A Methodology for Ontology-based Knowledge Acquisition and Structuring in an Industry-Academic-Government Project "Go Japan!"

Hideki Mima* Taesung Yoon**
*School of Engineering, University of Tokyo
**Open Knowledge Corp.

Abstract

The purpose of the study is to develop an integrated knowledge structuring system for the domain of engineering, in which ontology-based literature mining, knowledge acquisition, knowledge integration, and knowledge retrieval are combined using XML-based tag information and ontology management. The system supports combining different types of databases (papers and patents, technologies and innovations) and retrieving different types of knowledge simultaneously. The main objective of the system is to facilitate knowledge acquisition and knowledge retrieval from documents through an ontology-based dynamic similarity calculation and a visualization of automatically structured knowledge. Through experimentations we conducted using 100,000 words economic documents reported in the "Go! Japan" project for analyzing Japanese industrial situation, and 100,000 words molecular biology papers, we show the system is practical enough for accelerating knowledge acquisition and knowledge discovery from the information sea.

Key Words: Knowledge structuring, knowledge acquisition, information extraction, natural language processing, automatic term recognition, ontology, go Japan

1. Introduction

New scientific discoveries result in an abundance of documents, such as scientific papers and patents, verbalising these discoveries. These documents are created in an attempt to share new knowledge with other scientists. They are often reproduced in electronic form and placed on the Internet or other types of shared resources in order to make the new information widely and easily available. Electronically available texts are continually being created and updated, and, thus, the knowledge represented in such texts is more up-to-date than in any other knowledge media.

The sheer amount of published papers [1] makes it difficult for a human to efficiently localise the information of interest not only in a collection of documents, but also within a single document. The growing number of electronically available knowledge sources (KSs) emphasises the importance of developing flexible and efficient tools for automatic knowledge acquisition and structuring in terms of knowledge integration. Different text and literature mining techniques have been developed recently in order to facilitate efficient discovery of knowledge contained in large textual collections. The main goal of literature mining is to retrieve knowledge that is "buried" in a text and to present the distilled knowledge to users in a concise form. Its advantage, compared to "manual" knowledge discovery, is based on the assumption that automatic methods are able to process an enormous amount of texts. It is doubtful that any researcher could process such huge amount of information, especially if the knowledge spans across domains. For these reasons, literature mining aims at helping scientists in collecting, maintaining, interpreting and curating information.

One of the main problems when processing a collection of KSs is their heterogeneity and dynamic nature. Even when confined to a single domain, the KSs are autonomously developed and maintained by independent organisations for different purposes, hence resulting in a *heterogeneous* set of KSs. Moreover, this set is *dynamic* as a result of continuous attempts to synchronise its content with up-to-date knowledge. New information is being added and existing information is revised and often removed from the KSs. These two facts, heterogeneity and constant evolution of KSs, set a challenge to systems designed to assist users in locating and integrating knowledge relevant to their needs.

In this paper we introduce an integrated knowledge structuring (KS) system, in which ontology-based literature mining, terminology-driven knowledge acquisition (KA), knowledge integration (KI), and knowledge retrieval (KR) are combined using tag-based information management and ontology inference. The system incorporates an ontology development / management and a visualization of retrieved knowledge based on the ontology, which allow users to access KSs visually though sophisticated GUIs.

2. Related Work

2.1. Terminology management

Knowledge encoded in textual documents is organised around sets of specialised *terms* (e.g. in biomedical domain, terms represent names of

[1] For example, the MEDLINE database [1] currently contains over 12 million abstracts in the domains of molecular biology, biomedicine and medicine, growing by more than 40.000 abstracts each month.

proteins, genes, acids, etc.). Hence, KA relies heavily on the recognition of terms. Obviously, a scheme to integrate terminology management as a key prerequisite for KA and KI is needed.

There are several approaches to automatic term recognition (ATR), especially, in recent biomedicine and molecular biology domain. Some of them rely mainly on linguistic information, namely on morpho-syntactic features of domain terms. For instance, LaSIE [2], an adapted newswire name recogniser, uses a case-sensitive terminology lexicon of component terms, set of morphological cues (biochemical suffixes) and hand-constructed grammar rules in order to recognise terms belonging to specific terminological classes (e.g. enzymes, proteins, etc.). Another example of a rule-based system is PROPER [3], which uses "core" and "feature" terms to identify strings that correspond to proteins. "Core" terms are domain-characteristic words (containing capitals, numerals etc.) and "feature" terms are keywords that describe function and characteristic of a term (e.g. protein, receptor, etc.). Recently, hybrid approaches combining linguistic and statistical knowledge are increasingly used ([4, 5]). In order to assess the relevance of extracted term candidates, such methods calculate weights (i.e. termhoods) according to specific statistical measures. Machine learning techniques can be applied as well: for example, [6] presents a statistically based, unsupervised technique to acquire and disambiguate names of proteins, genes, and RNSs.

However, ATR is not the ultimate goal itself. The large number of new terms calls for a systematic way of accessing and retrieving the knowledge represented by them. Accordingly, the extracted terms need to be placed in an appropriate knowledge framework by discovering relations between them, and by establishing links between the terms and different factual databases.

In order to implement terminology-based knowledge structuring, several ontologies have been developed (e.g. MeSH terms, Gene Ontology, GENIA ontology, etc.). Each of them provides a top-down controlled framework, which aims to organise and describe the terminology in the domain. Ontologies implement a pre-defined classification system for terms and their relationships, as well as inference rules that are used to derive knowledge represented by them. However, ontology construction and maintenance are time-consuming activities, as terms are usually manually integrated into an ontology. This is one of the reasons why ontologies typically contain just a subset of existing terminology. In addition, no solution to the well-known difficulties in manual ontology development, such as ontology conflictions / mismatches [7], is provided. Therefore, techniques for automated ontology management [8] are required for efficient and consistent KA and KI.

2.2. Integration of knowledge sources

Different approaches to linking, integrating and interpreting relevant resources have also been suggested. For example, the Semantic Web framework [9] strives to link relevant XML-based resources in a bottom-up manner using the Resource Description Framework (RDF) and ontology information. Since XML allows introduction of new domain- and/or application-specific tags, RDF [10] is used to define their "meanings" and relationships to one another, while the corresponding ontology is used to combine and derive additional information (e.g. synonyms, hyponyms, etc.). In this sense, ontologies are used as a key domain knowledge repository. However, though the Semantic Web framework is powerful when it comes to expressing the content of resources to be semantically retrieved, manual description is needed when defining RDF descriptions and ontologies. If we, however, endeavour to process huge collections of new documents (which cover new knowledge), we need systems that do not rely solely on manual descriptions.

In this paper, we present our approach to terminology management and structuring / integration of knowledge sources adopted in the KS system.

3. An overview of the system

The KS system has been developed with the intention to address the problems of the ontology-driven literature mining and KA. Similarly to the Semantic Web framework, our system deals with XML documents by using domain-specific RDF descriptions and ontology-based inference. However, it facilitates KA tasks not only by using manually defined resource descriptions, but also by exploiting natural language processing techniques such as ATR and automatic term clustering (ATC), which are used for automatic population of the underlying ontology. Additionally, the system integrates an information retrieval engine and a similarity calculation engine that allow users to show not only relevant KSs to keywords but also relevance between KSs.

The system acts as an information extraction engine, which is based on managing XML tag information obtained from its subfunctional components. Typically, IE-based KA process within the system has the following course: first, a collection of documents is linguistically processed (part-of-speech (POS) tagging, shallow parsing, etc.). Further, the collection is terminologically analysed, i.e. relevant domain-specific terms are automatically recognised and structured (classified or incorporated into an ontology).

The system architecture is modular, and it integrates the following components (Figure 1):

- *Ontology Development Engine(s) (ODE)* – components that carry out the automatic ontology development which includes recognition and structuring of domain terminology;

- *Tag Data Manager (TDM)* – stores index of KSs and tag information in a tag information database (TID) and provides the corresponding interface;
- *Knowledge Retriever (KR)* – retrieves KSs from TID and calculates similarities between keywords and KSs. Currently, we adopt tf*idf based similarity calculation;
- *Similarity Calculation Engine(s) (SCE)* – calculate similarities between KSs provided from KR component in order to show semantic similarities between each KSs.
- *Graph Visualizer* – visualizes knowledge structures based on graph expression in which relevance links between provided keywords and KSs, and relevance links between the KSs themselves can be shown.

Linguistic pre-processing within the sysyem is performed in two steps. In the first step, POS tagging[2], i.e. the assignment of basic parts of speech (e.g. noun, verb, etc.) to words, is performed. In the second step, an ontology development engine is used to perform parsing, i.e. the recognition of basic syntactic structures (e.g. noun phrases). The parser is based on the LFG for English and Japanese, which is implemented as an unification based GLR parser with feature structures.

4. Terminological processing as an ontology development

The lack of clear naming standards in a domain (e.g. biomedicine) makes ATR a non-trivial problem [3]. Also, it typically gives rise to many-to-many relationships between terms and concepts. In practice, two problems stem from this fact: the same term may denote a number of concepts, and, conversely, the same concept may be denoted by more than one term. In other words, there are terms that have multiple meanings (*term ambiguity*), and, conversely, there are terms that refer to the same concept (*term variation*). Generally, term ambiguity has negative effects on IE precision, while term variation decreases IE recall.

These problems point out the impropriety of using simple keyword-based IE techniques. Obviously, more sophisticated techniques are needed. Such techniques should identify groups of different terms referring to the same (or similar) concept(s), and, therefore, could benefit from relying on efficient and consistent ATR/ATC and term variation management methods. These methods are also important for organising domain specific knowledge, as terms should not be treated isolated from other terms. They should rather be related to one another so that the relations existing between the corresponding concepts are at least partly reflected in a terminology.

Terminological processing in our system is carried out based on C / NC-value method [5] for ATR, and average mutual information based ATC (Figure 2). Its main purpose is to help domain experts in gathering and managing domain-specific terminology. It is used to automatically retrieve and cluster terms offline / on-the-fly and pass the XML-tagged results.

4.1. Term recognition

The ATR method used in the system is based on the C- and NC-value methods [4]. The *C-value* method recognises terms by combining linguistic knowledge and statistical analysis. The method extracts multi-word terms[3] and is not limited to a specific class of concepts. It is implemented as a two-step procedure. In the first step, term candidates are extracted by using a set of linguistic filters, implemented using a LFG-based GLR parser, which describe general term formation patterns. In the second step, the term candidates are assigned termhoods (referred to as C-values) according to a statistical measure. The measure amalgamates four numerical corpus-based characteristic of a candidate term, namely the frequency of occurrence, the frequency of occurrence as a substring of other candidate

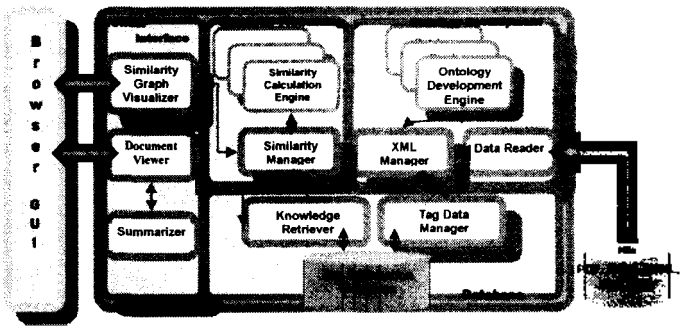

Figure 1: The system architecture

terms, the number of candidate terms containing the given candidate term as a substring, and the number of words contained in the candidate term.

The *NC-value method* further improves the C-value results by taking into account the context of candidate terms. The relevant context words are extracted and assigned weights based on how frequently they appear with top-ranked term candidates extracted by the C-value method. Subsequently, context factors are assigned to candidate terms according to their co-occurrence with top-ranked context words. Finally, new termhood estimations, referred to as NC-values, are calculated as a linear combination of the C-values and context factors for the respective terms. Evaluation of the C/NC-methods (see Section 6) has shown that

[2] We use EngCG tagger[12] in English and JUMAN / Chasen morphological analyzers in Japanese.

[3] More than 85% of domain-specific terms are multi-word terms [5].

contextual information improves term distribution in the extracted list by placing real terms closer to the top of the list.

4.2. Term variation management

Term variation and ambiguity are causing problems not only for ATR but for human experts as well. Several methods for term variation management have been developed. For example, the BLAST system [13] used approximate text string matching techniques and dictionaries to recognise spelling variations in gene and protein names. FASTR [14] handles morphological and syntactic variations by means of meta-rules used to describe term normalisation, while semantic variants are handled via WordNet.

The basic C-value method has been enhanced by term variation management [3]. We consider a variety of sources from which term variation problems originate. In particular, we deal with orthographical, morphological, syntactic, lexico-semantic and pragmatic phenomena. Our approach to term variation management is based on term normalisation as an integral part of the ATR process. Term variants (i.e. synonymous terms) are dealt with ir the initial phase of ATR when term candidates are singled out, as opposed to other approaches (e.g. FASTR handles variants subsequently by applying transformation rules to extracted terms). Each term variant is normalised (see table 1 as an example) and term variants having the same normalised form are then grouped into classes in order to link each term candidate to all of its variants. This way, a list of normalised term candidate classes, rather than a list of single terms is statistically processed. The termhood is then calculated for a whole class of term variants, not for each term variant separately.

Table 1: Term normalisation example

Term variants		Normalised term
human cancers cancer in humans human's cancer human carcinoma	→	human cancer

4.3. Term clustering

Beside term recognition, term clustering is an indispensable component of the literature mining process. Since terminological opacity and polysemy are very common in molecular biology and biomedicine, term clustering is essential for the semantic integration of terms, the construction of domain ontologies and semantic tagging.

ATC in our system is performed using a hierarchical clustering method in which clusters are merged based on average mutual information measuring how strongly terms are related to one another [15]. Terms automatically recognised by the NC-value

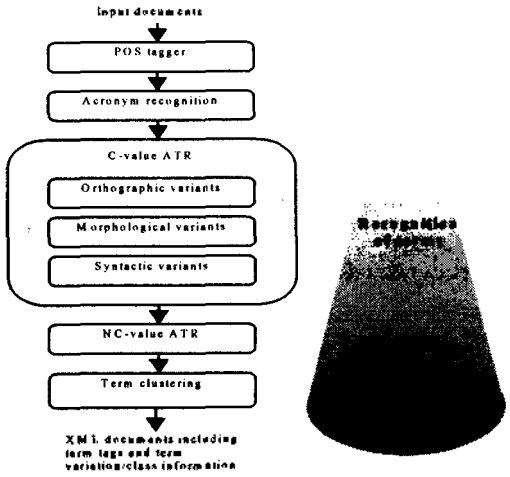

Figure 2: Terminology processing

method and their co-occurrences are used as input, and a dendrogram of terms is produced as output. Parallel symmetric processing is used for high-speed clustering. The calculated term cluster information is encoded and used for calculating semantic similarities in SCE component.

5. Knowledge Acquisition and Knowledge Structuring

Literature mining can be regarded as a broader approach to IE/KA. IE and KA in our system are implemented through the integration of tag- and ontology-based IE and semantic similarity calculation. Graph-based visualization for globally structuring knowledge is also provided to facilitate KR and KA from documents. Additionally, the system supports combining different types of databases (papers and patents, technologies and innovations) and retrieves different types of knowledge simultaneously and crossly. This feature can accelerates knowledge discovery by combining existing knowledge. For example, discovering new knowledge on industrial innovation by structuring knowledge of trendy scientific paper database and past industrial innovation report database can be expected. Figure 3 shows an example of visualization of knowledge structures in the domain of innovation and engineering. In order to structure knowledge, the system draws a graph in which nodes indicate relevant KSs to keywords given and each links between KSs indicates semantic similarities dynamically calculated using ontology information developed by our ATR / ATC components.

6. Experiments and evaluation in an Industry - Academic - Government Project "Go Japan!"

In this section we briefly explain our research project "Go Japan!" and explain the experiments we conducted using analysis reports about Japanese industrial economy to

show the practical performance of our ontology development. Experimental results using scientific papers to show the quality of the ATR/ATC results are also presented.

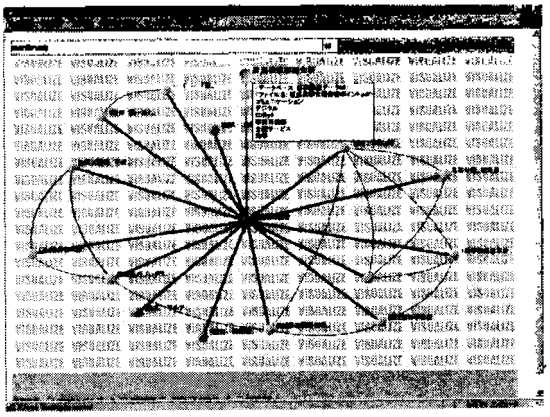

Figure 3: Visualization sample

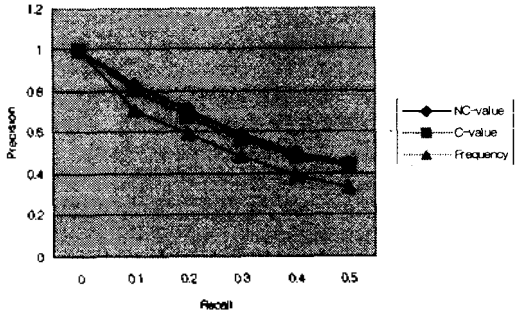

Figure 4: Precision and recall of C/NC-value methods

"Go Japan!" is a private-led project to seek ways to revitalize Japanese industries by studying and discussing the factors behind the decline in Japanese industries productivity and international competitiveness. The main contents of the project are 1) by utilizing innovation and business models based on Japanese potential science and technology, the project will show what industries with what economic effects will be created and what industries can be established locally. It will also show the process of revitalisation to realise such industries, 2) it will investigate science and technology trends, possibilities of industry development patterns of local economics, advantages of industrial competitiveness etc. in the U.S. and Europe, and study how revitalisation should be promoted in Japan. Thus, re-structuring industrial knowledge and accelerating new knowledge discovery are important topics in the project.

The experiments in ATR with the term variation management were conducted on a corpus containing 100,000 words from the "Go Japan!" project to analyze Japanese industrial economic situation, and 100,000 words corpus (2082 abstracts) from the MEDLINE database [1]. A sample of automatically recognised terms is presented in Table 2. As the table shows, reasonable terms to reflect Japanese current industrial situation were properly recognized. Figure 4 shows recall and precision of C- and NC-value methods compared to the frequency of occurrence. As one can see, the NC-value method increases slightly the precision compared to that of the C-value and conventional pure frequency-based methods[4].

Table 2: Sample of recognised terms

Automatically Recognised Terms	Termhood
構造改革 (structural reforms)	430.81
□ (research&development)	261.20
中小企業 (smaller businesses)	259.05
金融機□ (financial institutions)	230.57
東アジア (East Asia)	228.80
社□ (the social security system)	187.34
個人消費 (individual consumption)	171.00
不良債□ (a bad debt)	162.83

For the ATC experiment, we used the GENIA resources [16], which include 1,000 MEDLINE abstracts, with overall 40,000 (16,000 distinct) semantic tags annotated for terms in the domain of nuclear receptors. As a golden standard, we used the GENIA ontology. In the experiment, the test set contained 10,694 terms belonging to any of the three major GENIA classes (namely, *nucleic acid*, *amino acid*, *SOURCE*). These terms have been used as input for the ATC component and the corresponding dendrogram has been produced.

In order to calculate the quality of the dendrogram, we have adopted the average semantic similarity calculation method for measuring the similarity between terms [16]. The average similarity (AS) for two sets of terms is calculated as an average similarity between the corresponding terms:

$$(1)\quad AS(X,Y) = \frac{\sum sim(x,y)}{|X|+|Y|}$$

The similarity between two individual terms is determined according to their position in a dendrogram: a commonality measure is defined as the number of shared ancestors between two terms in the dendrogram, and a positional measure as a sum of their distances from the root. Similarity between two terms corresponds to a ratio between commonality and positional measure.

Table 3: AS-values for the GENIA classes

AS	nucleic acid	amino acid	SOURCE	terms
Nucleic acid	**0.498**	-	-	3108

[4] For detailed evaluation the reader is advised to see [4] and [5].

Amino acid	0.396	**0.492**	-	4284
SOURCE	0.390	0.388	**0.480**	3302

The AS values for all pairs of the three GENIA classes considered in this experiment were calculated (Table 3). The AS values for elements from the same class (i.e. when $X = Y$ in formula (1)) were greater than the values for elements from different classes. This means that terms belonging to the same GENIA class are more closely (i.e. more consistently) placed in the resulting dendrogram. In other words, the average distances between terms belonging to different classes are greater than the average distances within a class. Therefore, we assume that the organisation of terms within the dendrogram produced by ATRACT depicts the actual similarities between them.

7. Conclusion

In this paper, we presented a system for literature mining over large KSs. The system is an XML-based integrated KA system, in which we have integrated ATR, ATC, tagged data management and ontology-based knowledge structuring. It allows users to search and combine information from various sources. IE within the system is terminology-driven, with terminology information provided automatically in the XML format. Similarity based knowledge retrieval is implemented through various semantic similarity calculations, which, in combination with hierarchical, ontology-based matching, offers powerful means for KA through literature mining.

The preliminary experiments show that the system's knowledge management scheme is an efficient methodology to facilitate KA and IE in the field of engineering.

Important areas of future research will involve integration of a manually curated ontology with the results of automatically performed term clustering. Further, we will investigate the possibility of using a term classification system as an alternative structuring model for knowledge deduction and inference (instead of an ontology).

References

[1] National Library of Medicine, MEDLINE, www.ncbi.nlm.nih.gov/PubMed/, 2002.

[2] R. Gaizauskas, G. Demetriou, K. Humphreys, Term recognition and classification in biological science journal articles, Proc. of Workshop on Computational Terminology for Medical and Biological Applications, NLP-2000, Patras, Greece, 2000, pp. 37-44.

[3] K. Fukuda, T. Tsunoda, A. Tamura, T. Takagi, Toward information extraction: identifying protein names from biological papers, Proc. of PSB-98, Hawaii, 1998, pp. 3:705-716.

[4] H. Mima, S. Ananiadou, G. Nenadic, ATRACT workbench: an automatic term recognition and clustering of terms, in: V. Matoušek, P. Mautner, R. Mouček, K. Taušer (Eds.) Text, Speech and Dialogue, LNAI 2166, Springer Verlag, 2001, pp. 126-133.

[5] H. Mima, S. Ananiadou, An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese, Int. J. on Terminology 6/2 (2001), pp. 175-194.

[6] V. Hatzivassiloglou, P. Duboue, A. Rzhetsky, Disambiguating proteins, genes, and RNA in text: a machine learning approach, in: BIOINFORMATICS 17/1 (2001), pp. S97-S106.

[7] P.R.S. Visser, D.M. Jones, T.J.M. Bench-Capon, M.J.R. Shave, An analysis of ontology mismatches - heterogeneity versus interoperability, Proc. of AAAI 1997 Spring Symposium on Ontological Engineering, Stanford University, California, USA, 1997, pp. 164-172.

[8] J. Gamper, W. Nejdl, M. Wolpers, Combining Ontologies and Terminologies in Information Systems, Proc. of the 5th International Congress on Terminology and Knowledge Engineering, Innsbruck, Austria, 1999, pp. 152-168.

[9] T. Berners-Lee, The semantic Web as a language of logic, available at: www.w3.org/DesignIssues/Logic.html

[10] D. Brickle, R. Guha, Resource description framework (RDF) schema specification 1.0, W3C Candidate Recommendation, available at http://www.w3.org/TR/rdf-schema, 2000.

[11] P.G. Baker, A. Brass, S. Bechhofer, C. Goble, N. Paton, R. Stevens, TAMBIS: transparent access to multiple bioinformatics information sources - an overview, Proc. of 6th International Conference on Intelligent Systems for Molecular Biology - ISMB98, Montreal, 1998, pp. 25-34.

[12] A. Voutilainen, J. Heikkila, An English Constraint Grammar (ENGCG) a surface-syntactic parser of English, in: U. Fries et al. (Eds.) Creating and Using English language corpora, Rodopi, Amsterdam, Atlanta, 1993, pp. 189-199.

[13] M. Krauthammer, A. Rzhetsky, P. Morozov, C. Friedman, Using BLAST for identifying gene and protein names in journal articles, in: Gene 259 (2000), pp. 245-252.

[14] C. Jacquemin, Spotting and discovering terms through NLP, MIT Press, Cambridge MA, 2001, p. 378.

[15] A. Ushioda, Hierarchical clustering of words, Proc. of COLING '96, Copenhagen, Denmark, 1996, pp. 1159-1162.

[16] GENIA project, GENIA project home page, www-tsujii.is.s.u-tokyo.ac.jp/GENIA/, 2002.

[17] K. Oi, E. Sumita, H. Iida, Document retrieval method using semantic similarity and word sense disambiguation (in Japanese), J. of Natural Language Processing 4/3 (1997), pp.51-70.

Authors

Hideki Mima, Ph.D., has worked in the area of Natural Language Interface, Machine Translation, Information Retrieval and Automatic Term Recognition. He was a researcher at the ATR Interpreting Telecommunications Research Laboratories, a lecturer at the Department of Computing and Mathematics, Manchester Metropolitan University, and a research associate at the Department of Information Science, University of Tokyo, Japan. Currently, he is a research associate at the School of Engineering, University of Tokyo and is working on Knowledge Acquisition and Knowledge Structuring from various databases/documents in the genome / nano-technology domains. E-mail: mima@biz-model.t.u-tokyo.ac.jp