# Towards Multi-paper Summarization Using Reference Information

**Hidetsugu NANBA** and **Manabu OKUMURA**
School of Information Science
Japan Advanced Institute of Science and Technology
Tatsunokuchi Ishikawa 923-1292 Japan
Tel:(+81-761)51-1216, Fax: (+81-761)51-1370
{nanba, oku}@jaist.ac.jp

## Abstract

This paper presents a system to support writing a survey of a specific domain. The system utilizes reference information that consists of reference relationships between papers and the information derived from the description around citations. We think the following are inevitable for writing a survey : collecting papers of the specific domain, and understanding their essence and differences among them. Therefore, we firstly extract fragments of papers where the author describes the essence of a referred paper and the differences between his paper and it(we call them **reference areas**). Then with the information of reference areas, we identify the types of reference relationships that indicate the reasons for citations(we call them **reference types**). These types make it possible to collect papers in the same domain. The system can display the collection of the papers. It can also show abstracts and reference areas of the collected papers. With the system, we can understand the relationships between the collected papers.

## 1 Introduction

Recently, scientific information explosively increases, because of the increase of the number of researchers and the ramification of research domains, and it is difficult for researchers to read all the papers that they can get. In this situation, surveys of specific domains can make it easier to grasp the outlines of the domains. However, the number of surveys we can obtain is very small [Garvey, 1979], because it is quite time consuming to write them.

Now we are studying towards automatic survey generation. A survey can be considered to be a summary of multiple papers, and should describe their essence and differences among them. Furthermore, it is desirable that the author describes in the survey his viewpoint and what are needed to study in the domain. However, the task to generate such a survey automatically seems very difficult.

This paper presents a system to support writing a survey of a specific domain. The system utilizes reference information that consists of reference relationships between papers and the information derived from the description around citations. We think the following are inevitable for writing a survey : collecting papers of the specific domain, and understanding their essence and differences among them. Therefore, we firstly extract fragments of papers where the author describes the essence of a referred paper and the differences between his paper and it(we call them reference areas). Then with the information of reference areas, we identify the types of reference relationships that indicate the reasons for citations. These types make it possible to collect papers in the same domain. The system can display the collection of the papers. It can also show abstracts and reference areas of the collected papers. With the system, we can understand the relationships between the collected papers.

In the following sections, we first explain the essential points of multi-paper summarization, and then we describe reference information and its usage for summarization. We also report some experiments of reference information extraction and show the support system for writing a survey.

## 2 Multi-paper Summarization

### 2.1 Essential Points of Multi-paper Summarization

There have been several techniques proposed for summarizing a technical paper [Kupiec et al., 1995; Teufel and Moens, 1997; Mani and Bloedorn, 1998]. However, in case of multi-paper summarization, we have to perform more tasks than just extracting important fragments from each paper. First, we should take into account how to collect the target papers for summarization. Second, a multi-paper summary should clearly describe the similarity and differences among papers. Therefore, we can sum up the essential points of multi-paper summarization as in Figure 1.

- (a) *retrieval of papers in a specific domain*
- (b) *extraction of information from papers of the domain*
  - (b)−1 *detection of important fragments from each paper*
  - (b)−2 *detection of similarity among papers*
  - (b)−3 *detection of differences among papers*

Figure 1: The essential points of multi-paper summarization

## 2.2 Related Works

Kando proposed several rules using lexical cues to analyze the functional structure of technical papers, and used the structure for paper retrieval [Kando, 1997]((a) in Figure 1). She also pointed out that sentences that are assigned particular categories(e.g."Review of relevant previous research") in the collected papers are useful for grasping the outlines of the studies in a specific domain(b). Similarly, we use reference information for multi-paper summarization.

In the field of bibliometrics or citation analysis, there have been many related researches [Narin et al., 1994; Liu, 1993; White and Mccain, 1989]. Their basic idea is that the papers that jointly cite others are related to each other, and the purposes of these researches are to find an important citation from citation diagrams and to identify which papers should be included in a survey. This can be considered as the first step of our study to make the multi-paper summarization system. To realize this process, we develop the method to decide the reference types automatically.

# 3 Reference Information in Multi-paper Summarization

## 3.1 Reference Information

Consider 5 sentences in Figure 2 that are extracted around the citation of the referred paper [Murata and Nagao, 1993] in the referring paper [Bond, et al., 1996]. Both papers are on machine translation, and particularly deal with noun phrases including numerical expressions. Sentence (2) introduces the theme of the referred paper. Sentence (3) points out the problem of the referred paper. Then, sentence (4) describes that the referring paper copes with the problem pointed out in sentence (3).

By reading sentences (2),(3),(4), we can understand the relationships between [Murata and Nagao, 1993] and [Bond, et al., 1996]. We call the fragment **reference area**. With the information in it, we can also identify the reason for citation. We classify the reason for citation into the following three categories(we call these categories **reference types**), based on 15 categories proposed by Weinstock[Weinstock, 1971].

- **type B**
  The references to base on other researchers' theories or methods.
- **type C**
  The references to compare with related works or to point out their problems.
- **type O**
  The references other than types B and C.

in [Bond, et al., 1996]

(1)In addition, when Japanese in translated into English, the selection of appropriate determiners is problematic.
**(2)Various solutions to the problems of generating articles and possessive pronouns and determining countability and number have been proposed [Murata and Nagao, 1993].**
**(3)The differences between the way numerical expressions are realized in Japanese and English has been less studied.**
**(4)In this paper we propose an analysis of classifiers based on properties of both Japanese and English.**
(5)Our category of classifier includes both Japanese josushi 'numeral classifiers' and English partitive nouns.

reference area : sentences(2)∼(4)

Figure 2: The reference area of type C

- ($\alpha$) *introduction of previous research*
- ($\beta$) *description about the problem of previous research*
- ($\gamma$) *the purpose of the research in the referring paper*

Figure 3: The information obtained from type C reference area

We think the references of type C are more important than others, because from reference areas of type C, we can obtain information shown in Figure 3. In case of the example in Figure 2, sentences (2), (3), and (4) correspond to ($\alpha$), ($\beta$), and ($\gamma$), respectively. Here, ($\alpha$) can be considered as a kind of summary of the referred paper from the author's viewpoint. ($\alpha$) can also be regarded as a fragment that describes the similarity of research topics between two papers. Reading sentence (2), we can understand that the research topic of both papers is on generating articles and possessive pronouns and determining countability and number in English sentences. On the other hand, the problem of previous work and the purpose of research are described in sentences (3) and (4) respectively. These sentences can be regarded to describe differences between two papers.

## 3.2 Using Reference Information for Multi-paper Summarization

### Retrieval of Papers Using Reference Information

If we collect papers by tracing all reference relationships, many papers in other domains are also included in the

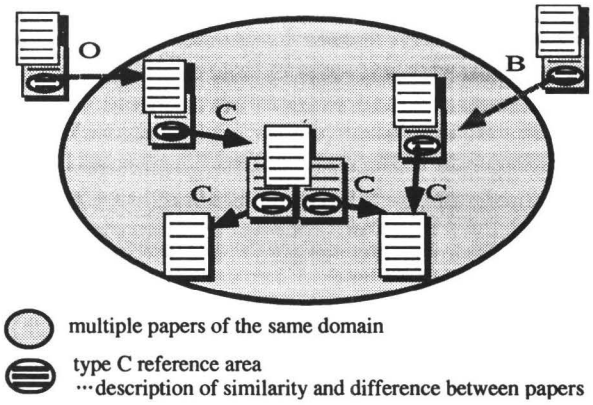


Figure 4: Similarity and difference between papers

collection. By tracing only type C reference relationships, however, we can collect the papers in the same domain. By investigating 31 cases of type C reference relationships in the database we will explain later, we actually find that both referring and referred papers belong to the same domain in 29 cases(94%). Figure 4 illustrates the target collection of papers for summarization(we call the collection in the shadow oval **reference graph**).

### Detection of Similarity and Difference among Papers

Making correspondence of the information in type C reference area (in Figure 3) to the essential points of multi-paper summarization(in Figure 1), $(\alpha)$ corresponds to $(b)-1,2$, and $(\beta),(\gamma)$ correspond to $(b)-3$. Therefore, extracting and displaying reference areas can be a good support for writing a survey.

## 4 Support System for Writing A Survey

In this section, we explain the method to realize the support system for writing a survey.

### 4.1 Analysis of Reference Relationships between Papers

We use as the database 450 papers in TEX style source on computational linguistics from e-Print archive[1]. Since TEX has commands to write the bibliography such as "`\cite`" or "`\bibitem`", by analyzing information in such commands, we can get the information of reference relationships among papers. For the database, we can obtain the relationships in the accuracy of 94%.

### 4.2 Extraction of Reference Areas

Reference areas can be considered as a succession of sentences that have a connection with the sentence including citation in the paragraph. Since we think that such a connection between sentences can be indicated by 6 kinds of cue words, we use those cue words for reference area extraction. Examples of cue words are shown in Table 1. The procedure to select cue words is as follows:

1. Create the reference area corpus by hand,
2. Apply n-word gram analysis to this corpus,
3. Select 86 cue words manually, by checking the list of frequently used expressions made in step 2.

The flow chart of reference area extraction is shown in Figure 5. Reference area extraction rules add to the reference area candidate one sentence before or after the candidate if it includes any cue words.

[1]http://xxx.lanl.gov/cmp-lg/

Table 1: Examples of cue words for reference area extraction

| | |
|---|---|
| **(1)anaphor** | In this, On this, Such |
| **(2)negative expression** | But, However, Although |
| **(3)1st person pronoun** | We, we, Our, our, us, I |
| **(4)3rd person pronoun** | They, they, Their, their, them |
| **(5)adverb** | Furthermore, Additionally, Still |
| **(6)other** | In particular, follow |

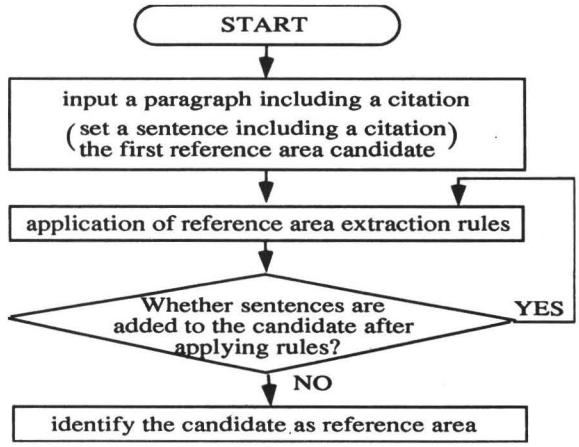


Figure 5: Flow chart of reference area extraction

### 4.3 Identification of Reference Types

In a reference area, if a negative expression appears at the beginning of the sentence after the sentence including citation, the reference area can be considered as type C. Similarly if the expression like "we adopt" or "we use" appears in the sentence including citation, the reference area can be considered as type B. Therefore, we prepare a list of cue words and make a set of rules using them for reference type identification.

Biber explicated the difference of linguistic features among standard four-part organization in medical papers: Introduction, Methods, Results, and Discussion [Biber and Finegan, 1994]. In making a list of cue words, we also pay attention to the difference among the sections. Type C reference areas tend to occur in Introduction, Related Works and Discussion. On the other hand, type B reference areas tend to occur in Introduction and

Table 2: Examples of cue words for type C

| | | |
|---|---|---|
| Although, | Though, | ,although |
| However, | however, their | however, the |
| but the | but it | But they |
| In spite of | Instead of | But instead |

Table 3: Examples of cue words for type B

| | | |
|---|---|---|
| based mainly on | basis | is based on |
| the basic | used in | uses of |
| used by | to use a | can use |
| that can | We can | We use |

Experiment. Therefore, we make a list of cue words with the following procedure:

1. Collect sentences for type B and C from corresponding sections,
2. Calculate n-word gram separately,
3. Apply cost criteria [Kita, et al., 1994], which tends to extract longer expressions, to the result of n-word gram statistics,
4. Select 76 cue words for type C and 84 for type B manually, by checking the list of frequently used expressions made in step 3.

Examples of cue words are shown in Table 2 and 3. Then we make 160 rules for reference type identification manually.

## 5 Experiments of Reference Information Extraction

We have conducted two experiments to evaluate the effectiveness of our methods.

### 5.1 Extraction of Reference Areas

For the experiment, we prepare 150 reference areas, that are manually identified from paragraphs including citation. We use 100 for making rules and 50 for evaluation. We use F-measure[van Rijsbergen, 1979] in formula(1) for evaluation, where b=1.

$$F(Fmeasure) = \frac{(1+b^2)PR}{b^2P+R} \quad (1)$$

where P and R are as follows:

$$R(Recall) = \frac{\left(\begin{array}{l}\textit{The number of sentences}\\ \textit{correctly extracted by rules}\end{array}\right)}{\left(\begin{array}{l}\textit{The number of sentences}\\ \textit{which should be extracted}\end{array}\right)}$$

$$P(Precision) = \frac{\left(\begin{array}{l}\textit{The number of sentences}\\ \textit{correctly extracted by rules}\end{array}\right)}{\left(\begin{array}{l}\textit{The number of extracted}\\ \textit{sentences by rules}\end{array}\right)}$$

Table 4: The accuracy of reference area extraction

| | Recall(%) | Precision(%) | F-measure |
|---|---|---|---|
| our method (in data for making rules) | 90.9 | 76.9 | 0.833 |
| our method (in data for evaluation) | 79.6 | 76.3 | 0.779 |
| baseline 1 | 100.0 | 36.4 | 0.534 |
| baseline 2 | 40.4 | 100.0 | 0.575 |

Table 5: The accuracy of reference type identification using evaluation data

| | | reference type identified by rules | | | accuracy for each type(%) |
|---|---|---|---|---|---|
| | | C | B | O | |
| correct reference type | C | **12** | 0 | 4 | 75.0 |
| | B | 2 | **25** | 5 | 78.1 |
| | O | 1 | 5 | **46** | 88.5 |

the accuracy of reference type identification in evaluation data: 83.0(%)

The results are shown in Table 4. To compare with the performance of our method, we calculate two baselines.

If we consider the whole paragraph as a reference area, all the sentences which should be extracted are extracted. In this case, the score of F-measure is 0.534(Recall/Precision: 100.0/36.4%). On the other hand, if we consider only the sentence including citation as a reference area, the extracted sentence is always correct as a sentence in the reference area. In this case, the score of F-measure is 0.575(Recall/Precision: 40.4/100.0 %).

As can be seen in Table 4, the performance of our method is better than the baselines.

### 5.2 Identification of Reference Types

For the experiment, we prepared 382 reference areas whose reference types are manually identified. Then we use 282 for making rules and 100 for evaluation. The results are shown in Table 5. The sum of bold numbers in Table 5 shows the number of reference areas whose reference types are correctly identified by rules. Therefore, we obtain the accuracy of 90.1(%) and 83.0(%) in the data for making rules and evaluation, respectively.

## 6 Using the Support System for Writing a Survey

We show in Figure 6 the process of using the support system. The process consists of two stages. One is the stage of paper retrieval, and the other is the stage of overviewing the similarity and difference among papers.

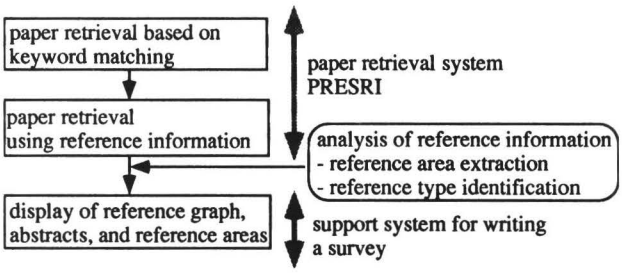


Figure 6: Paper retrieval system and support system for writing a survey

In the first stage, PRESRI(**P**aper **RE**trieval **S**ystem using **R**eference **I**nformation) provides 2 methods of paper retrieval. One is retrieval by queries (using authors' names and/or terms in titles). The other is retrieval by using reference relationships between papers. If the paper retrieved by queries has reference relationships with other papers in the database, the graph of the reference relationships around it can be displayed. By tracing this graph, we can retrieve other papers.

In the second stage, by clicking the button "show type C reference relationships" on the display, we can see the reference graph. By clicking icons of abstracts or reference areas, we can see abstracts and reference areas of the papers in the graph.

Figure 7 shows the display of the support system. The left window shows the reference relationships around [Murata and Nagao, 1993](9405019). 4 papers refer it, and [Bond96] (9601008) is darkened because three papers other than it refer [Murata and Nagao, 1993] in type C.

The right window displays three reference areas, where [Takeda, 1994](9407008), [Bond, et al., 1994](9511001), and [Bond, et al., 1996](9608014) describe [Murata and Nagao, 1993]. In this way, by displaying several abstracts and reference areas, the sytem can assist our overview of similarity and differences among papers. Therefore we think it is useful for writing a survey. Now this system is fully implemented and can be used on World Wide Web[2].

## 7 Conclusion

In this paper, as the first step towards automatic survey generation, we developed a support system for writing a survey by utilizing reference information between papers. By using the system, we can collect papers in the same domain. Additionally, the system displays the important parts and the descriptions of difference between papers, which are extracted from the collection of papers. We are now studying towards automatic survey generation based on the method described in this paper.

[2]http://galaga.jaist.ac.jp:8000/pub/tools/sum

## Acknowledgements

The authors would like to express our gratitude to E-Print archive administrators for allowing us to use the paper resources, Dr. Noriko Kando of NACSIS for her valuable suggestions of statistical analysis of cue words, and anonymous reviewers for their suggestions to improve our paper.

## References

[Biber and Finegan, 1994] Biber, D. and Finegan, E. *section13: Intra-textual variation within medical research articles.* Corpus-Based Research into Language. Oostdijk & de Haan(eds.) Amsterdam, Rodoph. pages 201–221, 1994.

[Garvey, 1979] Garvey, W. D. *Communication, the essence of science.* Oxford : Pergamon Press, 1979.

[Kando, 1997] Kando, N. *Text-level Structure: Implications for Information Retrieval and the Potential for Genre Analysis.* British Computer Society IR SG Annual Colloquium, 1997. (http://www.rd.nacsis.ac.jp/~kando/kando.ps).

[Kita, et al., 1994] Kita, K., Kato, Y.,Omoto, T., and Yano, Y. *A Comparative Study of Automatic Extraction of Collocation from Corpora: Mutual Information vs. Cost Criteria.* Journal of Natural Language Processing, 1(1), pages 21–33, 1994.

[Kupiec et al., 1995] Kupiec, J., Pedersen, J., and Chen, F. *A Trainable Document Summarizer.* In Proc. of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pages 68–73, 1995.

[Liu, 1993] Liu, M. *Progress In Documentation The Complexities of Citation Practice: A Review of Citation Studies.* Journal of Documentation, Vol.49, No.4, pages 370–409, 1993.

[Mani and Bloedorn, 1998] Mani, I. and Bloedorn, E. *Machine Learning of Generic and User-focused Summarization.* In Proc. of the 15th National Conference on Artificial Intelligence, pages 821–826, 1998.

[Narin et al., 1994]
Narin, F., Olivastro, D. and Stevens, K. A. *Bibliometrics/Theory, Practice and Problems.* Evaluation Review, Vol.18, No.1, pages 65–76, 1994.

[Teufel and Moens, 1997] Teufel, S. and Moens, M. *Sentence Extraction as a Classification Task.* Intelligent Scalable Text Summarization Proceeding of a Workshop ACL'97, pages 58–65, 1997.

[van Rijsbergen, 1979] van Rijsbergen. *Information Retrieval(2nd Edition).* Butterworths, London, 1979.

[Weinstock, 1971] Weinstock, N. *Citation indexes, in Kent A. (Ed.).* Encyclopedia of Library and Information Science, New York: Marcel Dekker, Vol.5, pages 16–41, 1971.

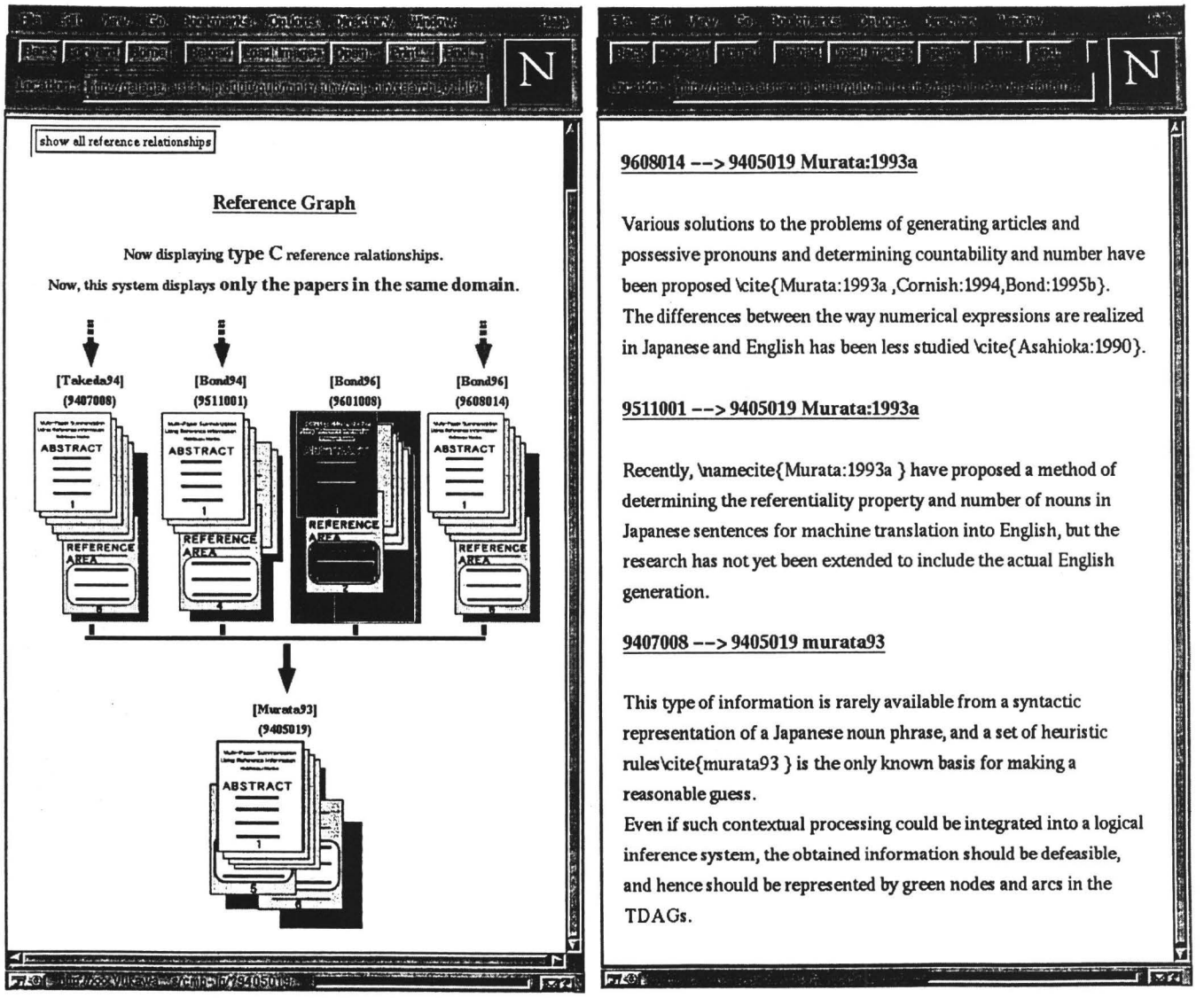


Figure 7: Survey writing support system

[White and Mccain, 1989] White, H. D. and Mccain, K. W. *Bibliometrics.* Annual Review of Information Science and Technology(ARIST), Vol.24, pages 119–186, 1989.

## Papers Used in the Explanation of Reference Information

[Bond, et al., 1996] Bond, F., Ogura, K., and Ikehara, S. *Classifiers in Japanese-to-English Machine Translation.* COLING'96 , pages 125–130, 1996. (http://xxx.lanl.gov/ps/cmp-lg/9608014)

[Bond, et al., 1994] Bond, F., Ogura, K., and Ikehara, S. *Countability and Number in Japanese-to-English Machine Translati on.* COLING'94, pages 32–38, 1994. (http://xxx.lanl.gov/ps/cmp-lg/9511001)

[Murata and Nagao, 1993] Murata, M. and Nagao, M. *Determination of referential property and number of nouns in Japanese sentences for machine translation into English.* TMI-93, 1993. (http://xxx.lanl.gov/ps/cmp-lg/9405019)

[Takeda, 1994] Takeda, K. *Tricolor DAGs for Machine Translation.* In Proceedings of ACL'94, 1994. (http://xxx.lanl.gov/ps/cmp-lg/9407008)