STUDIES ON RELEVANT DOCUMENT RETRIEVAL AND SUMMARIZATION

JANUARY, 2001

DOCTOR OF ENGINEERING

KIYONORI OHTAKE

TOYOHASHI UNIVERSITY OF TECHNOLOGY

関連文書検索とその自動要約に関する研究

（邦文要旨）

本論文の目的は，人間の知的活動を支援し，増強するためのいくつかの技術を発展させることである．そのために，ある文書に関連する文書群を検索するための手法を提案する．次に，関連する複数の新聞記事をまとめて要約する手法を提案する．さらに，より高品質な要約のために必要な語順を考慮した格フレームを自動獲得する手法について検討する．そして，格フレーム獲得において問題となる名詞の多義性を解消する手法を提案する．

計算機およびネットワークの発展によって，膨大な量の情報を容易に得ることができるようになった．その一方で，生物としての人間の情報処理能力はほとんど変化していない．そのため，膨大な情報の中から高速・高精度に必要とする情報を集め（検索），素早くその内容を理解（要約）することが高度情報化社会では要求される．さまざまな情報のなかでも文書は人間の知的活動の基礎となるため極めて重要な要素である．

これまでの多くの情報検索研究は検索質問に基づくシステムを想定している．一方で，検索質問としてキーワードではなく，文書そのものを示し，その文書に関連する文書を検索する手法が考えられる．このような手法はいくつか提案されてきたが，どのような索引語単位を用いるべきかについての研究はほとんどない．そこで，本研究では，索引語の単位として形態素の連接を用いた関連文書検索法を提案し，形態素のみを索引語の単位として用いる手法と比較した．その結果，本研究で提案する関連文書検索の有効性を確認した．

初期の自動要約に関する研究は，単一の文書においていかに重要な文を

選択するかに主眼が置かれたが，1995年以降になって，複数の文書をまとめて要約する場合の問題点も検討されるようになった．本研究では，対象を新聞記事としたとき，その表現上の特殊な構造から，ヒューリスティックスによって十分な要約手法が構築できると考え，手法を考案した．この複数記事要約手法をアンケートによって評価した結果，文章が自然で，適切な要約であることが明らかになり，本手法の有効性を確認した．

また，自動要約に関する研究が進む一方で，より自然で読みやすい要約の生成は依然として困難である．その原因のひとつとして構文解析誤りがある．日本語の構文解析においては，格構造解析が必要であり，そのために格フレームは重要な役割を果たす．しかしながら，既存の格フレーム辞書は人手により収集・整備されているため，量的に不十分である．したがって，格フレームの自動獲得が望まれる．そこで，本研究では，従来重要視されてこなかった語順に着目し，単一言語コーパスから格フレームを自動獲得する手法を検討した．また，コーパスからの格フレーム獲得において問題となる名詞の多義性を解消する手法を検討し，大規模なコーパスを用いた実験により提案手法の有効性を確認した．

Abstract

The goal of this dissertation is to develop methods for supporting and enhancing intellectual activities. For this purpose, firstly, we propose a method for retrieving relevant documents. Secondly, we propose a summarization method for multiple articles. Thirdly, we investigate a method of acquiring case frames from monolingual corpora. And, we present a word-sense disambiguation method required for acquisition of sequences of case elements.

A great deal of information can be acquired nowadays thanks to the recent drastic progress of computer and network technologies. On the other hand, human physical abilities of information processing have been hardly augmented since the dawn of its history. Therefore, efficient and effective retrieval of information which we need from a great deal of information, and its quick understanding of the retrieved information, which may be realized by summarization, are quite important to make critical decisions in this information-oriented society. There are some kinds of information(documents, speech, images and video etc.). Among them, documents are quite important as the basis for intellectual activities.

Most of earlier information retrieval studies suppose query-based systems, and they proposed language-independent methods. On the other hand, a different approach is possible by using the document itself as a query and by searching similar documents for the query. Some methods have been proposed for such retrieval, however, there have been few studies to investigate what unit of index terms should be employed for Japanese documents. Thus, in this dissertation, we propose a method of employing connections of morphemes as the unit of index

terms and compared the method with a conventional method which employs morphemes, not their connections, as a unit. The experimental results showed that the proposed method outperformed the conventional one.

Initial automated summarization researches focused on how to extract important sentences on a single document. However, from the middle of 90's, researchers have begun to tackle the question of multiple documents summarization. By taking advantage of a special structure of newspaper articles, we propose a method for multiple articles summarization. We evaluated the multiple articles summarization method by questionnaires, and the evaluated results showed that almost natural and valid summaries were produced by the proposed method.

However, a summary more natural and more readable is still hard to produce. One of the reasons is errors generated at parsing. In Japanese, case structure analysis is very important to handle several characteristics of Japanese such as scrambling, omission of case components, and disappearance of case markers. Case frames play a very important role in such case structure analysis, yet such case frames are insufficient for practical use because they are manually collected and maintained. Thus, the automated acquisition of case frames are desired. Many of past studies on automated acquisition of case frames supposed multi-lingual corpora as syntactic and semantic disambiguation is needed. However, it is hard to collect a great deal of multi-lingual corpora. Moreover, those past studies have neglected the case order. Hence, we investigate a method for acquiring case frames with case order from mono-lingual corpora. We also investigate a word-sense disambiguation method required to fill each slot of case frames. In order to test the effectiveness of our method, we conducted some experiments. The experimental results indicate that our method is very promising in acquiring sequences of case elements and case frames with case order.

List of Publications

(1) Journal Papers

1. Kiyonori Ohtake, Shigeru Masuyama, Kazuhide Yamamoto: “A Retrieval Method for Relevant Documents Employing Connective Information of Nouns”, Transactions of Information Processing Society of Japan, Vol. 40, No. 5, pp. 2460-2467(1999)(in Japanese). (Corresponding to Chapter 2)

2. Kiyonori Ohtake, Takahiro Funasaka, Shigeru Masuyama, Kazuhide Yamamoto: “Multiple Articles Summarization by Deleting Overlapped and Verbose Parts”, Journal of Natural Language Processing, Vol. 6, No. 6, pp. 45–64(1999)(in Japanese). (Corresponding to Chapter 3)

3. Kiyonori Ohtake, Masahiko Nezu, Shigeru Masuyama, Kazuhide Yamamoto: “Automated Acquisition of Case Frames with Case Order”, The Transactions of the Institute of Electronics, Information and Communication Engineers, Vol. J83-D-II, No. 3, pp. 1060-1063(2000)(in Japanese). (Corresponding to Chapter 4)

(2) International Conference Papers

1. Kiyonori Ohtake, Masahiko Nedu, Shigeru Masuyama and Kazuhide Yamamoto: “Automated Acquisition of Case Frames with Case Order”, Proceedings of NLPRS99, pp. 503-506(1999). (Corresponding to Chapter 4)

(3) Oral Presentations

1. 大竹清敬，山本和英，増山繁：“日本語新聞記事を対象とした関連記事検索の一手法”，情報処理学会第52回全国大会講演論文集 (3), pp. 19-20 (1996).

2. 大竹清敬，山本和英，増山繁：“名詞の連接に着目した日本語新聞記事の関連記事検索手法”，言語処理学会第3回年次大会 発表論文集, pp. 381-384 (1997).

3. 大竹清敬，増山繁，山本和英：“名詞を中心とした連接に着目した新聞の関連記事検索手法”, 情報処理学会 研究報告 97-NL-122 pp.77-82 (1997).

Acknowledgment

The author is heartily grateful to Professor Shigeru Masuyama of Toyohashi University of Technology for his enthusiastic guidance, discussion and persistent encouragement. Without his support, none of this work would have been possible. The author would also like to thank him for his careful reading of this manuscript and accurate comments.

The author is heartily grateful to Invited Researcher Kazuhide Yamamoto of ATR Spoken Language Translation Research Laboratories for his constructive suggestions and continuous encouragement, which enable the author to accomplish this work.

The author sincerely thanks to Professor Sadahiro Isoda, Professor Seiichi Nakagawa, Associate Professor Kyoji Umemura, and Lecturer Takehito Utsuro of Toyohashi University of Technology for their careful reading of the manuscript and valuable comments.

Many thanks are to his friends and colleagues in Professor Masuyama's laboratory. They kindly have discussions on this work, and keep and maintain computer environment comfortable for experiments and writing the manuscript.

Finally, I am thankful to all the people whom I have met in my life, in particular, to my family. To my parents for raising me to be as I am, to my daughter, Haru, for her smile, which encourages me every day and night, and to my wonderful wife, Yoko, for her patience and encouragement.

Contents

Chapter 1

Introduction

1.1 Social and Technological Backgrounds

A great deal of information is available nowadays thanks to the recent drastic progress of computer and network technologies. For example, the estimated number of World Wide Web(WWW) servers ranges from over 27.5 million according to Netcraft Web survey[47] (January 2001) to over 105.7 million according to NetSizer[35](January 2001), and there are a great number of Web pages more than that of servers. We sort out suitable information from these information repositories, and use it efficiently to solve our daily problems. On the other hand, human physical abilities of information processing have been hardly augmented since the appearance of Homo sapiens on the earth. Thus, it is natural for employing computers to support and enhance intellectual activities in this highly developed information-oriented society.

We can acquire a great deal of information, but retrieving required information for us from the great deal of information is becoming hard. Therefore, techniques for information retrieval(IR) from information repositories, in particular, information in the form of documents is quite important.

Under these circumstances, the first TREC(Text REtrieval Conference)[50] conference was held at the National Institute of Standards and Technology(NIST), in Maryland, November 1992. The TREC conference series is co-sponsored by the National Institute of Standards and Technology (NIST) and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) as part of the TIPSTER Text Program. The goal of the conference series is to encourage research in information retrieval from large text applications by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. Also in Japan, the first NTCIR(NII-NACSIS Test Collection for Information Retrieval Systems)[41] Workshop, the first evaluation workshop designed to enhance research in Japanese text retrieval, was held in Tokyo, August 30 - September 1, 1999. And another project, contest-styled, IREX(Information Retrieval and Extraction Exercise)[40] Workshop was held in September, 1999. These conferences show that the importance of large-scale standard test collections in IR research has been widely recognized.

IR with computers has a long history[3, 9, 11]. Also research on natural language processing(NLP) began soon after the advent of digital computers[34]. However, interactions between these two fields have been limited, partly because problems on IR were not seen as interesting in NLP communities, partly because statistical methods, the dominant approach in IR, were not popular in NLP[30] at that time. It is also the reason why interactions between them have been limited

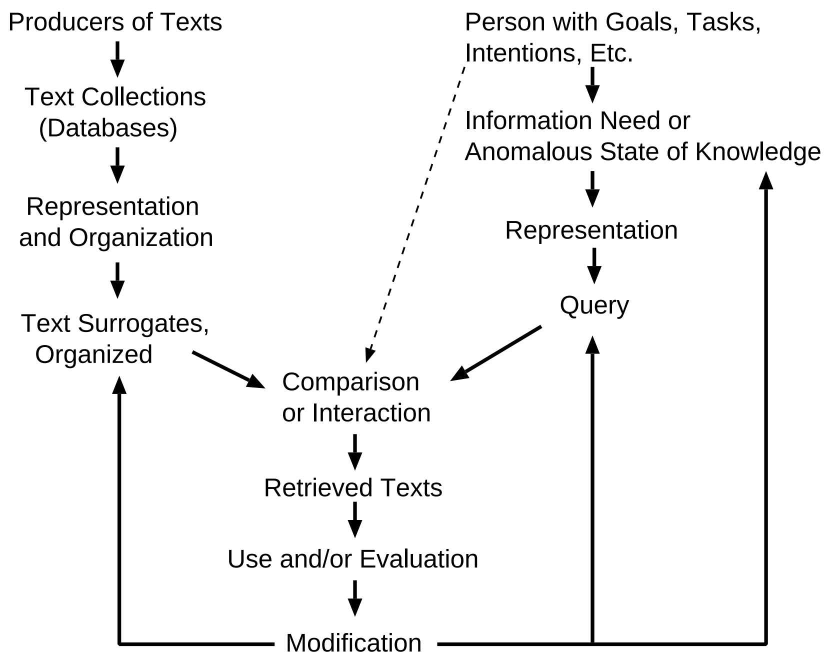

Figure 1.1: A general model of information retrieval[6]

that the NLP techniques were premature for practical use in IR.

The classical and practical IR model is shown in Fig. 1.1[6]. On the other hand, a different approach is possible which uses the document itself as a query and searches similar documents for the query. However, on the classical IR model, the user needs troublesome modification of the query to retrieve such documents. Users frequently encounter this kind of situation at IR process, and they must be worried about modification of the query. Thus a method which retrieves documents relevant to a given document is desired.

In contemporary societies having highly developed computers and networks, documents are machine readable, and more and more information is available and accessible through the Internet. This explosion of information has resulted in the well-recognized information overload problem. People have not enough time to read all documents potentially of interest. Because of these time-consuming reading tasks, the automated summarization is strongly desired. The automated summarization leads us to not only reduction of reading time but also an improvement of IR process by indicating summaries of each document on a retrieval result.

The explosion of on-line textual material and advances in text processing technology have created a renewed interest in text summarization. In May 1998, the U.S. Government conducted an evaluation of automatic text summarization systems, under the TIPSTER Text Program (Phase III). The TIPSTER Text Summarization Evaluation(SUMMAC) represents a significant step in the field of text summarization, as it is the first large-scale, developer-independent evaluation of text summarization systems.

In Japan, there has been a lot of research on automatic text summarization. However, since the evaluations of such systems were done individually with their own evaluation measures at universities and industrial research organizations, and there has been little discussions on the evaluation measures and methods. Thus, it is difficult to compare text summarization systems. In addition, we do not have enough language resources such as human-prepared summaries. Under these circumstances, the Text Summarization Challenge (TSC) has been held from July 20, 2000 to March, 2001 as an NTCIR-2 task in order for the researchers in the field to collect and share text data for summarization, and to clarify the issues of evaluation measures and methods for summarization of Japanese texts.

Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks). There are many uses of summarization in everyday activities, which are indicative of the types of functions that summarization can perform. We are all familiar with summaries such as[29]:

- headlines (from around the world)
- outlines (notes for students)
- minutes (of a meeting)
- previews (of movies)
- synopses (soap opera listings)
- reviews (of a book, CD, movie, etc.)
- digests (TV guide), etc.

In general, humans are extremely capable summarizers. We now turn to summarization by machines.

Research on automatic summarization in a broad sense, i.e., including extracting, abstracting, etc., has a long history[29], with an early burst of effort in the sixties following Luhn's pioneering work[27]. The goal of automated summarization research was to develop methods for extracting important sentences from a document. However, as automated summarization research becomes active, researchers put a great emphasis on more natural summarization[4, 20], on summarizing multiple documents[28, 31], and on how to evaluate summarization method[19, 37, 42]. For Japanese documents summarization, there is a pioneering work, which proposed a practical summarization system for editorials, done by Yamamoto et al.[53]. The feature of the system proposed by Yamamoto et al. is to employ several surface linguistic characteristics appeared in Japanese documents.

To make more natural summarization possible, a semantic analysis is required, yet we are facing the problems of parsing errors. One of the reasons of parsing errors is lack of case frames which play a very important role in syntactic parsing. Thus, the problem motivates the research on automated acquisition of Japanese verbal case frames. We aim to acquire case frames with case order from text corpora, because case order would be useful for summarization to determine important parts or unimportant one.

There is a problem on acquiring case frames from text corpora. The case frame acquisition process consists of two phases: extraction of case frame instances from corpora, and generalization of those instances to case frames. The generalization step is needed in order to represent the input case frame instances more compactly as well as to judge the acceptability of unseen case frame instances. To generalize case frame instances, we assign semantic features to each slot of case frames. When we assign semantic features to nouns, we could not avoid word-sense ambiguities. Thus, word-sense disambiguation method is needed for acquiring case frames.

There are now many computer programs for automatically determining which sense a word is being used in. One would like to be able to say which were better, which worse, and also which words, or varieties of language, presented particular problems to which programs. To this end, an evaluation exercise, SENSEVAL, was organized under the auspices of ACL SIGLEX (the Lexicons Special Interest Group of the Association for Computational Linguistics), EURALEX (European Association for Lexicography), ELSNET, and EU Projects SPARKLE. The first SENSEVAL took place in the summer of 1998, for English, French and Italian, culminating in a workshop held at Herstmonceux Castle, Sussex, England on September 2-4. The SENSEVAL-2 workshop will be held in conjunction with ACL/EACL 2001, in Toulouse, France, probably July 2001.

This dissertation proposes a method of retrieving relevant documents by which users already have on hand to reduce the user's query modification. Moreover, a multiple documents summarization method for Japanese newspaper articles is proposed to support users for grasping the abstract of a series of relevant articles. In addition, this dissertation investigates a method for acquiring case frames with case order and word-sense disambiguation method on acquiring sequences of case elements.

1.2 A Brief Review of Related Researches

The IR research has a long history, and a great number of researchers have tackled various problems on IR[3, 5, 9, 11]. Among them, some methods, vector space[5, 44] and cluster based methods[8, 13, 18], are developed to retrieve relevant documents. In either method, a document is represented by a surrogate expression and would be matched with a surrogate of a query. The surrogate is generally composed of words as the unit in general. However, when we consider

Japanese documents as the target of IR, it is controversial what granularity should be employed as the unit. Thus there have been a few research considering the granularity of the surrogate unit for Japanese texts[39]. In English, comparing the suitability of a word with a noun phrase as the unit of surrogate expression is done by Evans et al.[10]. They reported the noun phrase based IR slightly outperforms in recall and precision. A different approach, employing the passage, a series of words having a certain length, was reported by Callan[7]. The question for employing the passage is how to define the passage, as a method to define the passage is not established for Japanese.

Another approach to retrieve relevant documents employs a parser [2]. Employing a parser on IR may contribute to raise precision, on the other hand it causes the following problems: ambiguity, robustness, and requiring much time to analyze. Thus, we do not employ any parser to realize a method of robustness and to manage a great number of documents.

Research on automatic summarization has a long history from the Luhn's pioneering work[27]. The goal of summarization researches addressed how to extract important sentences. But since the 1990's, the summarization research has become very active and researches not only on extraction of sentences but also on simulation of summarization processes by human, summarization of multiple documents, and generation of a summary from information extracted from documents[43] are reported. Interest in the multiple summarization methods has been raised since the middle of 90's, and many studies have been done[28, 31, 46, 52].

Mani and Bloedorn[28] employed a graph representation whose nodes are term occurrences and whose edges are cohesion relationships (proximity, repetition, synonymy, hypernymy, and coreference) between terms, and utilizes a spreading activation technique to discover, in each document, nodes semantically related to the topic. The activated graphs of each document are then matched to yield a graph corresponding to similarities and differences between the pair, which is rendered in natural language.

Shibata et al.[46] employs frequency of appeared morphemes to determine overlapped sentences, and a summary would be generated from the one of overlapped sentences.

Most of usual researches are sentence selection-based summarization employing statistical methods, even though they aimed at summarization. However, Yamamoto et al.[52] proposed a method of employing heuristics with surface information which eliminates overlapped parts in a sentence. The method proposed by Yamamoto et al. considers an elimination of overlapped parts in the second one of two relevant articles, thus the method is different from ours in that the method does not consider a summary derived from all articles.

The proposed multiple articles summarization method in this dissertation is similar to that proposed by Yamamoto et al. in employing heuristics, but different from that in several points. One of those is: although Yamamoto et al. employs a clause as a unit of comparison and elimination, we employ key expressions appearing frequently in newspapers to determine overlapped parts.

In addition, there are researches employing parsers[32] to generate a summary. A parser, in general, requires much time than a morphological analyzer, and it has also high error rate on an analysis. If we employ a parser to generate a summary, the summary would be unnatural due to such errors although we spent much time to process. We do not employ any parser to generate a summary, because of the processing time and to avoid an unnatural summary caused by parsing errors.

One of reasons for parsing errors is lack of case frame dictionaries with wide-coverage. The case frames play an important role in syntactic analyses for Japanese. Moreover, a large amount of corpora is available nowadays, and researches to acquire automatically case frames from corpora have become popular[22, 33, 36, 51]. The problems on acquisition of case frames are as follows:

1. word-sense ambiguity,

2. Syntactic ambiguity

For example, Utsuro et al.[51] solved these problems by employing syntactically analyzed bilingual corpora. However, such existing syntactically analyzed corpora are too small to learn a dictionary for practical use. Thus, Kawahara et al.[22] employs a robust and accurate parser which does not utilize case frames to cope with the problems.

While a number of studies have been done on automatic acquisition of Japanese verbal case frames[22, 33, 36], there is no consideration on the case orders in the case frames, because changing a case order does not affect the meaning expressed by the case frame.

Although Japanese is believed to be a free word-ordered language, the most general order is as follows: "a time ingredient – a place ingredient – a nominative – a dative – an accusative – a verb[49]", and this fact is confirmed statistically in [17]. However, the investigation[17] of word orders ignored difference of verbs, and did not mention that whether patterns of word orders differ depending on a verb or not. If patterns of word orders differ depending on a verb, word order information for each verb must be useful on detailing a verbal semantic classification.

Reflecting the growing utilization of machine readable texts, a number of corpus-based word-sense disambiguation techniques have recently been proposed[12, 25, 30, 54]. The important points in study of word-sense disambiguation are the following three points[24]:

1. How to model a context.

2. How to define a word-sense.

3. How to realize an unsupervised learning.

Of course, we can consider a supervised method on word-sense disambiguation for acquiring sequences of case elements. However, it is hard to prepare a large amount of answer set.

1.3 Outline of The Dissertation

The purpose of this study is to develop some methods to support and enhance intellectual activities. Consequently, the importance of IR techniques to obtain efficiently and effectively the information which we need is widely recognized. In IR processes, a situation occurs frequently under which the user already has a part of answer documents and requires more documents relevant to that. Thus the method of retrieving relevant documents is important. However, in such a situation, a great number of relevant documents may be retrieved, and it is hard for users to read the all documents. Hence, the summary of outlining such multiple documents is very useful from the perspective of supporting intellectual activities. In addition, the really natural and readable summary is strongly desired. Realization of such summarization requires the robustness on natural language processing(NLP). The case frames, in particular in Japanese, are important for e.g., robust parsing. On the other hand, the case frames available now are not sufficient for actual NLP applications, because most of case frames have been collected and maintained manually. Thus, automatic acquisition of case frames will be important. These perspectives motivate us to study the methods of supporting and enhancing intellectual activities.

Firstly, in Chapter 2, we propose a method for retrieving relevant documents. The proposed method employs a morphological analyzer to deal with a number of documents, and to avoid influence by parsing errors. The features of the proposed method is to employ connections of morphemes in contrast to morphemes used in the conventional methods as the unit of index terms. When we consider Japanese as a target of retrieval, the question is what unit of index terms should be employed. To confirm the effectiveness of the proposed method, we executed experiments for comparing the proposed method with a method which employs just morpheme as the unit of index terms. The results showed that the proposed method outperformed the compared method.

Secondly, in Chapter 3, we propose a method which summarizes multiple documents being relevant to each other. In this chapter, we treat newspaper articles, because newspapers are frequently used as information sources. The method employs key expressions which are peculiar to newspapers to summarize the multiple documents. The proposed method is characterized by employing just heuristics. We confirmed by questionnaires that in summarization of such as newspapers having special structures of texts, heuristics are suitable for producing natural and readable summaries.

Thirdly, in Chapter 4, we go into an acquisition method of case frames which plays an important role in NLP such as parsing. Many of past studies on acquisition of case frames require multi-lingual parallel corpora. However, in this dissertation, we investigated a method which automatically acquires case frames from mono-lingual corpora. The feature of the proposed acquisition method is to consider case orders and approximating appearances of case elements as *bi*-grams. We executed acquiring experiments employing the same semantic feature of the IPAL verb dictionary[48]. From the experimental results, we could conclude that the method is promising, but the results unveiled the necessity of semantic disambiguation for practical use, and the semantic features of the IPAL verb dictionary are insufficient for automatically acquiring case frames.

Thus, in Chapter 5, we propose a method which acquires sequences of case elements, and a method to assign semantic features to nouns are also introduced. Chapter 4 unveiled that the IPAL verb dictionary has not sufficient semantic features for case frames. We have been able to employ the *Goi-Taikei*[15] as an approximate of semantic feature dictionary. However, a noun has some semantic features on the semantic feature dictionary in many cases. We employ statistical information to disambiguate word-sense. We executed acquisition and disambiguation experiments employing newspaper articles for 7 years. We compared the proposed method with base-line method, and confirmed effectiveness of the proposed method.

Finally, we conclude the dissertation in Chapter 6 and suggest some topics for future researches.

Chapter 2

Retrieving Relevant Documents

2.1 Introduction

A great deal of information can be acquired nowadays thanks to the recent drastic progress of computer and network technologies. One of the useful information sources among them is textual documents. Through commercial networks or CD-ROM devices, almost any documents are available.

Moreover, recent development of retrieval methods has enabled us to retrieve documents easily. However, most of those methods are query-based methods, and even if we already have one of answer documents on hand, we have to modify the query to retrieve documents being relevant to the document. Seeking such relevant articles is a time-consuming task due to such query modification. Under these circumstances, an easy to retrieve method of relevant documents are strongly desired.

Vector space methods for Japanese have employed a certain unit for index term, for example morphemes, strings composed of characters of the same description, and so on. However, since Japanese has various compound nouns, we have to cope with these nouns. For example, a compound noun: “自然言語処理 (natural language processing)” and a sentence: “ある言語におけるその処理は自然 (the processing on a language is natural)” shares the three morphemes “自然 (natural)”,“言語 (language)”, “処理 (processing)”, but they convey different meaning. Namely, we have to consider not only just morphemes but also local concatenation and order of morphemes. Moreover, even if we consider the concatenation, ellipsis of morpheme may cause disagreement on matching. Hence, the proposed method in this dissertation employs a connection of morphemes as the unit of index terms to cope with fluctuations caused by ellipsis.

The method proposed here is based on the vector space method to retrieve relevant documents, and it employs connections of morphemes as a unit for index term. To confirm the effectiveness of use of morpheme connections, we carried out comparative experiments with the method[1](Araya’s method), employing morpheme as the unit for index terms. The results showed us the proposed method gave higher value in each recall and precision, therefore we confirmed the effectiveness of employing connection of morpheme for a unit of index terms.

2.2 Method for Retrieving Relevant Documents

This section proposes a method to retrieve relevant documents. The method requires a morphological analyzer to extract connections of morpheme from documents, and a connection is composed of two morphemes.

2.2.1 Connections focused on a noun

We introduce the connections focused on a noun, and they are classified into the following four types:

MN-type A noun comes after an adjective. This type appears relatively rarely, but it would be useful for matching.

> Examples: 具体的な-措置 (concrete - action), 大規模な-援助 (large scale - aid)

NN-type A noun comes after a noun. This is introduced to manage combined nouns(e.g., compound nouns). This type appears frequently.

> Examples: 原子力-機関 (nuclear energy - agency), 大統領-選挙 (presidential - election), 環境-整備 (environmental - considerations)

NV-type A verb comes after a noun. This is introduced to judge if a verb-noun(サ変名詞) is employed as a verb or a noun. The verb corresponds to it's dictionary form(終止形).

> Examples：寄与-する (contribute), 確認-する (confirm), 開催-する (hold)

NP-type A period comes after a noun. This is introduced to cope with a sentence ending up by a noun(Normally, in Japanese, a sentence end up by a declinable word).

2.2.2 Sets of connections

Connections are extracted from morphologically analyzed documents, and its frequency is computed. We call the connections and its frequency a set of connections. A set of connections corresponds to a document.

2.2.3 Heuristics on connections

When connections are extracted, we employ the following heuristics to improve the performance of IR.

1. Each of ' の (*no*)', ',', and ' ・ ' existing between nouns is ignored. Thus two nouns putting one of those morphemes between them are put in a NN-type connection.

2. In three consecutive nouns, "N_1 / N_2 / N_3"(/ denotes morphological boundary), the first and the third nouns are treated as elements of a connection. Namely, we can extract not only N_1 - N_2 and N_2 - N_3 types of connections, but also N_1 - N_3 type connections from the three consecutive nouns.

3. When a parenthesis appears, just like "$M_1/(/M_2/)/M_3$", we extract M_1 - M_3 and M_2 - M_3 type connections. This is because, in other documents the consecutive morphemes, "$M_1/(/M_2/)/M_3$", may be described as M_1/M_3 or M_2/M_3.

4. If a document has a headline, that would be very useful for retrieval of relevant documents. Therefore, if a document has a headline, we would extract

nouns and their frequencies at the headline, and the headline information is distinguished from these from sets of connections.

2.2.4 Evaluation of relevance between documents

This section describes a method for evaluating relevance between documents.

Weighting connections

Each connection in a set of connections is assigned its frequency. However, since having a high frequency does not imply its importance, it is not sufficient to employ the frequency directly as an indicator of importance. Thus many weighting methods based on frequency have been proposed[1, 21, 26, 45]. We employ a general $TF \cdot IDF$ method[3] for weighting connections. The weight of connection c on a document x is defined by the following formula:

$$W(x,c) = \frac{TF(x,c)}{\sum_{c_0 \in SN_x} TF(x,c_0)} \log\left(\frac{M}{af(c)}\right), \tag{2.1}$$

where,

$TF(x,c)$: frequency of c in a document x,

SN_x: a set of connections corresponding to a document x,

M: the number of documents in a repository,

$af(c)$: the number of documents that have a connection c in a repository.

Weighing headline's noun

We already mentioned at Section 2.2.3 that the headline information is distinguished from that from a set of connections. Nouns in the headline are weighted by the following formula:

$$H(x,h) = \frac{TF_h(x,h)}{num_h(x)}, \tag{2.2}$$

where,

$TF_h(x,h)$: frequency of a noun h at the headline of a document x,

$num_h(x)$: a number of nouns in the headline of a document x.

Evaluating Relevance between documents

Given the set of connections and the headline information of Documents x,y, the relevance between x and y is given by the following function R(formula (2.3)). If R exceeds threshold θ, the documents x,y are put in relevance to each other.

$$\begin{aligned} R(x,y) = {} & \frac{\sum_{c_{x\cap y}} W(x, c_{x\cap y}) + \beta \cdot CON(SN_x, SN_y)}{\sum_{c_x} W(x, c_x)} \\ & \cdot \frac{\sum_{c_{x\cap y}} W(y, c_{x\cap y}) + \beta \cdot CON(SN_x, SN_y)}{\sum_{c_y} W(y, c_y)} \\ & + \alpha \left(\sum_{h_{x\cap y}} H(x, h_{x\cap y}) \sum_{h_{x\cap y}} H(y, h_{x\cap y}) \right), \end{aligned} \tag{2.3}$$

where,

x,y: documents,

SN_x, SN_y: sets of connections correspond to x, y, respectively,

c_x, c_y: connections which satisfy $c_x \in SN_x, c_y \in SN_y$,

$c_{x \cap y}$: connections shared by SN_x and SN_y,

$CON(SN_x, SN_y)$: the number of nouns shared by $(SN_x - SN_y)$ and $(SN_y - SN_x)$,

$h_{x \cap y}$: nouns shared by headlines of x and y,

α : a variable which represents an importance of headline information,

β: a variable which represents an importance of $CON(SN_x, SN_y)$.

The first term of the equation (2.3) calculates sharing ratio of connections at each document x, y, and multiplies those ratio. In addition to the product, by $CON(SN_x, SN_y)$ we can cope with not only connections but also morphemes, thus the method includes a conventional method employing just morphemes as an index term. The second term of the equation (2.3) implies the importance of each headline information.

2.3 Experiments

To examine the effectiveness of our proposed method, we executed experiments. The computer system being used is a PC(Dual PentiumII-300, 128MB Memory), and the programming language is *perl*.

We experimented against 28,588 articles from The Nihon Keizai Shimbun, the Japanese daily for business, in 1992 January and February, because newspapers provide useful information, and newspapers are retrieved very frequently.

2.3.1 Making indices

Firstly, we extracted sets of connections[1] from the articles with the morphological analyzer JUMAN.

Two indices are built using the hash method: one of them retains pairs of a connection and its frequency about each document and another is its inverted index[3]. We call an index retaining pairs of a connection and its frequency as the *regular index*. The regular index retains sets of connections each of which corresponds to an article and its headline information. The regular index is searched by an article ID as a key, on the other hand the inverted index is searched by a connection as a key. The size of the regular index constructed is 18.5MB and that of the inverted index is 19.9MB.

2.3.2 Procedure of retrieving relevant articles

A procedure of retrieving relevant articles is as follows.

Input: ID(s) of an article on user's hand.

Output: IDs of Relevant articles.

Step 1 Obtain a set of connections from the regular index.

Step 2 Obtain article IDs from the inverted index with each connection obtained in **Step 1**.

[1] nouns which compose a connection are defined by JUMAN's dictionary, but 形式名詞 (formal noun), 時相名詞 (tense noun), 数詞 (number), and 副詞的名詞 (adverbial noun) are excepted

Step 3 Pick up one ID from article IDs obtained in **Step 2**, and obtain the set of connections corresponding to the ID.

Step 4 Calculate the relevance from the two sets of connections obtained in **Steps 1** and **3**. If the relevance exceeds a threshold θ, the ID picked up in **Step 3** is just relevant article, and output the ID.

Step 5 If there are IDs which have not been calculated in IDs obtained in **Step 2**, then go to **Step 3**, otherwise finish this procedure.

Figure 2.1 shows a system overview. The proposed method is able to accept

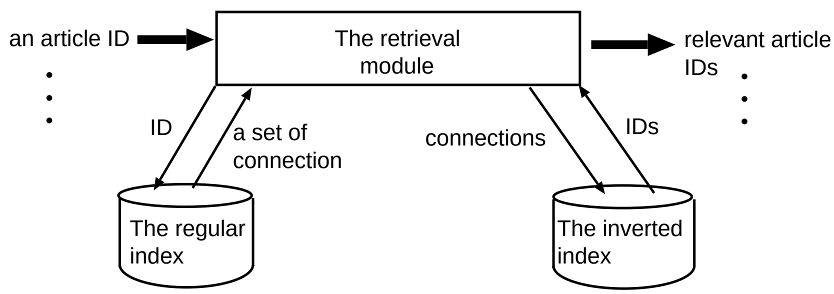

Figure 2.1: The system overview

multiple article IDs as a query, and then the method obtains a union of sets of connections corresponding to each article. When the union is obtained, if the articles share the same connection, the frequency of the connection would be summed up. Finally, the method employs the union as corresponding articles to retrieve relevance articles.

2.3.3 Articles for the experiments

The answer set of relevant articles is manually retrieved within a span of 16 days beginning at a query article appeared in advance. In addition, the variables α, β are defined by a preliminary experiment, and the values are $\alpha = 5$, $\beta = 2$, respectively.

We prepared 10 articles as queries. In addition to the queries, we retrieved relevant articles, answer sets for query articles, of each queries manually. Consequently, we obtained the following 10 groups:

Group A relevance to shipwreck of “TAKA GO”, its cause, and recovery of survivors. (length of the query article: 766, the number of articles within the span: 7,881, the size of the answer set: 29)

Group B relevance to negotiation of two Koreas for denuclearization at Korean Peninsula. (length of the query article: 664, the number of articles within the span: 6,142, the size of the answer set: 26)

Group C relevance to a graft scandal at a company. (length of the query article: 852, the number of articles within the span: 8,584, the size of the answer set: 62)

Group D relevance to the issue of comfort women and the Premier visiting to Korea. (length of the query article: 804, the number of articles within the span: 7,896, the size of the answer set: 31)

Group E relevance to the Summit of Security Council at the United Nations Headquarters (length of the query article: 1,167, the number of articles within the span: 7,721, the size of the answer set: 36)

Group F relevance to arguments on proposals by Provisional Commission for Study on Brain Death, and its answer. (length of the query article: 931, the number of articles within the span: 8,436, the size of the answer set: 44)

Group G relevance to a case of kidnapping of a doctor, releasing the hostage, putting the kidnapper on the wanted list, and arresting the kidnapper. (length of the query article: 1,423, the number of articles within the span: 8,436, the size of the answer set: 37)

Group H relevance to arresting a politician participating a graft scandal at a company. (length of the query article: 1,494, the number of articles within the span: 7,896, the size of the answer set: 77)

Group I relevance to movements of a company and the Russia relating to oil development at off the coast of Sakhalin. (length of the query article: 215, the number of articles within the span: 7,771, the size of the answer set: 28)

Group J relevance to issue on tariffication of rice relating to the Uruguay Round. (length of the query article: 1,471, the number of articles within the span: 7,929, the size of the answer set: 44)

2.3.4 Comparison with a previously proposed method

We compare the proposed method with the Araya's method[1]. The Araya's method weights on morphemes, and evaluate a relevance between documents employing top 25% of weighted morphemes. The proposed method is different from the Araya's one in that it employs connections of morphemes, and we considered the Araya's method is suitable for comparing with our method. The variables required in the Araya's method were decided from [1]. We employed morphological analyzer JUMAN(Version 3.4, the same for ours) for the Araya's method. We have indexed articles by the Araya's method in a manner similar to the proposed method, consequently the size of the regular index constructed is 6.5MB and the inverted index is 9.7MB.

2.4 Functions for Evaluation

We evaluate the effectiveness of retrieval by F-measure for group X on threshold θ. The definition of F-measure is shown by the following formula.

$$F(X,\theta) = \frac{(\gamma^2+1)\times P(X,\theta)\times R(X,\theta)}{\gamma^2\times P(X,\theta)+R(X,\theta)} \quad (\%),$$

where, the variable γ represents a relative importance of recall and precision. If $\gamma = 1$, the importance between recall and precision is equivalent. If $\gamma = 2$, precision is twice more important than recall. In this experiment we let $\gamma = 1$. P and R are the precision and the recall in group X at threshold θ, respectively. P and R are defined as follows.

$$precision : P(X,\theta) = \frac{|R_a|}{|A|}\times 100 \quad (\%),$$

$$recall : R(X,\theta) = \frac{|R_a|}{|R_r|}\times 100 \quad (\%),$$

where, let R_r be a set of relevant articles(the answer set), and A be a set of retrieved articles, and R_a be a set of intersection of the sets R_r and A.

2.4.1 Deciding thresholds

The thresholds on both of the Araya's method and ours have to be fixed to determine relevant articles. We divide the 10 groups into two, for training(A, B, C, D, E) and for test(F, G, H, I, J), to decide the threshold. We calculate $F(X,\theta)$ by changing θ on the training group, and if a θ maximizes the average of $F(X,\theta)$, then the threshold is set to be θ.

Firstly, Table 2.1 shows the results by the Araya's method on the training group. We experimented by changing the value of θ from 0.01 to 0.13 for every 0.01. As a result $\theta = 0.02$ maximized the average of $F(X,\theta)$.

Table 2.1: Retrieval result by the Araya's method(for training group)

Group	precision	recall	F-measure
A	78.6%(22/28)	75.9%(22/29)	77.2%
B	66.7%(22/33)	84.6%(22/26)	74.6%
C	97.4%(37/38)	59.7%(37/62)	74.0%
D	78.1%(25/32)	80.6%(25/31)	79.4%
E	43.5%(30/69)	83.3%(30/36)	57.1%
average	72.8%	76.8%	72.5%

Secondly, we experimented on our method, by changing θ from 0.1 to 1.5 for every 0.1. As a result $\theta = 0.5$ maximized the average of $F(X,\theta)$. Table 2.2 shows the results.

Similarly, we also experimented our method on $\alpha = 0$(do not employ any headline information) against the training group, then $\theta = 0.5$ maximized the av-

Table 2.2: Retrieval result by our method（for training group）

Group	precision	recall	F-measure
A	70.7%(29/41)	100.0%(29/29)	82.9%
B	80.0%(24/30)	92.3%(24/26)	85.7%
C	87.7%(57/65)	91.9%(57/62)	89.8%
D	57.4%(31/54)	100%(31/31)	72.9%
E	64.6%(31/48)	86.1%(31/36)	73.8%
average	72.1%	94.1%	81.0%

erage of $F(X,\theta)$ and the average is 78.7%.

2.4.2 Experimental results

We experimented against the test group. Table 2.3 shows results by the Araya's method.

Table 2.4 shows the result of our method($\alpha = 5$) against the test group.

Table 2.5 shows the result of our method($\alpha = 0$) against the test group.

In addition, we experimented to examine the behavior of the two method by shifting the threshold against the test group. Figure 2.2 shows the result of two methods by shifting the threshold against the test group.

Figure 2.3 illustrates a relation between threshold and F-measure by the proposed method($\alpha = 5$) against the test group.

We also measured an average retrieval time on the Araya's method and ours against all the groups, the results are the Araya's: 0.44 second and ours: 4.96

Table 2.3: Retrieval results by the Araya's method(for test group)

Group	precision	recall	F-measure
F	91.5%(43/47)	97.7%(43/44)	94.5%
G	96.3%(26/27)	68.4%(26/38)	80.0%
H	88.6%(31/35)	40.3%(31/77)	55.4%
I	76.9%(20/26)	71.4%(20/28)	74.1%
J	82.1%(32/39)	72.7%(32/44)	77.1%
average	87.1%	70.1%	76.2%

seconds. These results were obtained by repeating a measure 10 times against each of ten groups and got the average.

2.5 Discussions

2.5.1 Properties of proposed methods

From Tables 2.3, 2.4 and Figure 2.2, we can conclude that the proposed method outperforms the Araya's method for both cases when headline information is employed and is not employed. These results imply that employing connections of morphemes makes relevant articles retrieval more effective than just employing morphemes for index terms.

We can also point out the effectiveness of employing headline information from differences on the results between $\alpha = 0$ and 5. However, the contribution of headline information is not so large, because the average F-measures are 85.9% at

Table 2.4: Retrieval precision by our method(for test group, $\alpha = 5$)

Group	precision	recall	F-measure
F	97.5%(39/40)	88.6%(39/44)	92.9%
G	90.6%(29/32)	76.3%(29/38)	82.9%
H	88.2%(60/68)	77.9%(60/77)	82.8%
I	84.8%(28/33)	100.0%(28/28)	91.8%
J	82.9%(34/41)	77.2%(34/44)	80.0%
average	88.8%	84.0%	85.9%

$\alpha = 5$ and 83.5% at $\alpha = 0$.

Figure 2.3 shows a stability of the proposed method at the threshold from 0.4 to 0.8, and in this range, both of recall and precision are more than 80%.

2.5.2 Analyses of the results

In group D, Araya's method gives a high F-measure than ours for the reason that the precision with the proposed method is low. The low precision is caused by less important compound nouns like "官房長官 (Chief Cabinet Secretary)", "首脳会談 (meeting at the summit)", and so on. Other instance such as "第三管区海上保安本部 (the third Regional Maritime Safety Headquarters)" is a very long compound noun, and these compound nouns are overestimated at evaluating relevance due to too many connections composed by introduced heuristics in the proposed method.

If the heuristics work as we expected, the recall would be higher just like represented by group H. The recall by our method is higher than the Araya's

Table 2.5: Retrieval precision by our method(for test group, $\alpha = 0$)

Group	precision	recall	F-measure
F	100.0%(37/37)	84.1%(37/44)	91.4%
G	93.1%(27/29)	71.1%(27/38)	80.6%
H	89.8%(44/49)	57.1%(44/77)	69.8%
I	96.4%(27/28)	96.4%(27/28)	96.4%
J	81.0%(34/42)	77.3%(34/44)	79.1%
average	92.1%	77.2%	83.5%

method from Tables 2.4 and 2.5, because of the heuristic for headline information. In addition, the high recall is due to the heuristic for three consecutive nouns, such as “〇〇□□元長官” matched with “〇〇元長官” and these words are very important at each article in general.

An improvement of precision at the proposed method is expected by employing connections of morpheme as the unit of index terms in the vector space method. However, it was observed that the precision is not so high because of the heuristics which absorb fluctuations of descriptions to improve recall.

2.5.3 Retrieval time

There are two reasons for the large differences in retrieval time between the Araya’s method and ours. Firstly, the proposed method has to deal with much information than the Araya’s one. Secondly, the Araya’s method gives a weight to each morpheme in advance, on the other hand the proposed method gives a weight to

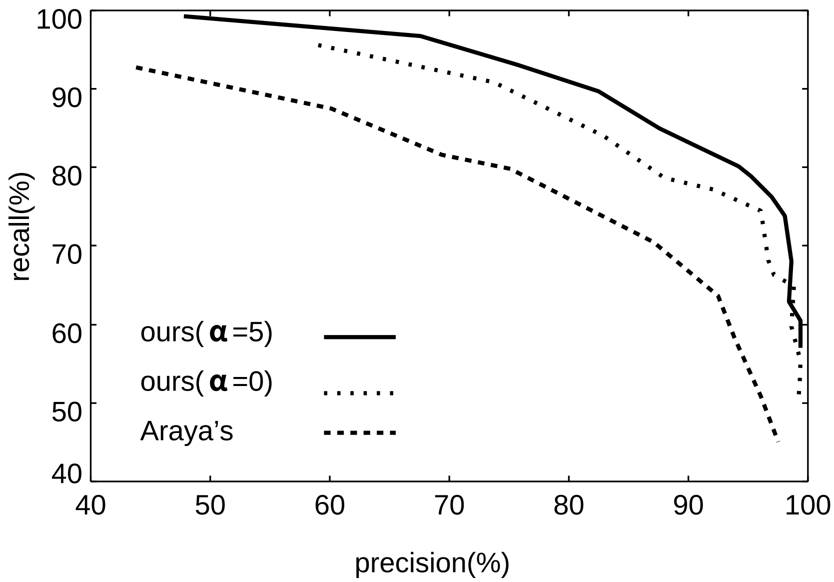

Figure 2.2: Relation between precision and recall average

each connection of morphemes at run-time to make possible multiple articles as a query. However, the drastic enhancement of computing powers nowadays is saving us from bothering such running time overhead, thus we consider the retrieval time would be practical.

2.6 Concluding Remarks

In this chapter, a method of retrieving relevant documents employing connections of morphemes as the unit of index terms for the vector space method is proposed. Flexibility in coping with compound nouns gave the proposed method a distinction. We experimented the proposed method with a prototype system on a

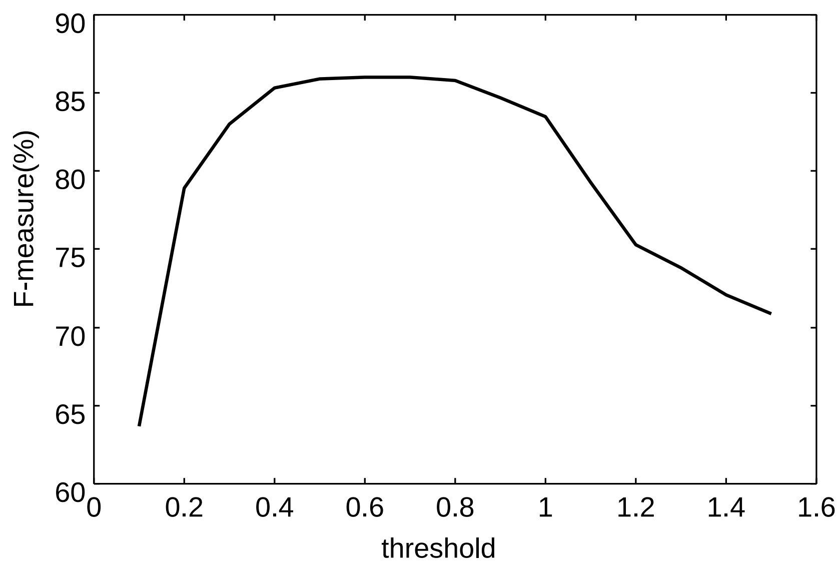

Figure 2.3: Relation between threshold value and F-measure

computer against newspaper articles, and we also compared the proposed method with the Araya's[1] one which gave weighting just only morphemes. The experimental results showed that the proposed method which employs connections of morphemes outperforms the conventional methods which employ weighting only on morphemes.

Chapter 3

Summarization Method for Multiple Japanese Newspaper Articles

3.1 Introduction

In this chapter, we propose a multiple summarization method for Japanese newspaper articles. With the rapid growth of World Wide Web and online information services, more and more information is available and accessible through networks. One of them, machine-readable newspaper articles, enable us to retrieve articles which we need very easily. On the other hand, when we retrieve relevant articles on a long term event, the results become enormous. We do not have sufficient time to read everything, and yet we have to make critical decisions based on whatever information is available. Hence, the multiple summarization method for these articles is indispensable. This chapter aims at realizing a summarization for multiple newspaper articles which are mutually relevant.

Although research on automatic summarization has a long history and a number of researches have been reported, most of them are customized for a single document. On multiple summarization of articles on the same event, there are similar sentences in different articles, and then applying single document summarization to each article would produce a verbose and redundant summary. Multiple-documents summarization has to cope with this problem.

Japanese newspaper articles have special structures[14]. A series of headlines about the same event gives us an outline of the event. When we want to know about the event, we have to read a body of each article, but thanks to the special structures that the first paragraph is mostly a summary of the article. Thus, an extract obtained by extracting the first paragraph from each article might be a good summary of the event. However, each article is written based on the assumption that an article is read as a single one, therefore the series of the first paragraph of relevant newspaper articles includes overlapped parts. Thus an extract obtained by extracting the first paragraph may be verbose and may hard to read. To avoid this problem, the method for multiple summarization has to summarize the articles detecting overlapped parts and deleting it to be more readable summary.

The method proposed here produces a summary by deleting unimportant parts inferred from the whole relevant articles. Unimportant parts are classified into the following two parts:

verbose part: unimportant parts in a single article.

overlapped part: overlapped parts between articles.

Note that the usual summarization method for a single document deletes the verbose parts, and the overlapped parts must be considered for multiple documents summarization.

The summary produced by the summarization method proposed has to satisfy the following requirements:

- summaries of each article are putted in order of time
- including no verbose part in a single article
- including no overlapped part in whole articles
- being able to understand an outline of the event by reading the summary
- summaries of each article have more information than each headline

The proposed method produces a summary from the first paragraphs of articles putted in order of time, and summarize the first paragraph in each article. Hence the summary produced has more detailed information than the series of headlines, and it is shorter than that obtained by compiling first paragraphs. Hence, we can grasp an outline being relevant to the event from the summary.

Appendix A shows us an example of a summary. That summary was generated by the method proposed in this dissertation. That example includes overlapped parts, which were deleted. Deleting overlapped parts is appropriate, as long as detecting the part is correct, because it is obvious that overlapped parts are known and unimportant. Actually, the example of summary evaluated about deleting all overlapped parts was suitable.

Detecting verbose parts requires a guiding importance, and the importance varied by points of view on summary, by necessary compression ratio, and so on.

Thus evaluating verbose parts detection may vacillate. This problem has been argued in evaluation on summarization for single document so far. Therefore, the example of summarization in Appendix A is proper about deleting overlapped parts, but the example is not proper about detecting verbose parts. However, the example, actually, is a satisfactory summary to grasp the outline of the event. In evaluation of the example summary, no invalid deletion was pointed out, but some parts were verbose.

In retrieving newspaper articles, it is natural that users require a summary of a series of relevant articles frequently as the number of relevant articles may become large. If the summary is provided, users would be able to know an outline of the relevant articles, and the summary gives users a useful information for next IR process. In addition, when a series of headlines in order of time has a little information for users but a series of the first paragraphs of the relevant articles has a large amount of text, the summary could provide a proper information with proper amount of text.

There is a possibility that the verbose parts are included in any article, but identifying the overlapped parts would be hard depending on how the articles were written. However, if an overlapped part exists, deleting the overlapped part is proper from the view point of multiple summarization. We can expect that an event or a big accident causes a long-term flow of newspaper articles, and there are many overlapped parts in such articles. There is a big demand for multiple-documents summarization for these articles, thus the summarization by deleting verbose and overlapped parts are of practical use. Actually, the example A exhibits efficiency argued above, and the example can be said to be a summary of purpose.

In this dissertation, we propose a method which employs heuristics for the process we mentioned above. We implemented the method on a computer, and confirmed an effectiveness of the method through an experimental evaluation.

3.2 The Proposed Summarization Method

The feature of the proposed multiple articles summarization method is that it consists of heuristics with morphologically analyzed articles. We believe that heuristics are sufficient for detecting overlapped and verbose parts in articles which are relevant to each other, and the summary by the heuristics is sufficiently readable for grasping the outline of the articles. Thus the proposed method does not employ conventional statistical methods for summarization. However, if we combine the proposed method with conventional methods, then more shorter summary would be generated.

The outline of the proposed summarization method is below:

Firstly, the first paragraphs of each article are arranged in time order. The series of the first paragraphs enables users to grasp easily the outline about the event.

Secondly, we detect guess sentences, which is verbose parts, by expressions at the end of sentence on each article and eliminate them. Moreover, the detailed expression of addresses appears frequently in newspapers, and we eliminate these expressions because they are not important.

Thirdly, we define an introduction part to detect overlapped parts. If nouns and verbs in an introduction part are included former articles, the introduction part has old information, hence the introduction part should be eliminated.

Finally, we examine overlaps of proper nouns, rewording by parenthesis, and sentences within an article and between articles. If there are overlapped parts through all articles, we would remain one part and others would be eliminated.

Through the process as mentioned above, the final results become a summary.

3.2.1 Prerequisites for summarization

The method generates a summary from morphologically analyzed articles, thus each article has to be analyzed in advance. We assume that the proposed method would be applied to articles derived from results of a retrieval of relevant documents and they have fully relevance to each other.

As generating a summarization, in general, there is a question what order of articles should be generated. The method outputs articles in time order, because time order of articles coincides with its description order, and the order assists users to understand the outline of the event.

The proposed method determines overlapped and verbose parts, and eliminates them against the first paragraphs of each article. Thus the method does not adjust the compression ratio.

3.2.2 Outline of summarization procedures

The proposed method is composed of a processing of verbose parts and a processing of overlapped parts. The processing of verbose parts investigates guess sentences and expressions of address. The processing of overlapped parts investigates introduction parts, parentheses, proper nouns, and overlapped sentences. The outline of the proposed summarization procedure is as follows:

1. Each article is morphologically analyzed.
2. Extract the first paragraph of each article, and the paragraphs are arranged in time order of articles.
3. Guess sentence are processed.
4. Overlapped sentences are processed.
5. Overlapped expressions of address are processed.
6. Overlapped proper nouns are processed.
7. Parenthesis are processed.
8. Introduction parts are processed.
9. Miscellaneous processing are done.
10. Output the summary.

Each processing is described in detail in the following pages.

3.2.3 Processing for guess sentences

A guess sentence is defined as a sentence which indicates a possibility of disagreement between the description and the reality. Thus guess sentences do not have much importance in newspaper articles. Firstly, to detect guess sentences, each of the end of sentences is examined whether the end matches with expressions shown in Table 3.1. If an end of a sentence matches with one of them, then the sentence would be taken as a provisional guess sentence. The expressions shown in Table 3.1 are collected from The Nihon Keizai Shimbun from 1990 to 1992. Secondly, the provisional guess sentences are eliminated in case when all the following conditions are satisfied.

1. Including no expression which indicates a basis such as “〜によると (...*niyoruto*)”,“〜ため (...*tame*)”.
2. Including no adversative conjunction or conjunctive particle such as “だが (*daga*)”,“しかし (*sikasi*)”, “が (*ga*)”.
3. Including no expression which indicates a condition or concession such as “ただ (*tada*)”, “ものの (*monono*)”.

3.2.4 Processing of overlapped sentences

In newspapers, we encounter almost the same sentences very rarely. These sentences tend to appear as the time passing from the advent of the first article on the event is long. An example is shown as follows: (The **boldface parts** indicate the almost the same sentences.)

> 米国が日本と共同開発中のＦＳＸの関連技術を入手できると定めた政府間合意に基づき，技術に付随する「試供品」としてハードウエアを輸出する．**日本企業が独自開発した本格的な軍用機**

Table 3.1: Expressions of the end of sentences at guess sentences

～可能性も出てきた (There is also a possibility of ...)	
～可能性もある (There is a possibility of ...)	
～可能性が大きい (to have strong possibility)	
～かもしれない (it may possibly be ...)	～情勢だ (being a situation)
～微妙 (subtle things)	～という (said to be ...)
～そうだ (show an inclination to ...)	～ちがいない (it must be ...)
～と思われる (be believed to ...)	～はずだ (it must be ...)
～見られている (being believed to ...)	～微妙だ (be subtle)
～見通しである (be expected to ...)	～微妙である (be subtle)
～見通しだ (be expected to ...)	～見通し (a prospect of ...)
～みられる (it seems ...)	～模様だ (look like ...)
～よう (it seems ...)	～ようだ (it seems ...)
～予想される (be expected to ...)	～らしい (it seems that ...)
～ろう (be expected to ...)	～そう (may ...)
～そうもない (not seem to ...)	

材を対米供与する初の事例になる．米空軍は将来の戦闘機開発で同レーダーの採用を検討するとみられ，同社は引き合いがあれば積極的に対応する方針だ．(Based on the inter-governmental agreement which decides that the U.S. is able to obtain the FSX technology jointly developed by Japan, a Japanese company will export a hardware as a “sample” accompanied with the technology. **This will be**

the first case of giving to U.S. full-scale military equipments originally developed by a Japanese company. The U.S. Air force seems to review in adoption of the radar at developing fighter airplane for the future, and the Japanese company will aggressively respond in receiving some inquiries.) (The Nihon Keizai Shimbun, January 26th, 1993)

米国が日米共同開発中のＦＳＸの関連武器技術を入手できると定めた政府間合意に基づき，技術に付随する参考品の形で供与した．**日本企業が独自開発した本格的な軍用機材を対米供与する初の事例となる．** (Based on the inter-governmental agreement which decides that the U.S. is able to obtain the FSX technology jointly developed by Japan, a Japanese company gave the technology in the shape of a sample accompanied with the technology. **This will be the first case of giving to U.S. full-scale military equipments originally developed by a Japanese company.**) (The Nihon Keizai Shimbun, August 4th, 1993)

These two articles have two sentences which share almost the same contents, when we summarize two articles, one of the sentences should be eliminated. There are some detection methods for these overlapped parts.

In this proposed method, detection and elimination for such overlapped sentences are described below. Given sentences $S1$ and $S2$, and let $n(S1)$ be the number of nouns in $S1$, and $n(S2)$ be that in $S2$, respectively, and let $m(S1,S2)$ be the number of nouns shared by $S1$ and $S2$. When the conditions

$$\frac{m(S1,S2)}{n(S1)} > \alpha, \quad \frac{m(S1,S2)}{n(S2)} > \alpha$$

are satisfied, $S2$ is eliminated.

In the proposed method, let the variable α be 0.8. The reason why α is set to be 0.8 is described below.

If the variable α is small, sentences would be easily eliminated. We observed invalid elimination caused by the small α in a preliminary examination, thus we let $\alpha = 0.8$ with the result so that the process eliminates almost the same sentence. It is possible that $S2$ has the same contents as that of $S1$, in other words $S1$ is subsumed by $S2$, and in this situation it may be natural to eliminate $S1$. However, in that situation, we observed that the elimination of $S1$ leads us to produce an unnatural article. Thus we defined the process as above simply. As we applied the process to the example shown above, the sentence: "日本企業が独自開発した本格的な軍用機材を対米供与する初の事例となる．(This will be the first case of giving to U.S. full-scale military equipments originally developed by a Japanese company)" in the second article is eliminated.

3.2.5 Processing of address expression

Detailed expressions of addresses appear frequently in newspapers. For example,

> 七千万円の債権を放棄させていた**北海道○○町△町一二三**，廃品回収業 (a junk dealer, **123 △ ○○-cho, Hokkaido,** have made someone abandon bonds of 70 million yen) (The Nihon Keizai Shimbun, February 13th, 1992)

東京都江戸川区西葛西七丁目の都道交差点で (at an intersection in **7 cho-me Nishi-Kasai Edogawa, Tokyo**) (The Nihon Keizai Shimbun, January 26th, 1990)

are shown(where the **boldface parts** indicate expressions of addresses). These expressions are detected by pattern matching and are eliminated. If the morpheme next to the expression of an address is a comma, the expression of an address and the comma would be deleted. If the morpheme next to that is 'の (*no*)', the part from the head of the expression to any one of "都 (*to*), 道 (*dou*), 府 (*fu*), 県 (*ken*), 管内 (*kan-nai*), 市 (*si*)" would be left, but the rest of the expression is eliminated. As this process is applied to the examples shown as above, the results are as follows:

七千万円の債権を放棄させていた廃品回収業 (a junk dealer have made someone abandon bonds of 70 million yen)

東京都の都道交差点で (at an intersection in Tokyo)

3.2.6 Processing of proper nouns

In newspaper articles, there are some parts explaining a proper noun in front and behind of that. The examples are as follows(where the **boldface parts** indicate such explaining parts):

前道議の○○○○容疑者**（５６）＝渡島管内△△町△△△＝** (a suspected ○○○○, **a pre-member of HOKKAIDO assembly, 56 years old, △ △△-cho, Oshima**) (The Nihon Keizai Shimbun, February 20th, 1992)

早大三年，○○○○さん**（２０）**ら四人は (four people including ○○○○, **third-year student of SODAI(Waseda University), 20 years old**) (The Nihon Keizai Shimbun, April 9th, 1991)

These parts appear frequently in a series of relevant articles. If a proper noun appears more than once within the whole articles, and they have such explaining parts, then those parts would be verbose and should be eliminated.

The explaining parts are adnominal parts for the proper nouns, an expression of age which is put into a parenthesis after the noun, and expressions of address which are put between '＝'. The proposed method eliminates these parts.

An adnominal part for a person's name is detected by tracing morphemes ahead from the person's name as long as the morpheme is one of a noun, particle 'の (*no*)', and a comma which exists at the next to a noun. An adnominal part for a place name is detected by tracing morphemes ahead from the place name as long as the morpheme is a noun. If the form "a noun - の (*no*) - a place name" appears, the method retains the noun, and if this form appears more than once and the nouns is the same, the part of the from " a noun - の (*no*)" is deleted. An example is shown below.

施行主体の広島市は (Hiroshima-city **as an executive entity**)(The Nihon Keizai Shimbun March 15th, 1991)

3.2.7 Processing of parenthesis

The processing of parenthesis in the proposed method deals with the following pattern:

A(B).

This pattern may be replaced with one of the following patterns.

1. B
2. A
3. A(B)

Pattern 1. is a case of word changing, and the word before the parenthesis is longer, this word changing occurs more frequently. For example, the pattern "石油輸出国機構（ＯＰＥＣ）" (Organization of Petroleum Exporting Countries(OPEC)) appears more than once in a series of relevant articles. In a case similar to this, the word existing before the parenthesis is verbose, and it is possible that the pattern is replaced as follows.

石油輸出国機構（ＯＰＥＣ） → ＯＰＥＣ

The conditions for applying this replacement is as follows:

- The word in the parenthesis is shorter than the word existing in front of the parenthesis, and
- the word in the parenthesis consists of one morpheme.

If these two conditions are satisfied, the replacement from A(B) to B would be done.

In the case of pattern 2., the words in parenthesis express an additional information of the word existing in front of the parenthesis. For instance, the pattern

“ダイエー（本社神戸市）(DAIEI(Kobe-based))” in a series of relevant articles appears more than once, the information conveyed by the words in the parenthesis is verbose. Thus, in these cases, the form is replaced as follows.

ダイエー（本社神戸市） → ダイエー

The conditions for applying this replacement is as follows:

- The words in the parenthesis are completely matched with the already appeared words in a parenthesis, or
- The words in the parenthesis is included partly as sequential strings in already appeared words in parenthesis.

If these two conditions are satisfied, the replacement from A(B) to A would be done.

Pattern 3. is a case of disagreement with conditions of patterns 1. and 2., and in this case, the method does nothing. This pattern 3. corresponds to a case where there is no relation between A and B. For example, the following pattern corresponds to this case.

した.（関連記事１面に）(...(related articles are on the 1st page))

There are other parenthesis patterns which are verbose in summaries and those are different from that mentioned above. The proposed method deletes these verbose parenthesis patterns such as the author’s information at the beginning of an article and relevant articles information for reference at the last of an article. Examples are shown as follows:

【パリ７日＝○○△△】（【at Paris, 7th, ○○△△】）

（関連記事１面に） ((related articles are on the 1st page))

3.2.8 Processing of introduction parts

In summarization of relevant articles, there are descriptions for premises, development of an event, and so forth, and there is a case when these descriptions overlap with each other, as each article is assumed to read by itself. These descriptions are verbose for multiple-articles summarization, thus the method detects it and eliminates.

In a series of articles in time order $A_1, A_2, \cdots, A_n$, there are descriptions already described in an old article. These descriptions generally appear in the beginning of an article, and we defined these parts as introduction parts. The definition of introduction parts is as follows:

1. The introduction part exists on the first sentence of an article.

2. The introduction part is defined as a string from the head of sentence to one of the following expressions[1]:

   〜したが (..., but), 〜問題で (on the issue/matter of ...),

   〜事件で (in the case of ...), 〜事故で (in the accident of ...),

   〜していたが (have been ...ing, but), 〜について (about ...)

[1]These expressions are manually collected by consulting The Nihon Keizai Shimbun from 1990 to 1992.

or to the front of “a noun - は (*ha*)”[2]. If there are more than one expressions matched with the above expressions, the shortest string would be taken.

3. If there is no part corresponding to the above condition, the article has no introduction part.

In a series of articles in time order $A_1, A_2, \cdots, A_n$, there is a case that the introduction part of articles $A_i (i \geq 2)$ is verbose. This is because, in articles $A_1, A_2, \cdots, A_{i-1}$, the introduction part of A_i may have already been described.

The conditions to decide whether an introduction part is overlapped are as follows:

- When the introduction part of A_i includes “〜事件で (at the event)” or “〜事故で (at the accident)”, 30% or more of the nouns[3] and the verbs of the introduction part are included in a sentence which exists on one of articles $A_1, A_2, \cdots, A_{i-1}$.

- Otherwise, at least 60% of the nouns and the verbs of the introduction part are included in a sentence which exists on one of articles $A_1, A_2, \cdots, A_{i-1}$.

If an introduction part satisfies the conditions shown above, the introduction part would be eliminated, where the rates(30%, 60%) are defined by preliminary examination.

[2]Here, the noun is detected by tracing morphemes ahead from the front of morpheme ‘は’ as long as the morpheme is one of a noun, ‘・’, a prefix, a suffix, and ‘と’ which follows a noun.

[3]In these conditions, the morphemes which are put in parenthesis and numbers are not considered as nouns.

Let $A_1, A_2, \cdots, A_n$ are a series of relevant articles in time order, then the processing of introduction part is shown as follows:

Input：Articles $A_1, A_2, \cdots, A_n$

Output：Articles $A_1, A_2, \cdots, A_n$

Step 1 $i := n$

Step 2 If i is larger than 1, then reiterate **Steps 2.1** to **2.4**.

Step 2.1 Detecting the introduction part on the article A_i

Step 2.2 If there is no introduction part, then goto **Step 2.4**.

Step 2.3 The introduction part of A_i is compared with each sentence of $A_1, A_2, \cdots, A_{i-1}$. If the conditions to decide whether the introduction part is overlapped are satisfied, the introduction part is deleted.

Step 2.4 $i := i - 1$, goto **Step 2**.

Step 3 Finish the procedure.

3.2.9 Miscellaneous processing

Here, the proposed method copes with miscellaneous phenomena as follows:

1. If there is a description such as “解説...面に (the explanation is on ... page)”, that would be eliminated.

2. Each sentence is examined from the beginning if a description ‘同日’(the same day) appears although there is no description of ‘a number + 日’ before

‘同日’. This situation is caused by elimination of a part which includes the ‘a number + 日’ and that corresponds to ‘同日’. Thus we detect the ‘a number + 日’ from before ‘同日’, and replace the ‘同日’ with ‘a number + 日’.

There are other ellipsis resolutions, but the proposed method copes with ellipsis of date information.

3.3 Evaluation by Experiments

We executed experiments on the proposed method, and evaluated by obtaining information by means of questionnaires. The computer system being used is a PC(PentiumII 300MHz, 128MB Memory), and the programming language is perl. We employed JUMAN3.5 as a morphological analyzer with no error correction.

We utilized The Nihon Keizai Shimbun in 1990 and 1992 for the experiments. We extracted 27 groups of relevant articles from the newspaper in advance. The average number of articles by which each group is composed of is 4.7, the maximum of that is 9 and the minimum of that is 3. The extracted articles are different from the articles that are referred to by which this method is composed of.

3.3.1 Experimental results

We summarized all 27 groups with the proposed method. As a result the average time to summarize is 0.8 second, and the average compression ratio is 82.1%.

3.3.2 Evaluation method

Evaluation problems on natural language systems has a long history particularly in fields of machine translation systems[23]. In fields of text summarization for a single document, researchers have compared produced summaries with manually generated summaries, and have computed the precision and the recall to evaluate[37].

It is possible that the texts manually summarized are various summaries due to such as the perspective of the summarizer. Namely, we cannot uniquely determine the best summary for a text. Thus it seems natural that an evaluation of automated summarization is done by some people. Actually, Yamamoto et al. evaluated their method with questionnaires to 18 people[53]. The questionnaires by Yamamoto et al. asked subjects about spontaneity, content suitability, and elimination of modifiers of the summaries automatically generated. However, these kinds of questionnaires are not able to unveil what element of the method works efficiently. Thus, in this dissertation, we obtained information by means of questionnaires by comparing the summary and source articles, and in which the subjects freely indicate the following points: (1) the parts which should be deleted, (2) the parts which should not be deleted.

3.3.3 Investigation by questionnaires

We chose the 21 groups, which have a compression ratio of not greater than 90%, from the all 27 groups. And, we chose 6 groups for the evaluation from the 21 groups. The outlines of these groups are shown in Appendix B. The average compression ratio of the six groups is 74.5%, minimum of that is 56.0%, and max of that is 83.1%.

The subjects are 11 students. We explained to them that the outline of the proposed method and the source articles consist of their first paragraphs, and the following assumptions imposed on the summary.

- The proposed method is applied to a result of retrieval of relevant articles.
- The user can eventually determine a group of the articles based on their headline to summarize.
- The user knows an outline of the articles, and he uses this method to get more precise information on those articles.
- The aim of the summarization system is to produce a summary which is more precise than that consisting of a series of their headline, but which is shorter than a series of their first paragraph.

It is also explained that the proposed method determines the overlapped and verbose parts and eliminates them.

It was shown for the subjects that the date, headline, and body(source and summary) of each article. The subjects freely pointed out the following two points:

1. the parts which should be deleted,
2. the parts which should not be deleted.

The evaluations were done with no limitation in time.

3.3.4 Results of the investigation

The indicated parts were counted by exact matching. For example, against the string C_1 C_2 C_3 C_4 C_5, a subject A pointed out from C_2 to C_4, and a subject B pointed out from C_1 to C_5, then the string from C_2 to C_4 is shared, but the strings are counted separately.

Firstly, we itemize the process of the proposed method applied for the six groups employed for the questionnaires, where the number in parenthesis shows a component percentage.

- Processing of guess sentences: 4(6.9%).
- Processing of expressions of address: 15(25.9%).
- Processing of proper nouns: 11 (19.0%).
- Processing of parenthesis: 9(15.5%).
- Processing of introduction parts: 19(32.8%)

Total 58 places were eliminated.

Firstly, the places, which should be deleted, indicated by subjects is shown in Table 3.2, where the total number of places is 101, and the average of the number of people overlapped is 2.1.

Here, the number of people overlapped means how many subjects indicate the same place, and the frequency means the number of places indicated by the same number of subjects.

Table 3.2: The places which should be deleted are indicated

the number of people overlapped	1	2	3	4	5	6	7
its frequency	56	16	11	6	6	3	3

Secondly, the places, which should not be deleted, indicated by subjects is shown in Table 3.3, where the number of places is 13, and the average of the number of people overlapped is 1.5.

Table 3.3: The places which should not be deleted are indicated

the number of people overlapped	1	2	3
its frequency	8	3	2

We itemize the process relevant to the 13 places shown above as follows. (the number in parenthesis shows a number of people overlapped)

- Processing of guess sentences: 4 (3) (2) (1) (2)
- Processing of expressions of address: 2 (1) (1)
- Processing of introduction parts: 7 (1) (2) (3) (1) (1) (1) (1)

3.4 Discussions

3.4.1 Validity of the proposed method

Firstly, we have discussions on the indication of the places which should be deleted. The six places which were indicated by more than half of subjects are itemized as follows.

1. The part from the beginning to "... とされた (...*tosareta*)" in an article. (1 place)

2. The part, "悪質な犯行で大きな社会的関心を呼んだ (aroused much public concerns with a malignant crime)", in the beginning of an article. (1 place)

3. The part which matches partially some other sentences in other articles. (1 place)

4. The part which should be processed as a person's name is not processed as that. This is caused by errors of morphological analyzer. (1 place)

5. The part which is relevant to the process of parenthesis. (2 places)

Here, the number in parenthesis means its frequency.

In the case of 1., the expression,"... とされた", should be included in introduction part, because that describes a past fact.

In the case of 2., the expression is the so-called clause of an adnominal form, and it is hard to cope with such a clause for the proposed method, because to cope with such a clause efficiently requires a parser for dependency analyses.

In the case of 3., the results were expected, because the proposed method does not cope with partial matching of sentences as we mentioned in the previous Section 3.2.4. It is able to apply to these phenomena by employing surface information that the clauses quasi-matched by Yamamoto et al[52].

In the case of 4., the morpheme should be a person's name, but it analyzed as a place name, and such errors easily happen in Japanese. To avoid these errors, we can employ heuristics which employ morphemes existing nearby the nouns, and expressing a property of person's name(e.g., "... 容疑者 (suspected ...)" or "従業員,... (... employee)").

In the case of 5., the indicated two places are classified into two types. One of them has an embedded structure of parenthesis, and the proposed method did not cope with such expressions. It is easy to fix this problem. And another is pointed out the two expressions, "関税貿易一般協定・多角的貿易交渉（ガット・ウルグアイ・ラウンド）(General Agreement on Tariffs and Trade / Multilateral Trade Negotiations(G.A.T.T. Uruguay Round))" and "ガット・ウルグアイ・ラウンド（関税貿易一般協定・多角的貿易交渉）". This is a paraphrase. In such a case, expressions should be unified a shorter expression "ガット・ウルグアイ・ラウンド". It is not so hard to realize this procedure.

Secondly, we have a discussion on the indication of the places which should not be deleted. We itemize the process of all the 13 places indicated by subjects, where the number in parenthesis means the number of people overlapped.

1. The process of introduction parts "... について (about ...)": 1 (1).

2. The process of introduction parts "a noun + は (*ha*)": 6 （1）（1）（1）（1）

（2）（3）.

3. The process of expressions of address: 2 (1) (1).

4. The process of guess sentences "... ようだ (it seems ...)": 1 (3).

5. The process of guess sentences "... なろう (be expected to ...)": 2 (2) (1).

6. The process of guess sentences "... そうだ (show an inclination to ...)": 1 (2).

In the case of 1., one person pointed it out. We guess the reason why this indicates that the subject felt the topic of the summary is not understandable due to the elimination of "... について (about ...)". It also may be a cause that the subjects are not notified enough on the assumptions imposed on the questionnaire.

The case of 2. was pointed out at 6 places, and these introduction parts are classified into two types(a. and b.) which has 3 places, respectively. In type a., there is "... による (by ...)" in front of "a noun + は (*ha*)", and an absence of "a noun + による (by ...)" leads the subjects to a feeling unnatural. It is easy to change the procedure of introduction parts to deal with this expression, when necessary. Next, in type b., there is "a noun + の (*no*)" in front of "a noun + は (*ha*)", and the subjects probably felt that an insufficient information of the summary is caused by the absence of "a noun + の (*no*)". It is also easy to change the procedure.

In the case of 3., the expression of the address exists on the article which is the first one of the series of articles. That expression has a place name which is included in the headline. The elimination of such nouns results in the fact that subject felt the expression is unnatural. Thus, the procedure should leave the expression which is included in its headline.

In the case of 4., there is a proper noun expression, “超伝導超大型粒子加速器（ＳＳＣ）(superconducting supercollider(SSC))”, in the sentence. The sentence is suitable for the guess sentence, but an existence of the proper noun in the eliminated part gave a few subjects unnatural impressions. To avoid these eliminations, we can employ a conventional method which utilizes a measure of importance for sentences[37]. It is also pointed out that the sentence exists on the first one of the series of articles, but there is no description relevant to the noun, thus the elimination is unnatural.

In the case of 5., the two places are pointed out, but the procedure deleted one sentence. The guess sentence has an expression, “... おり, ... なろう (*...ori*, *...narou*)”, and the number of people who pointed out the part, “... おり (*...ori*)”, is 2, and the number of people who pointed out the part, “... おり, ... なろう”, is 1. The sentence may not have to be eliminated due to the expression, “... おり”, which tells us a fact indeed. The question whether the sentence should be eliminated or not in such a situation is left for our future work.

In the case of 6., the sentence is the last sentence of the last one of the series of the articles. It is pointed out that such a sentence should be left.

As we have seen, the elimination of the guess sentences is proper from the form of the sentence. We have to employ a gauge of importance for sentences to improve the procedure of guess sentences. We consider the expression of address has less importance from the investigation.

From these discussions above, we conclude that the proposed method is valid.

3.4.2 The relevance between the compression ratio and style of articles

The compression ratio of a summary produced by the proposed method depends on the source articles. The experimental results enable us to say that the proposed method satisfies the required conditions of a summary which are to eliminate unimportant information and to leave important information. Whether the target articles include many verbose and overlapped parts or not, the proposed method summarizes them properly.

The proposed method is able to apply to any article. However, the compression ratio depends on the articles as we mentioned above. This implies a limit of the proposed method, but eliminations of overlapped and verbose parts aimed at this dissertation work properly.

Articles for an event or an accident tend to include many overlapped and verbose parts. In the experimental evaluation, the articles about an event achieved approximately 55% compression ratio(cf. Appendix A). Conversely, there are articles which include few verbose and overlapped parts, then the compression ratio attains more than 95%. However, we consider that people frequently require a development or an outline of the articles for an event or an accident. Thus the proposed method works correctly for the articles which are required to summary.

3.5 Concluding Remarks

We proposed a method to summarize multiple newspaper articles, and conducted experiments. The experimental results showed us that the proposed method summarizes the articles at about compression ratio 80% on the average by elimination of verbose and overlapped parts. It also showed by investigation of the questionnaires that the elimination of the proposed method is almost natural and the elimination is almost valid.

The processes of verbose and overlapped parts are able to be composed by heuristics and many of them are proposed in this chapter. The investigation of the questionnaires unveiled that there are the guess expressions which are hard to cope with. However, a conventional method which employ a gauge of importance for sentences[37] enable us to tackle them. How to reflect the gauge to our method is left for our future work.

Chapter 4

Automated Acquisition of Case Frames with Case Order

4.1 Introduction

In Japanese, case structure analysis is very important to handle several troublesome characteristics of Japanese such as scrambling, omission of case components, and disappearance of case markers. However, for lack of wide-coverage of case frame dictionaries, it has been difficult to perform case structure analysis accurately. This is because most of the past works collecting case frames were performed manually(e.g., [15, 48]).

We propose a model which considers the order of the case elements in a simple sentence and a method acquiring verbal case frames statistically. Here, a case element is composed of the semantic features and the case marker, and we call it a simple sentence if a sentence has only one verb and there has been no noun after the position of the verb.

Acquisition of verbal case frames with case orders enables us to examine that the possibility of the case frames to detail a verbal semantic classification by case orders. In addition, if there is a framework to retain case order, we would be able to utilize not only generation of languages having the most general case order, but also changing the case order to indicate an emphasis.

The acquisition of case frames with case order has a great significance from the practical point of view. We hope that case frames with case order will be applied to:

- acquisition of information, such as stress, which will contribute to a pragmatic context analysis,
- detailing a verbal semantic classification, and
- generation of sentences with natural case order.

4.2 Case transition network

We propose the case transition network as a framework for representing case order in addition to a case frame. We consider a method of acquiring case frames statistically from instances on a monolingual corpus.

We expand the *bi*-gram model employing a case element as a unit to reflect the word order precisely. The method proposed in this paper acquires surface verbal case frames by learning from a monolingual corpus on the case transition network.

The case transition network is roughly illustrated in Fig. 4.1. Outline of the learning on the model is as follows:

1. The model scans a case element from the beginning of a sentence.
2. The model transits a state to another by the appearance of a case element. Then the weight on the arc is calculated.
3. Finally, transition reaches to the terminal state by the appearance of a verb, and learning for a sentence is completed.

A path from the starting state to the terminal state, which consists of arcs having non-zero value, represents a case frame.

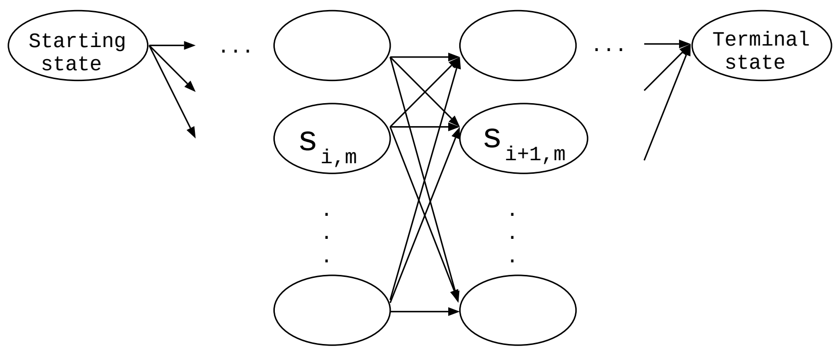

Figure 4.1: Overview of the case transition network model

The case transition network model is formally defined as follows. When a verb v, and a number of case elements n are given, the case transition network $N(v,n) = (S, A, w)$ is composed of three components, set S of states, set A of arcs, and weight w of arcs.

State: There are three kinds of states: starting, intermediate, and terminal states. A starting state corresponds to the beginning of a simple sentence, an intermediate state corresponds to a case element, and a terminal state corresponds to a verb. Each case transition network has exactly one starting and one terminal state. Intermediate states are distinguished by the appearance position, and are divided into layers. Each layer consists of all case elements. An intermediate state is represented by $s_{i,m}(1 \leq i \leq n)$, where i denotes the position of a case element in a simple sentence, and m denotes a case element.

Arc: An arc represents the relation between two states. The transition, from $s_{i,k}$ to $s_{i+1,l}$, is represented by arc $(s_{i,k}, s_{i+1,l})$. An arc $(s_{i,k}, s_{i+1,l})$ has a weight $w(s_{i,k}, s_{i+1,l})$.

The weight on arcs: The weight $w(s_{i,k}, s_{i+1,l})$ is given by the conditional probability $p(s_{i+1,l}|s_{i,k})$, where $p(s_{i,k}, s_{i+1,l})$ corresponds to the transitional probability from $s_{i,k}$ to $s_{i+1,l}$ and $p(s_{i,k})$ corresponds to the appearance probability of $s_{i,k}$. Then the $p(s_{i+1,l}|s_{i,k})$ is given by formula (4.1).

$$w(s_{i,k}, s_{i+1,l}) = p(s_{i+1,l}|s_{i,k}) = \frac{p(s_{i,k}, s_{i+1,l})}{p(s_{i,k})} \tag{4.1}$$

A sequence of case elements is represented by a path from the starting state to the terminal state. We call it learning to give the weights to the arcs.

The presuming probability in appearance for a sequence of case elements by a case transition network is given by the product of the weights on arcs at the path. For the definition of the case transition network, the network may have a case frame which does not appear in the corpus for learning. We expect that when the network has a case frame which does not appear in the corpus but has a high appearance probability, the case frame will be practically useful.

4.3 Experiments

To examine the appropriateness of the case transition network, we carried out two experiments. Firstly, to examine how much information of a sequence of case elements is preserved by the case transition network, we executed experiments of comparing the frequency of a sequence of case elements with the presumed probability by the case transition network. This is because, the case transition network is based on a bi-gram model, therefore the network does not handle long distance dependencies of cases. Secondly, we investigated whether the case order information by a case transition network is proper or not.

We used articles in The Nihon Keizai Shimbun, a Japanese daily newspaper for business, as a corpus. We obtained 557,048 sentences as a consequence of extracting simple sentences from the corpus.

We must assign semantic features to a noun. We adopted eighteen semantic features mentioned in IPAL verb dictionary[48] which is collected manually by IPA(The Information-technology Promotion Agency, Japan). We constructed a dictionary to assign semantic features to nouns by allocating semantic features to categories of the 'Kadokawa Ruigo Shinjiten' (Kadokawa New Thesaurus)[38], and we revised a part of the dictionary. We also added some nouns to the dictionary.

We picked up nine postpositional particles: "は (*ha*), が (*ga*), を (*wo*), に (*ni*), から (*kara*), へ (*he*), と (*to*), より (*yori*), で (*de*)" as case markers. Note that, although '*ha*' is not a case marker, we handle, in this paper, the '*ha*' for special postpositional particle as a case marker.

The method of extracting a sequence of case order for a verb is given as follows:

1. We obtain simple sentences in advance where a verb is designated from the corpus analyzed by morphological analyzer JUMAN[1].
2. The simple sentence is parsed by KNP[1].
3. We mark the case elements which consist of a noun and a case marker in the parsed sentence. If there is a suffix, the noun and the suffix are put together as a noun.
4. We assign semantic features to each noun in case elements by exact matching. If we can not assign semantic features to a noun by exact matching, we try the longest-first method from behind of the noun. Finally, when we can not assign semantic features to a noun, then we mark the noun as impossible assignment. It is possible that a noun is assigned several semantic features.
5. We output the sequence of case elements.

When a noun is assigned several semantic features in a sequences of case elements, sequences of case elements in compliance with the number of semantic features are produced.

We choose the verbs, "開く (to open), 始める (to begin), まとめる (to organize), 出る (to exit), 入る (to enter)", which have been ranked higher in the corpus as a subject of our investigation.

[1]http://pine.kuee.kyoto-u.ac.jp/nl-resource/

4.3.1 Investigation of probability presumed

We experimented on condition that a verb is '開く (to open)', and the number of case elements in a simple sentence is equal to three. Fig. 4.2 demonstrates the relevance between presumed sequences of case elements by the network and frequencies of sequences of case elements in the learning corpus. In Fig. 4.2, the horizontal axis indicates the frequency of a sequence of case elements, and the vertical axis indicates the probability of appearance for a sequence of case elements presumed by the case transition network.

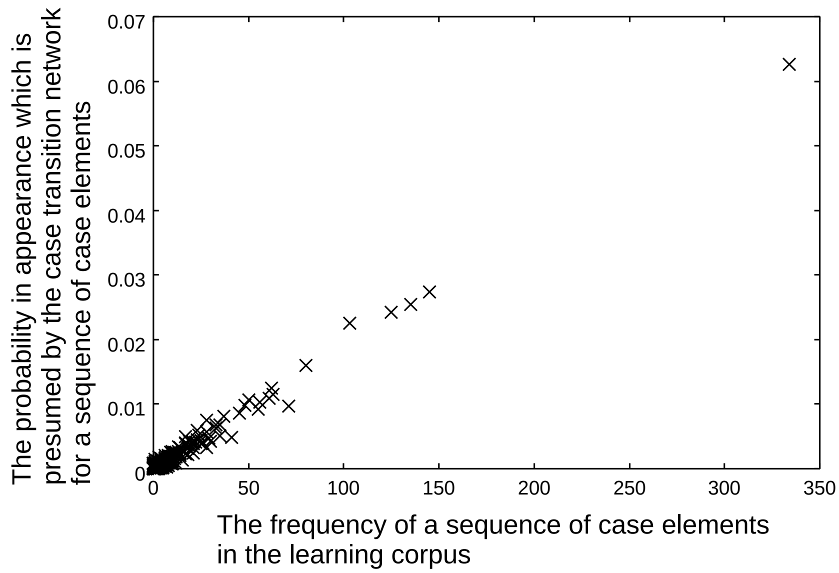

Figure 4.2: The relation between presumed sequences of case elements by the network and the sequences in the corpus for the verb, '開く (to open)'

When we carried out the investigation, we adopted three as the number of case elements by the following reasons:

- by the definition of the case transition network which is based on *bi*-gram, if the case elements number is one or two, there is no difference between the sequences of case elements in the corpus and the sequences of case elements by the networks, and

- we investigated the number of case elements in IPAL's verb dictionary[48] and found that the sentence having two case elements appears the most frequently and three case elements appears the second most frequently (Table 4.1 shows a result of the investigation, where the maximum number of case elements is five.).

We also investigated other verbs. Though the frequency and presumed appearance probability differ, the result shows that the presumed appearance probability becomes higher in proportion to the frequency of sequences by case elements in the corpus in a manner similar to the case of Fig. 4.2.

Table 4.1: A component ratio for the number of case elements in a simple sentence using the IPAL's verb dictionary

the number of case elements	1	2	3	4	5
obligatory case(%)	15.0	64.4	20.5	0.2	0
+optional case(%)	6.8	50.5	36.7	5.4	0.5

4.3.2 Investigation of case orders

To evaluate how case orders are preserved by the case transition network, we examined sentences having three case elements.

We defined preservation ratio to evaluate it as follows. Firstly, when a number of case elements is three, a sequence α_i of case elements, and sequences obtained by changing case order of α_i, consist of six sequences, then we let $\{\alpha_{i1},\ldots,\alpha_{i5}\}$ stand for the sequences except α_i. In addition, we defined a function g:

$$g(\alpha_i) \equiv \begin{cases} 1, & \forall j,\ p(\alpha_i) > p(\alpha_{ij}), \\ & \qquad s.t.\ f(\alpha_i) > f(\alpha_{ij}), \\ 0, & \text{otherwise.} \end{cases}$$

Here, $f(\alpha_i)$ denotes the frequency of α_i in the corpus, and $p(\alpha_i)$ denotes the presuming probability by the case transition network of α_i. Let Z be a set of sequences obtained by learning. Then the preservation ratio of case orders is defined by the following formula.

$$\frac{\sum_{\forall\alpha_i\in Z} g(\alpha_i)}{|Z|} \times 100(\%) \tag{4.2}$$

The results on investigation of five verbs are shown in Table 4.2.

4.4 Discussions

Fig. 4.2 shows that the case transition network presumes a sequence of case elements in proportion to the frequency of the sequence of case elements in the corpus for learning. We can conclude by Table 4.2 that the transition network is a

Table 4.2: The preservation ratio of case orders

the verb	(A) (%)	(B)
開く (to open)	95.0	2143
始める (to begin)	98.7	1304
まとめる (to organize)	97.2	205
出る (to exit)	97.6	336
入る (to enter)	95.7	419
average	96.8	881.4

(A): the preservation ratio of case order

(B): the number of simple sentences

model which sufficiently preserves the case order information. These results lead us to the conclusion that the transition network has a sufficient capacity to acquire case frames with case order.

When the number of case elements in a simple sentence is not less than three, the case transition network can cope with the problem of data sparseness by definition. However, the verification of the capacity to cope with the data sparseness is difficult owing to the ambiguity in the dictionary of semantic features.

In this model, we can assign several semantic features to each noun. Consequently, several semantic features on a noun demand the several learning sequences of case elements. Hence the number of learning sequences is very large compared with the number of simple sentences. For example, if a verb is '開く (to open)', and the number of case elements in a simple sentence is three, then 5,256 learning sequences of case elements are obtained from 2,143 simple sentences. It is easy for us to imagine that the above fact will do harm to the learning process of the case transition network. For the reasons mentioned above, employing a semantic feature dictionary with high quality will be needed to improve the case transition network. How to tackle these problems remains for future work.

Unfortunately, we could not examine the ability of the case transition network for detailing verbal semantic classification. However, the verb '開く (to open)' has two examples "N1 *ga* N2 *de* N3 *wo*" and "N1 *ga* N2 *wo* N3 *de*" as semantically different examples in the IPAL verb dictionary[48]. The fact encourages us to believe the possibility of verbal semantic classification by the case transition network.

4.5 Concluding Remarks

We proposed the case transition network model for acquiring case frames with case order automatically. The case transition network is remarkable as it preserves case order on case frames. The following conclusions were derived from our experimental results and discussions.

1. The proposed case transition network presumes a sequence of case elements in proportion to the frequency of case elements sequence in a learning corpus.
2. The network has a sufficient ability to keep case order.
3. Employing a semantic feature dictionary with high quality and wide-coverage will be required for improving the case transition network for practical use.

In the next Chapter, thus, we will investigate a method acquiring sequences of case elements with the *Goi-Taikei* as a semantic feature dictionary which has high quality and wide-coverage.

Chapter 5

Semantic Disambiguation on Acquiring Sequences of Case Elements from Corpora

5.1 Introduction

In this chapter, we propose a method of acquiring case sequences, which are utilized as case frames, from mono-lingual text corpora. We will also propose a word-sense disambiguation method which is required for acquiring case frames.

Chapter 4 revealed the necessity of word-sense disambiguation for acquiring case frames with case order. Thus, this chapter tackle on the problem of semantic ambiguity on acquiring sequences of case elements.

Although the importance of case frames in natural language processing is well understood, there is no case frame dictionary with high quality and wide-coverage. Some research institutes have constructed Japanese case frame dictionaries manually[15, 48]. However, it is quite expensive or almost impossible to construct a wide-coverage case frame dictionary manually. In addition, it is hard to maintain the case frame dictionary consistent and objective.

Thus, a number of studies have been done on automatic acquisition of Japanese verbal case frames[22, 33, 36]. The problems on acquiring case frames are as follows:

- Syntactic ambiguity
- Semantic ambiguity, especially polysemy of nouns

We aim to acquire conclusively case frames with case order. The case frames have to retain dependency relation between case elements and a verb and case order. In this study, we assume that a case element is composed of a noun and a case marker. Case frames with case order require sequences of case elements(SCE) which express the dependency relation between case elements and a verb with its order. Namely, SCE are sequences listing case elements which depend on a verb by its appearance order.

Kawahara et al.[22] conducted a study on acquiring case frames automatically from text corpora. However, case frames acquired by Kawahara et al. did not consider the case order.

To acquire case frames which describes relations between a verb and nouns from text corpora, a parser is required to analyze the dependency relation between them. However, parsed results include some errors, and the errors would be a hindrance to acquire case frames. In addition, when we assign semantic features to each noun to obtain case frames with semantic features, the word-sense ambiguity would be a problem.

In this study, syntactic ambiguities are reduced by utilizing simple sentences. We propose a method which employs statistical information for word-sense disambiguation to assign semantic features to nouns.

5.2 Extracting Sequences of Case Elements

In this section, we consider acquiring case sequences from corpora. We call the sequences of case elements(SCE) obtained from corpora by "example SCE" to distinguish from SCE assigned semantic feature. The overview of extracting SCE is as follows:

1. Extract simple sentences from corpora.
2. The simple sentences are parsed by a parser.
3. Extract the SCE from the parsing results.

Each processing is described below:

Extracting simple sentences: To avoid syntactic ambiguities, we do not utilize all sentences, but we employ simple sentences on corpora. The definition of a simple sentence is "a sentence which has one verb, and which has no noun behind the verb," where the verb is analyzed by a morphological analyzer, and we assume that the verb is placed at the end of the sentence. If a sentence has some declinable word(e.g., *sahen* noun + comma, adverb

+ conjugational stem, etc.) in front of the verb, the sentence would not be a simple sentence. The extraction of simple sentences employs the only morphological analyzer.

Parsing: The extracted simple sentences are parsed by a parser.

Extracting the SCE: We extract 3-tuples (verb, the number of case elements, the SCE) from parsed simple sentences.

Examples of the extracted example SCE are shown as follows:
見込む (to expect), 1, 四億円を (400 million yen-‘wo’)
集める (to attract), 2, ホタルが (fireflies-‘ga’), 人気を (popularity-‘wo’)
計画する (to plan), 2, 青森県六ケ所村に (Rokkasyo, Aomori-‘ni’), 建設を (construction-‘wo’)

5.3 Assignment of Semantic Features for Nouns and Coping with its Ambiguity

The extracted example SCE are not able to apply for natural language processing due to sparseness. Thus we assign a semantic feature to each noun. We employed the semantic codes of *Goi-Taikei*–A Japanese Lexicon[15] as approximation of semantic features. The *Goi-Taikei*[15] has three semantic attribute systems(3,000 semantic categories): the general noun semantic attribute system, the proper nouns semantic attribute system, and the verb semantic attribute system. We employed the general noun semantic attribute system, the proper nouns semantic attribute system, and its the Japanese semantic word dictionary which lists 300,000 Japanese words their pronunciation, part of speech, and semantic categories(described as semantic code).

5.3.1 Assignment of semantic features for nouns

The *Goi-Taikei*(a semantic feature dictionary) enables us to assign semantic features to nouns. However, all nouns are not always exist on the semantic feature dictionary. Thus, we employ the longest-first method from the back end for the subject noun. For example, if the noun "ダボス会議 (Davos meeting)" does not exist on the semantic feature dictionary, we will look at the dictionary "ボス会議 (BOSU-meeting)","ス会議 (SU-meeting)","会議 (meeting)" in this order. Consequently, if the string "会議 (meeting)" is exist on the dictionary, we would assign the semantic feature of the string "会議 (meeting)" to the noun "ダボス会議 (Davos meeting)."

If a string exists on the semantic feature dictionary, the string might have some parts of speech. The semantic feature dictionary has the following 8 parts of speech "noun, *sahen* noun, numeral, time noun, pronoun, proper noun, suffix, prefix." If a string has some parts of speech, the order of priority would be low in order shown above. For instance, a string has two parts of speech "*sahen* noun" and "suffix", then the "*sahen* noun" is adopted as the part of speech of the string. However, if a subject string has been short in the longest-first method, the string would be a "suffix" with highly prospects, and then we would give "suffix" the highest priority. To put it concretely, if the length of the original string is more than 3, and then the length of a part of the string would be not more than 2, the part of speech “suffix” would be given the highest priority. If the length of the original string is 2, and then the length of a part of the string would be 1, the part of speech “suffix” would be given the highest priority.

We call the SCE assigned semantic features to each noun “the assigned example SCE.”

Examples of the assigned example SCE are shown as follows:

見込む (to expect), 1, 2570 2595 2590 1190:を (‘wo’)

集める (to attract), 2, 549:が (‘ga’), 1150|2527:を (‘wo’)

計画する (to plan), 2, 23K:に (‘ni’), 1980 2075:を (‘wo’)

These are based on the examples of example SCE shown in Section 5.2.

5.3.2 Coping with word-sense ambiguity

A noun has some semantic features on the semantic feature dictionary in many cases. For example, the noun “単語 (word)” has the semantic feature “1084 語 (word).” On the other hand, the noun “タンゴ (tango)” has the three feature “1060 舞踊, 1675 舞踊 (dance)・演劇 (the theater)・諸芸 (arts), 1055 楽曲 (musical composition).” These word-sense ambiguity causes us a problem in natural language processing.

If an entry on the semantic feature dictionary had some semantic features, the features are arranged in order from basic meaning to derivative one. Thus, we can take an approach adopting the most basic semantic feature, which is described at the leftmost of the entry, for the noun’s one. We have investigated this approach, and the accuracy of disambiguation showed almost 70%.

However, there must be statistical differences between semantic features on condition which limit and determine the verb and the case. We investigate a method for word-sense disambiguation based on the following policies:

- Utilizing the order of semantic features on the semantic feature dictionary.

- Utilizing the examples of case elements which have no ambiguity.

The method employs statistical information obtained easily from the assigned example SCE. We expect that the accuracy of word-sense disambiguation with this method will be almost 80%.

We obtain the following three kinds of information from the assigned example SCE.

1. Semantic features having no ambiguity and its frequency: Firstly, we limit and determine a verb(v) and a case marker(m). Secondly, we investigate the noun which has been assigned single semantic feature(f_i) and its frequency($f(v,m,f_i)$). We calculate $f(v,m,f_i)$ with all available v, m and f_i.

2. Semantic features having ambiguity and its co-occurrence frequency: Firstly, we limit and determine a verb(v) and a case marker(m). Secondly, we investigate the noun which has been assigned some semantic features($f_i,\ i=1,\ldots$), and co-occurrence frequency between f_i and f_j, where $i \neq j$. The frequency is described as $f_{co}(v,m,f_i,f_j)$. For example, a noun was assigned three semantic features "f_1, f_2, f_3", and then f_1-f_2, f_1-f_3, f_2-f_1, f_2-f_3, f_3-f_1 and f_3-f_2 co-occur each time. In addition, the number of kinds that occurs with a semantic feature f_i

is described as $N_{co}(v,m,f_i)$. We calculate $f_{co}(v,m,f_i,f_j)$ and $N_{co}(v,m,f_i)$ with all available v, m, f_i and f_j.

3. The statistically weights of a semantic feature: Firstly, we limit and determine a verb(v) and a case marker(m). Secondly, we calculate the statistically weights $W(v,m,f_i)$ of a semantic feature f_i by the following formula:

$$\sum_{ce \in S(v,m,f_i)} \sum_{j=1}^{N(ce)} \frac{1}{p(ce_j,f_i)^3 \cdot N(ce)^{1.5}} \times \frac{1}{N_{co}(v,m,f_i)} \tag{5.1}$$

where, $S(v,m,f_i)$ corresponds to all case elements which has the case marker m and semantic feature f_i on the assigned example SCE for verb v. In other words, $S(v,m,f_i)$ is a multiple set.

The ce corresponds to a case element. There are some nouns in case elements which have some entry on the semantic dictionary even if the string and the part of speech are the same. For instance, "なまず" is a noun, and the noun has two entry. One corresponds to "鯰 (catfish)" which has an entry "543 魚 (fish) 842 魚介類 (fish and shellfish)", and the another corresponds to "癜 (leukoderma)" which has an entry "2419 病気類 (a kind of illness)." The number of entries which corresponds to a noun in a case element ce is described as $N(ce)$. Each element in ce is described as $ce_j (1 \leq j \leq N(ce))$.

The $p(ce_j,f_i)$ corresponds to the position of f_i in ce_j. For example, an entry "543 842 123" and the semantic feature 842 are given, then

$p(\text{“543 842 123”}, 842) = 2$.

The statistical weight W, given by formula (5.1) would be small if the position, expressed by $p(ce_j,f_i)$, is large. It means that more derivative word-sense is less important. The W also would be small if the $N_{co}(v,m,f_i)$ is large, and it means that a word-sense which tend to co-occur with many other senses is less important. In the formula (5.1), dividing by $N(ce)^{1.5}$ means a correction for addition on some entries.

The overview of disambiguation method proposed for a case element, where the verb is v and the case marker is m, is shown below:

1. If the case element(ce) has some entry, the ce would be divided into ce_j $(j = 1,...,N(ce))$.

2. For all ce_j, on each semantic feature (f_i $(i = 1,...) \in ce_j$), we calculate $f(v,m, f_i)/p(ce_j, f_i)$.

3. If the value of the greatest $f(v,m, f_i)/p(ce_j, f_i)$ is more than one, the f_i would be decided on the semantic feature of the case element.

4. For all ce_j, on each semantic feature (f_i $(i = 1,...) \in ce_j$), an f_i which has the greatest $W(v,m, f_i)$ is decided on the semantic feature of the case element.

5. If the case element is not disambiguated by processing above, we would make the first semantic feature of the first element for the semantic feature of the case element.

The disambiguated SCE is called semantic featured SCE.

5.4 Experiments

To examine the effectiveness of our proposed method, we executed some experiments. The corpora, we employed, are The Nihon Keizai Shimbun, the Japanese daily for business, CD-ROM edition from 1990 to 1996. We employed the morphological analyzer JUMAN and the parser KNP. The case markers on these experiments are eight case markers: "が (*ga*) を (*wo*) に (*ni*) から (*kara*) へ (*he*) と (*to*) より (*yori*) で (*de*)." The reason why these case markers were selected is that the parser KNP is able to utilize the IPAL verb dictionary[48], and we followed the dictionary. However, there are *bunsetu*s which have no surface case marker, but they are obviously case elements. In parsing results by KNP, those *bunsetu*s depend on a verb, and they have *bunsetu* pattern: $<$ 時間 (time) $><$ 数量 (quantity) $><$ 係:無格 (case:no-case) $>$. We treat these *bunsetu*s as case elements concerned with time. In these experiments, we choose the SCE which has the verb "開く (to open, to hold)", because the verb has high frequency in corpora.

Firstly, we extracted simple sentences from the corpora for five years(1990 to 1994), and we obtained 924326 sentences. These sentences were parsed by KNP. We extracted 9000 example SCE which have the verb "開く (to open, to hold)". We assigned semantic features to each case element on the example SCE, and then we could not assign semantic features to the total 130 nouns. From these assigned example SCE, we obtain $f(v,m,f_i)$, $f_{co}(v,m,f_i,f_j)$, $N_{co}(v,m,f_i)$ and $W(v,m,f_i)$.

Secondly, we extracted simple sentences, which have the verb "開く (to open, to hold)", from the 1995 corpus, and we obtained 1705 sentences. From these sentences, we obtained assigned example SCE in the same way as above, and then we could not assign semantic features to the total 30 nouns. The reason why we can not assign some nouns is that many proper nouns do not have entry on the semantic feature dictionary. Thus we tried to expand the semantic feature dictionary as much as possible. We also adjusted the parameters required for our method at the same time. To adjust the parameters, we applied the disambiguation method to the assigned example SCE obtained from the 1995 corpus. The disambiguation was done based on the statistical information derived from the 1990-1994 corpora. Finally, we determined the parameters as values mentioned in section 5.3.2.

After the expansion of the semantic feature dictionary, the total numbers of nouns which we could not assign semantic features are 32 for the 1990-1994 corpora, and 16 for the 1995 corpus.

Firstly, to evaluate the proposed method, we obtained the assigned example SCE from the 1996 corpus in the same way as above. The number of SCE is 2058, and we applied the disambiguation method for these SCE. Secondly, we consider a disambiguation method which decides the semantic feature by choosing the leftmost semantic feature as a base-line method, and we compared this base-line method with ours.

We choose 90 SCE derived from the 1996 corpus at random. In this 90 SCE, a case element is removed from subjects of evaluation if a case element is applicable to any one of the following four conditions:

1. The case element is not correct. A parsing error causes this case element.
2. The case element is failed on assigning semantic features, and there is no suitable semantic feature.
3. The case element is assigned semantic features, but there is no suitable semantic feature.
4. The semantic features are nearly the same, and if we choose any of them,

there might be no difference.

Finally, we obtained 119 case elements for evaluation.

The evaluated results by our method and the base-line's are shown in Table 5.1. The breakdown for information which is utilized by the proposed method is also shown in Table 5.2: (A) frequency of nouns which are assigned single semantic feature($f(v,m,f_i)$), and (B) the statistically weights ($W(v,m,f_i)$).

Table 5.1: Disambiguation results

Method	correct	incorrect	accuracy(%)
Ours	94	25	78.99
Base-line	84	35	70.59

Table 5.2: Breakdown for information utilized

information	correct	incorrect	total	accuracy(%)
(A)	79	21	100	79.00
(B)	15	4	19	78.95

5.5 Discussions

From Table 5.1, the proposed method has effectiveness for the word-sense disambiguation. Table 5.2 shows that many of what utilized for disambiguation are the frequencies of nouns which are assigned single semantic feature(100/119). If the frequencies $f(v,m,f_i)$'s are fully obtained for all f_i, the disambiguation would be done by utilizing only $f(v,m,f_i)$ with almost accuracy 80% (79/100). However, it is hard to imagine such a situation. In experiments, the 19 case elements did not disambiguate by $f(v,m,f_i)$, but disambiguated by $W(v,m,f_i)$ proposed in this paper with about accuracy 80%(15/19).

It is very important to assign semantic features to nouns, because the assignments affects the whole SCE from which we obtain statistical information. In the proposed method, we look at the semantic feature dictionary by character based matching to assign semantic features to nouns. On the other hand, we can consider a method which looks at semantic feature dictionary with morphological analyzer not character based matching. However, in either method, it is difficult to cope with proper nouns. The biggest problem on assigning semantic features to proper nouns is that a proper noun does not have entry on the semantic feature dictionary. Thus, to obtain practical SCE, the very large semantic feature dictionary which has fully proper nouns entry or method identifying proper nouns automatically is required.

At disambiguations by $f(v,m,f_i)$ in incorrect cases, the value calculated by $f(v,m,f_i)/p(ce_j,f_i)$ tend to be small. If the threshold value for decision is changed, the accuracy might be changed. In the proposed method, such parameters are defined empirically. However, the most suitable parameters may be fluctuated by amount of corpora, the subject of verb, etc. Coping with these problems are left for our future works.

5.6 Concluding Remarks

In this chapter, we proposed a method for acquiring sequences of case elements(SCE) from text corpora. One of problems, the syntactic ambiguity, on acquiring SCE is suppressed by utilizing simple sentences. To acquire case frames which assigned semantic features, we investigate a method for assigning semantic features to nouns. We proposed a method utilizing statistical information for assigning semantic features to nouns. By comparing the proposed method with a base-line method, the proposed method is very promising in word-sense disambiguation.

Chapter 6

Conclusion

Throughout this dissertation, we have prepared to develop methods for supporting and enhancing intellectual activities. In particular, we have proposed methods for retrieving relevant documents and for summarizing multiple articles. We also have investigated a method for acquisition of case frames with case order.

Chapter 2 proposed a method for retrieving relevant documents by a document. The feature of the proposed method is to employ connections of morphemes as the unit of index terms. The proposed method was compared by experiments with a conventional method, proposed by Araya et al[1]., employing morphemes as the unit of index terms. The experimental results showed that the proposed method outperformed the conventional method.

Chapter 3 proposed a method for summarizing multiple articles. The summarization method is characterized by employing heuristics, and the proposed method summarized the articles at compression ratio approximately 80% on average by elimination of the verbose and overlapped parts. We confirmed by questionnaires that in summarization of such as newspapers having special structures of texts, heuristics are suitable for producing natural and readable summaries.

Chapter 4 investigated a method for acquiring case frames with case order from mono-lingual text corpora. The feature of the proposed acquisition method is to consider case orders and approximating appearances of case elements as *bi*-grams. We executed acquisition experiments employing the same semantic feature of the IPAL verb dictionary[48]. From the experimental results, we conclude that the method is promising, but the results unveiled the necessity of semantic disambiguation for practical use, and the semantic features of the IPAL verb dictionary are insufficient for automatically acquiring case frames.

Chapter 5 presented a method for acquiring sequences of case elements with semantic features. The feature of the proposed method is unsupervised word-sense disambiguation employing statistical information. We executed experiments employing The Nihon Keizai Shimbun, from 1990 to 1996, as corpora for disambiguation. By comparing the proposed method with a base-line method, we confirmed that the proposed method is very promising.

Before closing this dissertation, we would like to point out the following several possible topics for future research, which may be important and interesting.

- To explore and explain essentially effective units of index terms for Japanese documents. To begin with, do such units really exist? If such units exist, it will depend on a domain or not? Along this line, a very interesting work has done by Ozawa et al.[39].

- To build a multiple summarization method which summarizes each sentence or paraphrases some sentences into a sentence(e.g., [20]), and users are able to set the compression ratio freely.

- To acquire case frames with case order by the proposed method mentioned in Chapter 4 from sequences of case elements brought by the techniques presented by Chapter 5.

- To evaluate the sequences of case elements obtained at Chapter 5 by comparing with the verb semantic attribute system of *Goi-Taikei*.

- And to investigate how many semantic features should we consider on acquiring case frames. Ikehara et al.[16] has employed approximately 3000 semantic attributes for Japanese-English machine translations.

Appendix A

Examples of multiple summarization

The examples by multiple summarization method which mentioned in Section 3.2 are shown as follows. The boldface parts are eliminated by the system. The compression ratio of these articles is 56.0%.

1. No. 1. (February 13th, 1992)

   Headline: 北海道警,「○○○○」から４億１０００万，恐喝の会社社長ら逮捕.

   北海道警捜査四課と豊平署は十二日，廃バッテリー回収設備の工事をめぐって九○年四月と十一月ごろ東証一部上場の化学会社○○○○（本社・札幌市，○○△社長）から現金三億四千万円を脅し取ったり，七千万円の債権を放棄させていた**北海道○○町□町一三四，**廃品回収業「□□」社長，□□▼▼容疑者（５１）と○○**市△△町八，**無職●▽▽容疑者（５５）の二人を恐喝容疑で逮捕した.

2. No. 2. (February 18th, 1992)

   Headline: ○○○○恐喝事件で道警，■■前道議を逮捕.

   東証一部上場の高圧ガス，産業機器メーカー「○○○○」（本社札幌市）恐喝事件で道警捜査四課と札幌・豊平署は十七日，恐喝の疑いで新たに○○**管内**○○**町**□□**二，**前道議■■◎◎容疑者（５６）を逮捕した. 同事件の逮捕者は三人となった.

3. No. 3. (February 20th, 1992)

   Headline: ■■前道議を送検，○○○○恐喝事件.

   東証一部上場の化学会社「○○○○」（本社札幌市，社長○○△氏）が廃バッテリー回収設備の工事をめぐって現金約三億四千万円を脅し取られたり工事代金の債権（約七千万円）を放棄させられていた事件で道警捜査四課と札幌豊平署は十九日，恐喝容疑で逮捕した**前道議の■■**◎◎容疑者**（５６）＝**○○**管内**○○**町**□□**二＝**を札幌地検に送検した．同課は○○○○恐喝の中で■■容疑者が果たした役割などを本格的に追及，事件の解明を急ぐ.

4. No. 4. (February 23rd, 1992)

   Headline : ○○○○恐喝，新たに会社社長逮捕.

   東証一部上場の化学メーカー，「○○○○」恐喝事件で道警捜査四課と札幌・豊平署は二十二日，新たに○○**市**△△**三丁目，**会社社長，●●□□容疑者（４３）を恐喝の疑いで逮捕した．同事件の逮捕者はこれで四人目となった.

5. No. 5. (March 5th, 1992)

   Headline: ○○○○恐喝，前道議ら３人起訴――札幌地検，余罪裏付け

急ぐ.

東証一部上場の高圧ガス，産業機器メーカー「○○○○」（本社札幌市）恐喝事件で札幌地検は四日，恐喝罪で**北海道□□郡○○町△△二，前北海道議会議員，■■◎◎（５６）**，**同郡○○町□□□八四，廃品回収業，□□▼▼（５１）**，**○○市△△町八ノ八，無職，●▽▽（５５）**の三容疑者を起訴した.

6. No. 6. (April 8th, 1992)

   Headline: ○○○○恐喝で札幌地検，●被告を追起訴.

   東証一部上場の高圧ガス，産業機器メーカー「○○○○」（本社札幌市）恐喝事件で札幌地検は七日，先に恐喝罪で起訴していた**○○市□□町八ノ八，無職●▽▽**被告**（５５）**を同罪で追起訴し，事件の捜査を終了した．被害総額は約七億四千万円となった.

Appendix B

Overview of 6 groups utilized at the questionnaires

1. An extortion case of "○○○○".

   The number of articles: 6, The compression ratio: 56.0%

2. A picture book of the Madonna inspected at the customs.

   The number of articles: 4, The compression ratio: 77.5%

3. A small group of officers and soldiers revolted at Mindanao, Philippines.

   The number of articles: 5, The compression ratio: 82.5%

4. A hit-and-run fatalities of a mother and her children at Tokyo.

   The number of articles: 5， The compression ratio: 80.4%

5. A series of arson cases at Osaka and Nara.

   The number of articles: 6， The compression ratio: 68.3%

6. Japan-U.S. summit.

The number of articles: 7， The compression ratio: 83.1%

Bibliography

[1] Ken Araya, Tatsuhiko Tsunoda, Takumi Ooishi, and Makoto Nagao. A retrieval method of relevant newspaper articles using word's cooccurrence frequency and location. *Transactions of Information Processing Society of Japan*, 38(4):855–862, 1997. (in Japanese).

[2] Kazuaki Aso, Tsunenori Mine, and Masato Amamiya. A Japanese text retrieval system with dependency structures on WWW. In *Proceedings of The Third Annual Meeting of The Association for Natural Language Processing*, pages 253–256, 1997. (in Japanese).

[3] Rcardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[4] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 550–557, 1999.

[5] Nicholas J. Belkin and W. Bruce Croft. Retrieval techniques. *Annual Review of Information Science and Technology(ARIST)*, 22:109–145, 1987.

[6] Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? *Communication of the ACM*, 35(12):29–38, 1992.

[7] James P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the seventeenth annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 302–310, 1994.

[8] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Devlopment in Information Retrieval*, pages 318–329, 1992.

[9] David Ellis. *New Horizons in Information Retrieval*. Library Association Publishing, 1990.

[10] David A. Evans and Chengxaing Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of the 34th Annual meeting of Association for Computational Linguistics*, pages 17–24, Santa Cruz, California, June 1996.

[11] William B. Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval Data Structures & Algorithms*. Prentice Hall, 1992.

[12] Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. A comparative evaluation of recent corpus-based word sense disambiguation techniques. In *IPSJ SIG Notes 97-NL-119*, pages 45–52, 1997.

[13] Marti A. Hearst and Jan O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR-96)*, pages 76–84, 1996.

[14] Masao Hirai. *The Encyclopedia for Writing*. SANSEIDO, 1984. (in Japanese).

[15] Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentarou Ogura, Yoshifumi Oyama, and Yoshihiko Hayashi, editors. *Goi-Taikei – A Japanese Lexicon*. Iwanami Publishing, CD-ROM edition, 1999.

[16] Satoru Ikehara, Masahiro Miyazaki, and Akio Yokoo. Classification of language knowledge for meaning analysis in machine translations. *Transactions of Information Processing Society of Japan*, 34(8):1692–1704, 1993. (in Japanese).

[17] The National Language Research Institute. *Vocabulary and Chinese Characters in Ninety Magazines of Today (3)–Analysis–*, pages 171–239. SYUEI Publishing, Japan, 1964. (in Japanese).

[18] Makoto Iwayama and Takenobu Tokunaga. Associative document search using a probabilistic document clustering. *Journal of Natural Language Processing*, 5(1):101–117, 1998. (in Japanese).

[19] Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Summarization evaluation methods: Experiments and analysis. In *Intelligent Text Summarization*, pages 51–59. AAAI Press, 1998.

[20] Hongyan Jing and Kathleen R. McKeown. Cut and paste based text summarization. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'00)*, pages 178–185, 2000.

[21] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[22] Daisuke Kawahara, Nobuhiro Kaji, and Sadao Kurohashi. Japanese case structure analysis by unsupervised construction of a case frame dictionary. In *Proceedings of the 18th International Conference on Computational Linguistics(COLING 2000)*, pages 432–438, 2000.

[23] Margaret King. Evaluating natural language processing system. *Communications of the ACM*, 39(1):73–79, 1996.

[24] Kenji Kita, Satoshi Nakamura, and Masaaki Nagata. *Spoken Language Processing*. MORIKITA Publising, 1996.

[25] Hang Li and Naoki Abe. Generalizing case frames using a thesaurus and the mdl principle. *Computational Linguistics*, 24(2):217–244, 1998.

[26] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, 1957.

[27] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–168, 1958.

[28] Inderjeet Mani and Eric Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings of the 14th National Conference on Artificial Intelligence*, pages 622–628, 1997.

[29] Inderjeet Mani and Mark T. Maybury, editors. *Advances in Automatic Text Summarization*. The MIT Press, 1999.

[30] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[31] Kathleen McKeown and Dragomir R. Radev. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, 1995.

[32] Makoto Mikami, Shigeru Masuyama, and Seiichi Nakagawa. A summarization method by reducing redundancy of each sentence for making captions of newscasting. *Journal of Natural Language Processing*, 6(6):65–81, 1999. (in Japanese).

[33] Tsunenori Mine, Masaru Higashi, and Makoto Amamiya. Case frame acquisition and verb sense disambiguation on a large scale electronic dictionary. In *Proceedings of NLPRS '97*, pages 221–226, 1997.

[34] Makoto Nagao, editor. *The IWANAMI Software Science Series 15 Natural Language Processing*. IWANAMI SHOTEN, 1996. (in Japanese).

[35] NetSizer. http://www.netsizer.com/.

[36] Akira Oishi and Yuji Matsumoto. Lexical knowledge acquisition for Japanese verbs based on surface case pattern analysis. *Transactions of Information Processing Society of Japan*, 36(11):2597–2610, 1995. (in Japanese).

[37] Manabu Okumura and Hidetsugu Nanba. Automated text summarization: A survey. *Journal of Natural Language Processing*, 6(6):1–26, 1999. (in Japanese).

[38] Susumu Ôno and Masando Hamanishi. *Kadokawa Ruigo Shinjiten (Kadokawa New Thesaurus)*. Kadokawa Publishing Co., 1981. (in Japanese).

[39] Tomohiro Ozawa, Mikio Yamamoto, Kyoji Umemura, and Kenneth W. Church. Japanese word segmentation using similarity measure for IR. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 89–96, 1999.

[40] IREX Home Page. http://cs.nyu.edu/cs/projects/proteus/irex/.

[41] NTCIR Home Page. http://research.nii.ac.jp/ntcir/index-en.html.

[42] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings in ANLP/NAACL2000 Workshop on Automatic Summarization*, pages 21–30, 2000.

[43] Dragomir R. Radev and Kathleen R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500, 1998.

[44] Gerard Salton, editor. *THE SMART RETRIEVAL SYSTEM / Experiments in Automatic Document Processing*. Prentice-Hall, Inc., 1971.

[45] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[46] Shogo Shibata, Takaya Ueda, and Yuji Ikeda. A fusion of overlapped documents. In *IPSJ SIG Notes 97-NL-120*, pages 77–82, 1997. (in Japanese).

[47] Netcraft Web Server Survey. http://www.netcraft.com/survey/.

[48] Information technology Promotion Agency(IPA). *IPA Lexicon of the Japanese Language for computers IPAL (Basic Verbs)*, 1987. (in Japanese).

[49] Hideo Teramura, Yasushi Suzuki, Naofumi Noda, and Makoto Yazawa, editors. *Case Study Nihon Bunpou(Case Studies on Japanese Grammar)*. O-U-FU-U, 1987. (in Japanese).

[50] TREC. http://trec.nist.gov/.

[51] Takehito Utsuro, Yuji Matsumoto, and Makoto Nagao. Verbal case frame acquisition from bilingual corpora. *Transactions of Information Processing Society of Japan*, 34(5):913–924, 1993. (in Japanese).

[52] Kazuhide Yamamoto, Shigeru Masuyama, and Shozo Naito. An empirical study on summarizing multiple texts of Japanese newspaper articles. In *Proceedings of Third Natural Language Processing Pacific-Rim Symposium(NLPRS '95)*, volume 1, pages 461–466, 1995.

[53] Kazuhide Yamamoto, Shigeru Masuyama, and Shozo Naito. GREEN: An experimental system generating summary of Japanese editorials by combining multiple discourse characteristics. *Journal of Natural Language Processing*, 2(1):39–55, 1995. (in Japanese).

[54] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, 1995.