR method is based on the *C/NC-value* method (Frantzi et al., 2000). The original method has been augmented with acronym acquisition and term variation management (Nenadic et al. 2002), in order to link different terms that denote the same concept. Term variation management is based on term normalisation as an integral part of the ATR process. All orthographic, morphological and syntactic term variations and acronym variants (if any) are conflated prior to the statistical analysis, so that term candidates comprise all variants that appear in a corpus.

Besides term recognition, term clustering is an indispensable component in a knowledge management process (see figure 2). Since terminological opacity and polysemy are very common in molecular biology, term clustering is essential for the semantic integration of terms, the construction of domain ontology and for choosing the appropriate semantic information.

The ATC method is based on Ushioda's AMI (Average Mutual Information)-hierarchical clustering method (Ushioda, 1996). Our implementation uses parallel symmetric processing for high speed clustering and is built on the *C/NC-value* results. As input, we use co-occurrences of automatically recognised terms and their contexts, and the output is a dendrogram of hierarchical term clusters (like a thesaurus). The calculated term cluster information is stored in LiLFes (see below) and combined with a predefined ontology according to the term classes automatically assigned.

1.3 LiLFes

LiLFes (Miyao et al., 2000) is a Prolog-like programming language and language processor used for defining definite clause programs with typed feature structures. Since typed feature structures can be used like first order terms in Prolog, the LiLFes language can describe various kinds of applications based on feature

structures. Examples include HPSG parsers, HPSG-based grammars and compilers from HPSG to CFG. Furthermore, other NLP modules can be easily developed because feature structure processing can be directly written in the LiLFes language. Within TIMS, LiLFes is used to: 1) infer similarity between terms using hierarchical matching, and 2) parse sentences using HPSG-based parsers and convert the results into an XML-based formalism.

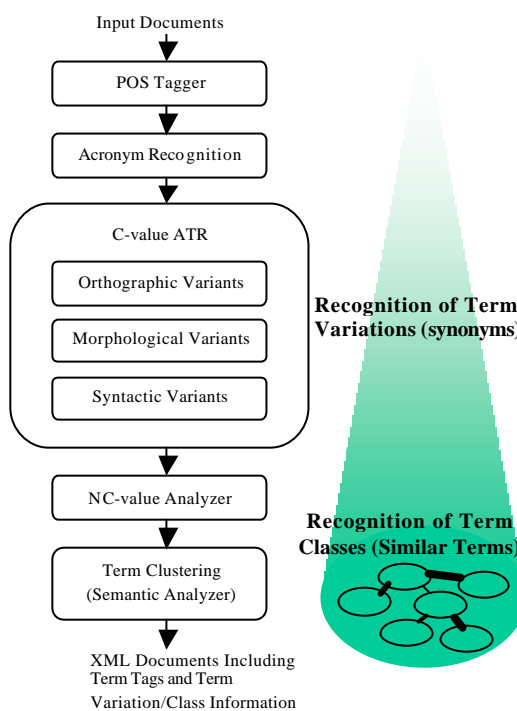


Figure 2. Term Ontology Development

2 Knowledge Integration and Management

Knowledge integration and management in TIMS is organised by integrating XML-data management (section 2.1) and tag- and ontology-based information extraction (section 2.2). Figure 3 illustrates a model of the knowledge management based on the knowledge integration and question-answering process within TIMS. In this scenario, a user formulates a query, which is processed by a query manager. The tag data manager retrieves the relevant data from the collection of documents via a tag database and ontology-based inference (such as hierarchical matching of term classes).

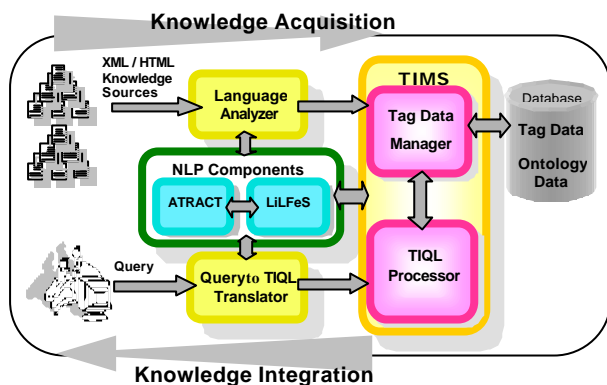


Figure 3: Question-answering process in TIMS

2.1 XML-tag data management

Communication within TIMS is based on XML-data exchange. TIMS initially parses the XML documents (which contain relevant terminology information generated automatically by ATTRACT) and “de-tags” them. Then, like in the TIPSTER architecture (Grishman, 1995), every tag information is stored separately from the original documents and managed by an external database software. This facility allows, as shown in figure 4, different types of tags (POS, syntactic, semantic, etc.) for the same document to be supported.

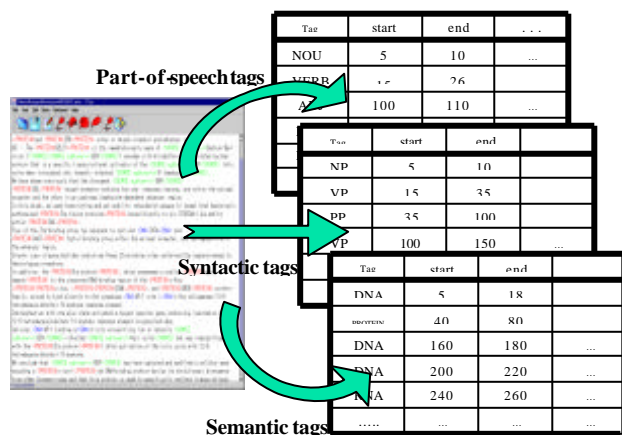


Figure 4: Tag data management

2.2 Tag- and ontology-based IE

The key feature of KA and KI within TIMS is a facility to logically retrieve data that is represented by different tags. This feature is implemented via *interval operations*. The main

assumption is that the XML tags specify certain *intervals* within documents. Interval operations are XML specific text/data retrieval operations, which operate on such textual intervals. Each interval operation takes two sets of intervals as input and returns a set of intervals according to the specified logical operations. Currently, we define four types of logical operations:

- *Intersection* ‘ \otimes ’ returns intersected intervals of all the intervals given.
- *Union* ‘ \oplus ’ returns merged intervals of all the intersected intervals.
- *Subtraction* ‘ \ominus ’ returns differences in intervals of all the intersected intervals.
- *Concatenation* ‘+’ returns concatenated intervals of all the continuous intervals.

For example, the interval operation ¹ $\langle VP \rangle \otimes (\langle V \rangle \cup \langle term \rangle)$ describes all verb ($\langle V \rangle$)-term ($\langle term \rangle$) pairs within a verb phrase ($\langle VP \rangle$). Similarly, suppose X denotes a set of intervals of manually annotated tags for a document and Y denotes a set of intervals of automatically annotated tags for the same document. The interval operation $((X \otimes Y) \oplus \{X \cup Y\})$ results in the differences between human and machine annotations (see figure 5).

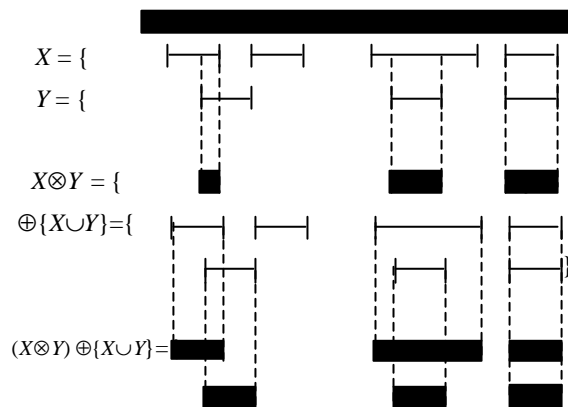


Figure 5. $(X \otimes Y) \oplus \{X \cup Y\}$

Interval operations are powerful means for textual mining from different sources using tag information. In addition, LiLFes enables tag (interval) retrieval to process not only regular pattern/string matching using tag information, but also the ontological hierarchy matching to

¹ ‘ \cup ’ denotes a merged set of all the elements.

subordinate classes using either predefined or automatically derived term ontology. Thus, *semantically*-based tag information retrieval can be achieved. For example, the interval operation² $\langle VP \rangle \otimes \langle nucleic_acid^* \rangle$ will retrieve all subordinate terms/classes of *nucleic acid*, which are contained within a VP.

The interval operations can be performed over the specified documents and/or tag sets (e.g. syntactic, semantic tags, etc.) simultaneously or in batch mode, by selecting the documents/tag sets from a list. This accelerates the process of KA, as users are able to retrieve information from multiple KSs simultaneously.

2.3 TIQL - Tag Information Query Language

In order to integrate and expand the above components, we have developed a tag information query language (TIQL). Using this language, a user can specify the interval operations to be performed on selected documents (including the ontology inference to expand queries). The basic expression in TIQL has the following form:

```

SELECT [n-tuple variables]
FROM [XML document(s)]
WHERE [interval operation]
FROM [XML document(s)]
WHERE [interval operation]
.....

```

where, [n-tuple variables] specifies the table output format, [XML document(s)] denotes the document(s) to be processed, and [interval operation] denotes an interval operation to be performed over the corresponding document with variables of each interval to be bound.

For example, the following expression:

```

SELECT x1, x2
FROM "paper-1.xml"
WHERE
  <VP>⊗{x1:<EVENT*>∪x2:<nucleic_acid*>}
FROM "paper-2.xml"
WHERE
  <VP>⊗{x1:<EVENT*>∪x2:<nucleic_acid*>}

```

extracts all the hierarchically subordinate classes matched to ($\langle EVENT \rangle$, $\langle nucleic_acid \rangle$) pair

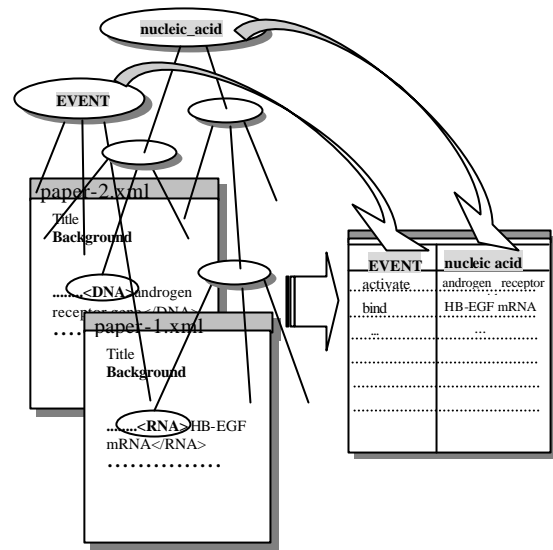


Figure 6. Ontology-based Tagged Information Retrieval

within a VP from the specified XML-documents, and then automatically builds a table to display the results (see figure 6).

Since formulating an appropriate TIQL expression using interval operations might be cumbersome, in particular for novice users, TIMS was augmented with a capability of “recycling” predefined queries and macros.

3 Evaluation and discussion

We have conducted preliminary experiments using the proposed framework. In this paper we briefly present the quality of automatic term recognition and similarity measure calculation via automatically clustered terms. After that, we discuss the practical performance of tag manipulation in TIMS compared to string-based XML tag manipulation to show the advantage of the tag information management scheme.

The term recognition evaluation was performed on the NACSIS AI-domain corpus (Koyama et al., 1998), which includes 1800 abstracts and on a set of MEDLINE abstracts. Table 1 shows a sample of extracted terms and term variants. The ATR precisions of the top 100 intervals range from 93% to 98% (see figure 7; for detailed evaluation, see Mima et al. (2001b) and Nenadic et al. (2002)).

² ‘*’ denotes hierarchical matching.

terms (and term variants)	term-hood
<u>retinoic acid receptor</u>	
retinoic acid receptor	6.33
retinoic acid receptors	
RAR, RARs	
<u>nuclear receptor</u>	
nuclear receptor	6.00
nuclear receptors	
NR, NRs	
<u>all-trans retinoic acid</u>	
all trans retinoic acid	4.75
all-trans-retinoic acids	
ATRA, at-RA, atRA	
<u>9-cis-retinoic acid</u>	
9-cis retinoic acid	4.25
9cRA, 9-C-RA	

Table 1: Sample of recognised terms

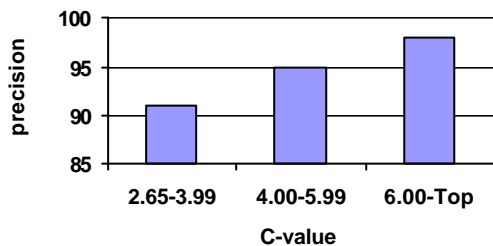


Figure 7: ATR interval precision

For term clustering and tag manipulation performance we used the GENIA resources (GENIA corpus, 2002), which include 1,000 MEDLINE abstracts (MEDLINE, 2002), with overall 40,000 (16,000 distinct) semantic tags annotated for terms in the domain of nuclear receptors. We used the similarity measure calculation as the central computing mechanism for inferring the relevance between the XML tags and tags specified in the TIQL/interval operation, determining the most relevant tags in the XML-based KS(s). As a gold standard, we used similarities between the terms that were calculated according to the hierarchy of the clustered terms according to the GENIA ontology. In this experiment, we have adopted a semantic similarity calculation method for measuring the similarity between terms described in (Oi et al., 1997). The three major sets of classes (namely, *nucleic_acid*, *amino_acid*, *SOURCE*) of manually classified terms from GENIA ontology (GENIA corpus, 2002) were

used to calculate the average similarities (AS) of the elements. ASs of the elements within the same classes were greater than the ASs between elements from different classes, which proves that the terms were clustered reliably according to their semantic features.

In order to examine the tag manipulation performance of TIMS, we measured the processing times consumed for executing an interval operation in TIMS compared to the time needed by using string-based regular expression matching (REM). We focused on measuring the interval operation ‘ \otimes ’ with intervals (tags) $\langle title \rangle$ and $\langle term \rangle$ (i.e. extracting all terms within titles). In the evaluation process, we used 5 different samples to examine IE performances according to their size (namely the number of tags and file size in Kb).

	Sample1	Sample2	Sample3	Sample4	Sample5
TIMS (millisec.)	16	28	40	44	62
REM (millisec.)	24	38	58	80	104
# of tags	1146	2383	3730	4799	5876
Size (K bytes)	92	191	298	382	470

Table 2: TIMS - practical performance

Table 2 and Figure 8 show the results: the processing times of TIMS were about 1.4-1.8 times faster (depending on number of tags and corpus length) than those of REM. Therefore, we assume that the TIMS tag information management scheme can be considered as an efficient mechanism to facilitate knowledge acquisition and information extraction process.

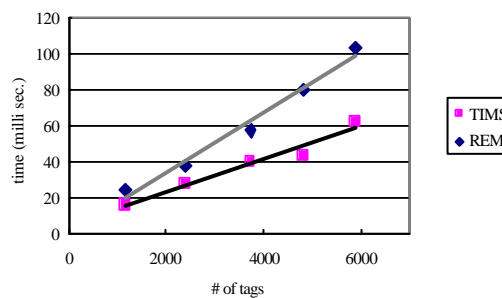


Figure 8. IE performance (TIMS vs. REM)

Conclusion

In this paper, we presented a methodology for KA and KI over large KSs. We described TIMS, an XML-based integrated KA aid system, in which we have integrated automatic term recognition, term clustering, tagged data management and ontology-based knowledge retrieval. TIMS allows users to search and combine information from various sources. An important source of information in the system is derived from terminological knowledge, which is provided automatically in the XML format. Tag-based retrieval is implemented through interval operations, which – in combination with hierarchical matching – prove to be powerful means for textual mining and knowledge acquisition.

The system has been tested in the domain of molecular biology. The preliminary experiments show that the TIMS tag information management scheme is an efficient methodology to facilitate KA and IE in specialised fields.

Important areas of future research will involve expanding the scalability of the system to real WWW knowledge acquisition tasks and experiments with fine-grained term classification.

References

- Baker P. G., Brass A., Bechhofer S., Goble C., Paton N. and Stevens R. (1998) *TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources*. An Overview in Proc. of the Sixth International Conference on Intelligent Systems for Molecular Biology, ISMB98, Montreal.
- Berners-Lee, T. (1998) *The Semantic Web as a language of logic*, available at: <http://www.w3.org/DesignIssues/Logic.html>
- Brickle, D. and Guha R. (2000) *Resource Description Framework (RDF) Schema Specification 1.0*, W3C Candidate Recommendation, available at <http://www.w3.org/TR/rdf-schema>
- Frantzi K. T., Ananiadou S. and Mima H. (2000) *Automatic Recognition of Multi-Word Terms: the C-value/NC-value method*, in International Journal on Digital Libraries, Vol. 3, No. 2, 115–130.
- Gene Ontology Consortium (2002) *GO ontology*. available at <http://www.geneontology.org/>
- GENIA corpus (2002) *GENIA project home page*. <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>
- Grishman R (1995) *TIPSTER Phase II Architecture Design Document*. New York University, available at <http://www.tipster.org/arch.htm>
- Jacquemin C. and Tzoukermann E. (1999) *NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax*. In T. Strzalkowski (editor), Natural Language Information Retrieval, Kluwer, Boston, pp. 25-74.
- Koyama T., Yoshioka M. and Kageura K. (1998) *The Construction of a Lexically Motivated Corpus - The Problem with Defining Lexical Unit*. In Proceedings of LREC 1998, Granada, Spain, pp. 1015–1019.
- MEDLINE (2002) National Library of Medicine, <http://www.ncbi.nlm.nih.gov/PubMed/>
- Mima H., Ananiadou S. and Nenadic G. (2001a) *TRACT Workbench: An Automatic Term Recognition and Clustering of Terms*, in Text, Speech and Dialogue - TSD2001, Lecture Notes in AI 2166, Springer Verlag
- Mima H. and Ananiadou S. (2001b) *An Application and Evaluation of the C/NC-value Approach for the Automatic term Recognition of Multi-Word units in Japanese*, in International Journal on Terminology, Vol. 6(2), pp 175-194.
- Mima H., Ananiadou S. and Tsujii J. (1999) *A Web-based integrated knowledge mining aid system using term-oriented natural language processing*, in Proceedings of The 5th Natural Language Processing Pacific Rim Symposium, NLPRS'99, pp. 13–18.
- Miyao Y., Makino T., Torisawa K. and Tsujii J. (2000) *The LiLFeS abstract machine and its evaluation with the LinGO grammar*. Journal of Natural Language Engineering, Cambridge University Press, Vol. 6(1), pp.47-62.
- Nenadic G., Spasic I. and Ananiadou S. (2002) *Automatic Acronym Acquisition and Term Variation Management within Domain Specific Texts*, in Proc. of LREC 2002, Las Palmas, Spain, pp. 2155-2162.
- Oi K., Sumita E. and Iida H. (1997) *Document Retrieval Method Using Semantic Similarity and Word Sense Disambiguation* (in Japanese), in Journal of Natural Language Processing, Vol.4, No.3, pp.51-70.
- Visser P.R.S., Jones D.M., Bench-Capon T.J.M. and Shave M.J.R. (1997) *An Analysis of Ontology Mismatches; Heterogeneity versus Interoperability*. In AAAI 1997 Spring Symposium on Ontological Engineering, Stanford University, California, USA.
- Ushioda A. (1996) *Hierarchical Clustering of Words*. In Proc. of COLING '96, Copenhagen