2-2 Information Access Technologies for Processing a Very Large Number of Natural Language Documents

MURATA Masaki

We have developed various information access technologies for information retrieval, information extraction (text mining), question answering, and document classification used in processing natural language documents. The effectiveness of these technologies was confirmed when they produced the highest level of precision in the NTCIR evaluation workshop. As the number of electronic documents continues to increase, these information access technologies will become increasingly useful.

Keywords
Information retrieval, Information extraction, Text mining, Question answering, Document classification

1 Introduction

As the number of existing electronic documents grows day by day, there is a corresponding increase in the need for technologies that will enable access to the information within these documents. To respond to this need, the authors have developed a variety of natural-language information-access techniques for applications including information retrieval, information extraction, question answering, and automatic document classification. The effectiveness of these techniques was confirmed through demonstration of many instances of highest level of precision in an NTCIR evaluation workshop. This paper presents a description of these techniques.

2 Information retrieval

Information retrieval is a technique for retrieving documents that contain information users wish to retrieve, described in the form of key words or phrases, from a large group of documents. In particular, in the context of natural language processing in recent years, this process involves the retrieval of documents containing desired information described by users more commonly in the form of phrases, rather than through key words[1][2].

For example, suppose a user enters sentences to describe the information to be retrieved, as in Fig. 1. The user retrieves documents that match this information. That is, a document (article) such as that shown in Fig. 2

Topic: Corporate merger
Description: An article announces a corporate merger, and I can research the names of the companies participating in the merger. Additionally, I can research the business field of the merged company, the purpose of the company's establishment, and other specifics. Here, the concept of corporate merger includes absorption, integration, and acquisition.

Fig. 1 Information a user wishes to retrieve

is retrieved.

In typical information retrieval, information that users wish to retrieve (such as the information shown in Fig.1), is morphologically analyzed by separating words and determining parts of speech. Words that are nouns are isolated, and documents containing relatively more of these words (called “key words”) are retrieved. Additionally, words that rarely appear in documents are heavily weighted while very common words that appear in most documents are lightly weighted. Documents containing the more heavily weighted words are then retrieved. Several word-weighting methods are available; we employed the BM 25 method[3], previously demonstrated as an effective weighting technique. Tf-idf is another method applied to information retrieval. In the tf-idf method, tf represents the number of key words appearing in documents for retrieval, and df is the number of documents in the database in which the key words appear. With N as the total number of documents, key word weights are calculated as tf · log (N/df). BM 25 is an improved version of the tf-idf method using an equation in which the effects of tf are weaker than under the tf-idf method.

Additionally, we applied a process known as the automatic feedback method[4]. With this method, documents are initially retrieved and additional words are identified that appear frequently among the most relevant documents retrieved. These words are then added to the key words for the initial search, and document retrieval is performed again. Words among the initial-search key words that appear frequently in the documents in the most relevant documents retrieved are given greater weight. Documents are retrieved again after the newly identified words have been weighted, and these secondary results are considered the final results of document retrieval. In the results of initial document retrieval, words that appear frequently in the most relevant documents retrieved are often synonyms of the original key words, and adding these key words improves the effectiveness of document retrieval. This is recognized as an effective method, and it is among the methods we adopted.

Kygnus to become a wholly owned subsidiary of Tonen

On the 16 th, Tonen announced that it will make group member company Kygnus (capital: ¥ 1 billion; Headquarters: Kawasaki, Kanagawa; President: Toshihide Mori) a wholly owned subsidiary. Stakes in Kygnus are currently held by Tonen (70%) and Nichimo (30%), and Tonen will purchase all 600,000 shares held by Nichimo for ¥12.5 billion.

Tonen is a leading oil producer with an 8% share of facilities as of the end of fiscal 1993. The acquisition of Kygnus will bring the share to 9.4%.

Nichimo has decided to transfer the stocks to raise funds for the purpose of covering expenses, including those incurred by plant closures in Yamaguchi Prefecture.

With the impending annulment of the Provisional Measures Law on the Importation of Specific Petroleum Refined Products next spring, the oil industry is promoting cost-cutting and improved efficiency, and restructuring — including integration of group companies — is finally beginning in earnest.

Fig.2 Retrieved article

In addition to BM 25 and the automatic feedback method previously described, we employed a method using document title information. Documents in which the search key words appear in the title are more likely to contain the information sought. Thus, we devised a method for retrieval of documents in which heavily weighted key words appear in the title. The application of this method led to the establishment of a technique of demonstrably effective information retrieval[1][2].

To evaluate the various processes related to information retrieval, workshops, organized under the acronym NTCIR[5], are held. Here, several groups answer the same question, and the precision of their results is compared. Evaluating the precision of information retrieval involves determining how many documents that should have been retrieved are actually found in the top retrieval results. In this workshop, we noted many instances of the highest precision[1][2][6], thus proving the effectiveness of our technique.

3 Information extraction (text mining)

Information extraction is a technique for extracting useful information from a large amount of natural language data. The process is also referred to as “text mining”. As one form of information extraction, we studied research trends by analyzing bibliographical document information and similar data. Here we introduce results pertaining to the journal of the Association for Natural Language Processing, *Natural Language Processing*[7][8].

As the basis of this text mining process, we determined word frequency over a period of time. We performed analysis to determine which words appeared more or less frequently over this period to investigate how topics of interest had shifted.

We selected words appearing in titles of the papers published in the journal of the Association for Natural Language Processing. Taking these words as key words indicating the area of research, we investigated the frequency of papers in which the key word appeared in the title. The results are shown in Fig. 3. Key words that were not indicative of an area of research (such as the Japanese adjectival suffix “teki”, or the word “study”) were removed manually. The study focused on the association’s annual conference papers over ten years. Publication dates extended roughly from 1994 to 2003.

In Fig. 3, results of analysis are expressed using contour lines. The height of the contour lines (indicated by shading) shows the number of papers in which the words appear. In the data for the publication frequency of each area of research, we determined the average value, value of greatest frequency, and median for each year of publication, and then the average for these three values, which then formed the basis for the order of the listing in the figure (in increasing order). Each word representing an area of research is labeled with the total number of publications and the average, as described above. Accordingly, in the figure

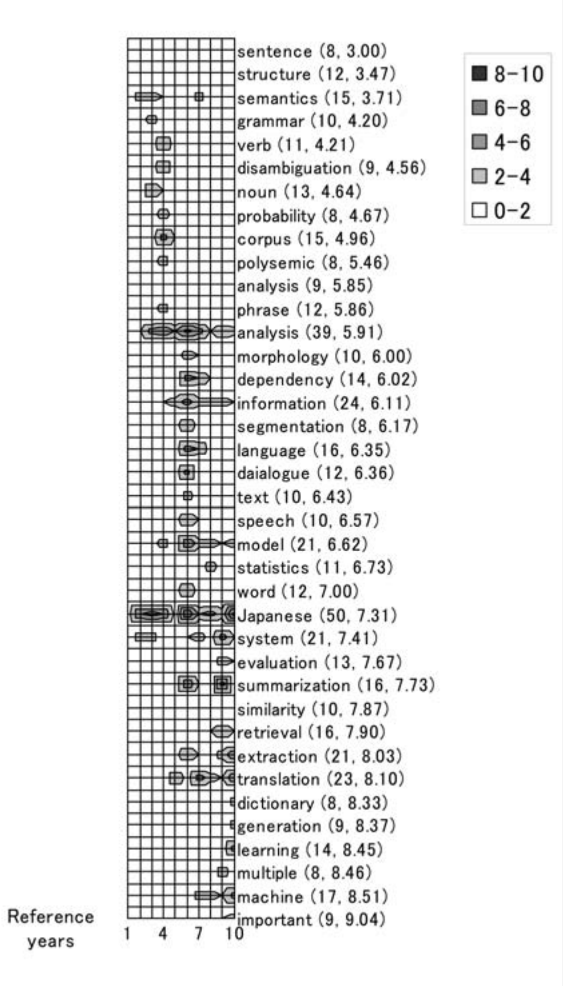


Fig.3 Trends in publication frequency, by research area

The height of the contour lines (indicated by shading) shows the number of papers in which the words appear. In the figure, words representing each research area are labeled with two numbers, the first being the total number of papers published and the second being the average in the year of most publications. (For detailed definitions, refer to the body text.)

the research areas at the top have a higher publication frequency in the earlier years of publication, and areas at the bottom have a higher publication frequency in the more recent years of publication. This method of representing the data effectively enables confirmation of which areas of research were popular in earlier years of publication and which were popular in later years. Similarly, this approach can be viewed as an effective method of representation in various text-mining applications. A patent has been filed with respect to this particular representation technique.

From Fig. 3, it is clear that the key words “Japanese” and “analysis” were most prevalent. It is also clear that because the research areas designated with “verb”, “noun”, “disambiguation”, “probability”, “corpus”, and “polysemic” appear higher on the list, these areas were studied intensively in the earlier years for publication in journal papers. Similarly, it is clear that the research areas of “morphology”, “dependency”, “dialogue”, and “speech” were studied more intensively in the sixth year, while the areas of “summarization”, “retrieval”, “translation”, and so on were the focus of study in later years. Because special issues on “summarization” were published in the sixth and ninth years, this research area was the topic of more papers in these periods. We anticipate that, in light of the increasing publication of research in the area of “translation”, this topic will continue to be the focus of many papers in the future.

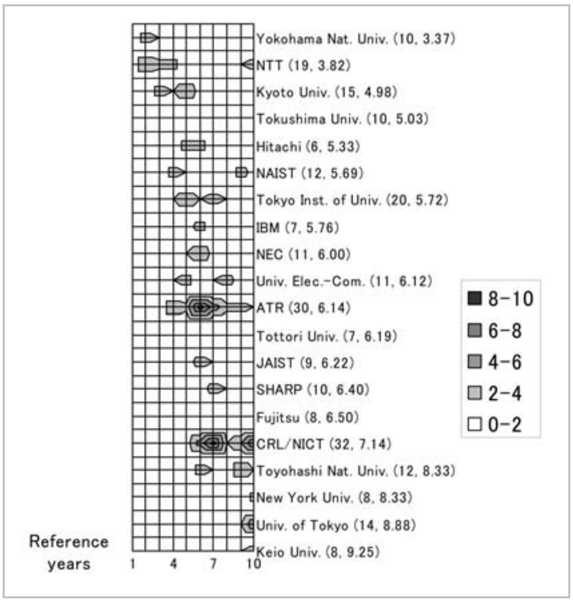


Fig.4 *Trends in publication frequency, by research organization*

The height of the contour lines (indicated by shading) shows the number of papers in which the organization names appear. In the figure, research organizations are labeled with two numbers, the first being the total number of papers published and the second being the average in the year of the most publications. (For detailed definitions, refer to the body text.)

Next we investigated research organizations themselves, counting the number of research papers published by each. The results are shown in Fig. 4.

In the data for the publication frequency of each organization, we determined the average value, value of greatest frequency, and median for each year of publication, and then the average for these three values, which then formed the basis for the order of the listing in the figure (in increasing order). Each research organization is labeled with the total number of publications and the average, as described above. Accordingly, in the figure, organizations at the top have a higher publication frequency in the earlier years of publication, and organizations at the bottom have a higher publication frequency in the more recent years of publication. Here, only the organizations with the highest total number of publications are shown. Organizations that were renamed are identified by the name under which their publication record is most extensive.

From the figure, it is clear that organizations with the highest publishing volume were the Communications Research Laboratory (now the National Institute of Information and Communications Technology) and ATR (now affiliated with the National Institute of Information and Communications Technology, centered in the Natural Language Processing Group). In the area of natural language processing, our own research organization (National Institute of Information and Communications Technology) showed excellent results. Additionally, it is clear that while NTT and ATR were leading publishers in earlier years, publication by InfoCom Research and the University of Tokyo is concentrated in later years. We anticipate the current increasing trend in publications by InfoCom Research and the University of Tokyo to continue. As for the remaining organizations, the figure clearly indicates the years in which the respective publications have been the most extensive.

4 Question answering system

A question answering system is a system that accurately responds to questions from people. For example, in response to the question “How high is Mt. Fuji, in meters?” the answer given is “3,776 meters” (Fig. 5). Instead of involving a database prepared in advance with knowledge of specific questions and answers, these systems are distinguished by their extraction of answers from a large amount of natural-language sentences. In some cases, the automatic answering of such questions truly conveys the impression of intelligence on the part of the responding computer or other device. Question answering systems represent a truly fascinating research topic, in some ways related to research on artificial intelligence aimed at imparting human levels of cognition to computers[9].

4.1 Importance of question answering systems

The following three considerations underscore the importance of question answering systems.

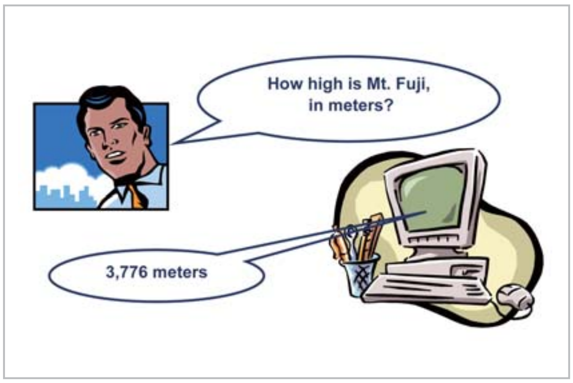


Fig.5 Question answering system

(1) More convenient than information retrieval (document retrieval)

Because question answering systems present the exact answer to user questions, users need not read through numerous documents, enabling easier retrieval of the desired information. Information retrieval systems only present documents that appear relevant, and users must read through each such document to find the desired information.

(2) Able to extract knowledge from vast amounts of natural language data

It is clearly convenient to be able to obtain information freely from vast amounts of data stored in the form of natural language. In question answering systems, this acquisition occurs through the process of asking a natural language question in order to elicit an answer.

(3) Capable of incorporation within other knowledge-processing systems

The service provided by question answering systems-producing answers to natural language questions-renders these systems potentially useful in other knowledge-processing systems. For example, question answering systems can even fulfill a role in the context of guessing referents in instructions. To guess what “saline solution” refers to in the sentences “Mix salt and water. Using this saline solution . . .” we can ask the system “What is made by mixing salt and water?” to obtain the answer “Saline solution.” Uses such as this are not limited to problems involving guessing referents; potential applications include a variety of natural-language processing and knowledge-processing systems. Question answering systems thus have potential for use in other intelligent processing systems. Question answering systems can be viewed as the minimum technological requirement when developing artificial intelligence systems and may in fact form the basis of future intelligent processing and knowledge processing systems.

4.2 Typical structure of question answering system

The typical structure of current question answering systems is as follows.

(1) Estimation of the form of the answer

Based on the form of the question (which may use an interrogative pronoun) and other factors, the system estimates the form of the answer (that is, what kind of linguistic expression it is). For example, if the supplied question is “About how much area does Japan cover?” the expression “about how much” forms the basis for estimating that the answer will be expressed numerically.

(2) Document retrieval

The system extracts key words from the question and these key words are then used for document retrieval. Retrieval presents a group of documents that seem likely to include the answer. For example, if the supplied question is “About how much area does Japan cover?” then “Japan” and “area” are extracted as key words, and documents that include these words are retrieved.

(3) Answer extraction

From the group of documents that seem likely to include the answer, the system extracts linguistic expressions matching the estimated form of the answer and presents these as answers. For example, in response to the question “About how much area does Japan cover?” the linguistic expressions presented as answers are extracted from the group of documents (which contain “Japan” and “area”) retrieved. These expressions match the form of the estimated answer (which is a numerical expression).

4.3 The authors’ question answering system

Our question answering system also adopts the method described above. Additionally, our system’s method of answer-extraction presents, out of the linguistic expressions matching the estimated form of the answer, expressions that closely reproduce the key words taken from the question. Furthermore, we devised a simple and easy method for increasing the effectiveness of the method through the comprehensive use of information in multiple groups of articles. Often the answer to a question appears in multiple articles. In this case, using multiple articles instead of one when estimating answers can yield better answers. Thus, in our method, individual candidate answers obtained from multiple articles are given point values, are then summed, then added to the list of potential answers. However, merely summing points related to candidate answers would in fact dilute the effectiveness of the system. Thus, our method compensates for the drawback of merely summing points of candidate answers by reducing the list of candidates as newer candidates are added. Specifically, our question answering system uses the following methods[10].

Evaluative workshops on information retrieval (NTCIR) also include question answering tasks. Here, too, several groups respond to the same question in a question answering session, and the precision of their results is compared. At this workshop, we achieved many instances of the highest level of precision[11][12]. These results prove the effectiveness of our technique.

5 Automatic document classification

Here, “automatic document classification” refers to a technique for automatically classifying documents according to a predetermined classification system. As an effective way to classify and organize large numbers of documents, automatic document classification is an important technique in the context of information access. The automatic document classification we studied involves automatic classification of patent documents (Fig. 6).

In Japan, patent documents are classified by the assignment of an F-term. A single patent is assigned multiple F-terms. The F-term patent classification system is unique to patents in Japan, where it provides a useful perspective on patents from a variety of standpoints, including purpose or application, structure, and material. Using F-terms, patent characteristics can be analyzed in much greater detail than before. For example, if a chart were to be created based on particular prob-

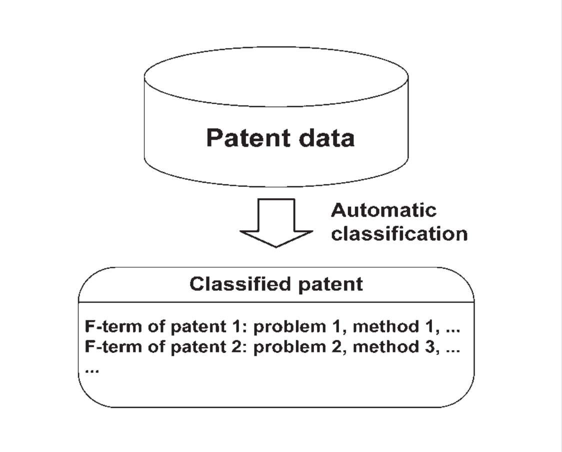
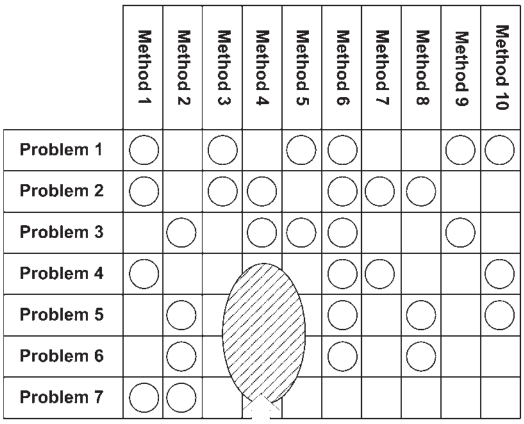

Patent data
Automatic classification
Classified patent
F-term of patent 1: problem 1, method 1, ...
F-term of patent 2: problem 2, method 3, ...
...

Fig.6 Automatic patent classification

lems and methods of solving them, one could in theory estimate what kinds of patents may still be obtained (Fig. 7).

We investigated the estimation of the F-terms that should be applied to patents using an improved k neighborhood method. With the k neighborhood method[13], k represents the collected number of patent cases that most resemble the cases to be classified. The classification most commonly applied to these patent cases is applied to the patent cases awaiting classification. To improve on this method, we devised a way to facilitate answers, such that it is as easy to select these answers as it is to assign classifications to patent cases, or to identify patent cases that are quite similar. We added a framework for automatically learning to what extent these easily answered classifications should be considered to appropriate answers. When collecting a number k of similar patent cases, it is also necessary to determine similarity among these patent cases, so BM 25 (viewed as highly effective) or other methods were used in information retrieval. These methods were used for automatic assignment of F-terms.

The NTCIR[5] evaluative workshop on information retrieval also included tasks in F-term classification of patent documents. Here, several groups offer a solution to the same question on F-term classification of patent documents, and the precision of their results is compared. In this workshop, by employing the improved k neighborhood method described above, we were able to demonstrate the greatest precision[14]. Other participating teams used methods such as vector space models with word vectors or support vector machines based on mechanical learning. In comparison with the other teams, our results were approximately 15% more precise.



	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10
Problem 1	○		○		○	○			○	○
Problem 2	○		○	○		○	○	○		
Problem 3		○		○	○	○			○	
Problem 4	○					○	○			○
Problem 5		○				○		○		○
Problem 6		○				○		○		
Problem 7	○	○								

The part in the gray oval is not covered by any patents, which indicates the possibility of filing promising new patents, that cover handling problems 4 to 7 using methods 3 to 5.

Fig.7 Discovery of the potential for a new patent

6 Conclusions

The authors have developed a variety of natural language information access techniques for applications including information retrieval, information extraction, question answering, and automatic document classification. As the number of existing electronic documents grows day by day, there is a corresponding increase in the need for technologies that will enable access to the information within these documents.

References

1 Masaki Murata, Qing Ma, Kiyotaka Uchimoto, Hiromi Ozaku, Masao Utiyama and Hitoshi Isahara, “Information Retrieval Using Location and Category Information”, Journal of Natural Language Processing, Vol.7, No.2, pp.141-160, 2000. (in Japanese)

2 Masaki Murata, Qing Ma, Kiyotaka Uchimoto, Hiromi Ozaku, Masao Utiyama and Hitoshi Isahara, “Japanese Probabilistic Information Retrieval Using Location and Category Information”, IRAL’2000, Hong Kong, Sep. 30, 2000.

3 Stephen E. Robertson, Steve Walker, Susan Jones, Micheline. Hancock-Beaulieu, and Mike Gatford, “Okapi at TREC-3”, Proceedings of the third Text REtrieval Conference (TREC-3), pp.109-126, 1994.

4 Lisa Ballesteros and W. Bruce Croft, “Phrasal translation and query expansion techniques for cross-language information retrieval”, In Proceedings of the 20 th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR’97), pp.84-91, 1997.

5 Yasushi Ogawa, Yutaka Sasaki, Shigeru Masuyama, Masaki Murata, Masaharu Yoshioka, NTCIR from Participants’ Viewpoints, Journal of the Japanese Society for Artificial Intelligence, Vol.13, No.3, pp.306-311, 2002. (in Japanese)

6 Masaki Murata, Qing Ma, and Hitoshi Isahara, “Applying Multiple Characteristics and Techniques to Obtain High Levels of Performance in Information Retrieval”, Proceedings of the NTCIR Workshop 3 (CLIR), 2002.

7 Masaki Murata, Koji Ichii, Qing Ma, Tamotsu Shirado, Toshiyuki Kanamaru and Hitoshi Isahara, “Trend Survey on Journal and Conference Papers Published by the Association for Natural Language Processing Over the Last Decade”, Web Site of the Association for Natural Language Processing, (http://www.nak.ics.keio.ac.jp/NLP/trend-survey.html), 2007. (in Japanese)

8 Masaki Murata, Koji Ichii, Qing Ma, Tamotsu Shirado, Toshiyuki Kanamaru and Hitoshi Isahara, “Trend Survey on Japanese Natural Language Processing Studies over the Last Decade”, The Second International Joint Conference on Natural Language Processing, Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts, pp.252-257, Jeju Island, Korea, Oct. 2005.

9 Masaki Murata, “Question Answering System: The Current Status and Perspective”, The Institute of Electronics, Information and Communication Engineers, Vol.86, No.12, pp.959-963, 2003. (in Japanese)

10 Masaki Murata, Masao Utiyama, and Hitoshi Isahara, “Use of Multiple Documents as Evidence with Decreased Adding in a Japanese Question-answering System”, Journal of Natural Language Processing, Vol. 12, No. 2, pp.209-247, 2005.

11 Masaki Murata, Masao Utiyama and Hitoshi Isahara, “Decreased-Adding-Based Question-Answering System Using Simple Connection Method For Contextual Questions”, Journal of the ISCIE (the Institute of Systems, Control and Information Engineers), Vol. 20, No. 8, 2007. (in Japanese)

12 Masaki Murata, Masao Utiyama and Hitoshi Isahara, “Japanese Question-Answering System For Contextual Questions Using Simple Connection Method”, Decreased Adding with Multiple Answers, and Selection by Ratio, Asia Information Retrieval Symposium (AIRS) 2006, Shangri-La’s Rasa Sentosa Resort, Singapore, Oct.16, pp.601-607, 2006.

13 Yiming Yang and Xiu Liu, “A re-examination of text categorization methods”, Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999.

14 Masaki Murata, Toshiyuki Kanamaru, Tamotsu Shirado, and Hitoshi Isahara, “Automatic F-term Classification of Japanese Patent Documents Using the k-Nearest Neighborhood Method and the SMART Weighting”, Journal of Natural Language Processing, Vol.14, No.1, pp.163-190, 2007.

***MURATA Masaki**, Ph. D.*

Senior Researcher, Computational Linguistics Group, Knowledge Creating Communication Research Center (Former: Senior Researcher, Computational Linguistics Group, Keihanna Human Info-Communication Research Center, Information and Network Systems Department)

Natural Language Processing