NAIST-IS-DT9961016

Doctor's Thesis

# A Study on Operations used in Text Summarization

Kazuhiro Takeuchi

February 5, 2002

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

Doctor's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
DOCTOR of ENGINEERING

Kazuhiro Takeuchi

Thesis committee: Yuji Matsumoto, Professor
Shunsuke Uemura, Professor
Katsumasa Watanabe, Professor

# A Study on Operations used in Text Summarization[*]

Kazuhiro Takeuchi

## Abstract

Current automatic summarization systems basically consist of extraction and concatenation of phrases and sentences of particular features, but humans seem to summarize in a more fine-grained manner. In order to improve the quality of machine summarization, not only lexical but also textual features relevant to summarization as well as the operations humans adopt for summarization need to be determined, and each of them must be carefully examined from the viewpoint of implementation. This dissertation consists of 3 major topics as follows:

First, we carefully observe the pair of the original texts and their summaries by humans in terms of text structure, paying special attention to sentence reduction and sentence combination. Sentence reduction is an operation that shortens the original sentence by removing some of its constituents. Sentence combination is an operation that makes two or more original sentences are combined into a summary sentence. Both operations are constantly used by humans in order for a summary sentence to be concise and consistent, but they presuppose the understanding of the relations between sentences, the text structure. We discover, by manually analyzing human summaries of newspaper articles, that even human subjects have difficulties in judging the relatedness between sentences when the related sentences are not adjacent, though humans are far more accurate than automatic summarization systems in determining the relatedness between adjacent sentences. We also discover that relatedness between adjacent sentences is crucial in summarization. Using a machine learning method, we confirm that a set

[*] Doctor's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT9961016, February 5, 2002.

of linguistic features can characterize how strongly an adjacent pair of sentences relates to each other.

Second, in order to analyze summary operations in detail, we introduce an algorithm dealing with the dependency structure of sentences for aligning a summary expression with the corresponding original expression in the source text. We use this algorithm to investigate human summaries of newspaper editorials. We discover that most of the summary expressions keep their dependency structure in the original sentences, and thus the proper sentence combination plays a crucial role in generating a consistent summary. We also categorize new expressions in summary sentences from the viewpoint of paraphrasing.

Finally, we discuss an experimental implementation of sentence reduction. Support vector machines (SVMs) are used for acquiring knowledge for sentence-reduction operation. The training data are extracted from human summaries with an automatic alignment program. In result, we determine that linguistic features acquired with the help of SVM include the knowledge required for sentence reduction.

**Keywords:**

Text Summarization, Sentence Generation, Automated Alignment,Text Structure,Paraphrasing

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation

Due to the rapid growth of the Internet and the emergence of low-cost, large-capacity storage devices, we are now exposed to a large amount of on-line information in daily life. This situation makes it difficult for us to find and gather the exact information we need. Automatic text summarization is a key technology to overcome this difficulty; with the properly summarized information, we can quickly and easily understand what the major points of the original document are and find how relevant the original document is to our own needs.

However, summarization is a hard problem of the Natural Language Processing (NLP), for summarization, in principle, presupposes a fair understanding of the content of the original document. When we summarize a document, we firstly try to understand what the document says, then try to extract the ”more important” parts from it, and finally try to compose a consistent passage, the summary. However, the current NLP technology can deal only with very basic, if at all, meaning of the text, most of which are lexical or logically semantic; it cannot understand where the ”more important” parts are in a true sense; and there still exists many difficulties in the coherent text generation. In a sense, the true text summarization, which humans do, goes far beyond the current NLP technology. Still, even when a machine, or even a human, cannot understand the meaning of a text, it can distinguish, for example, the more frequent, seemingly important, parts from others, which enables the extraction of seemingly important sentences.

At the same time, the itemized pieces of information, each of which are not so coherent, may well work for obtaining the overview of the original document. From this viewpoint, most automated summarization systems today focus on the proper extraction of important phrases or sentences in the original document by using various types of formal but not semantic features, though their achievements are still below our satisfaction.

In order to extend such limits of the extraction and of the technique of text processing, we have to know how a human produces a summary from the original. The purpose of this dissertation is to model human process of generating summaries by an investigation of human-written summaries.

## 1.2 Attempts on Automated Text Summarization

Although the attempts at automated text summarization have been done since 1950s, there is a gap between the summaries produced by current automatic summarization systems and the summaries written by humans. Earliest attempts at summarization [19] essentially relied on lexical and locational information within a text. Such an approximation called extraction is still a fundamental method today. To create an extract, a system simply needs to identify the most important/topical/central topic(s) of the text, and return them to the reader.

Compared with answering queries in database, it is very difficult to define what the correct summary is in the summarization system because the notion of importance is very ambiguous. Clearly, the importance of a text varies with its genre, domain, and so on. Furthermore, the importance also varies with what kind of summaries a user wants. One of the general types of summaries that have been identified is a contrast between indicative and informative. The summaries of the former can be used to indicate what topics are addressed in the source text, and can be used to alert the user to the source content. The purpose of the latter is to suggest the contents of the article without giving away detail on the article content. It can serve to entice the user into retrieving the full form. Book jackets, card catalog entries and movie trailers are examples of indicative summaries. In order to decline the discussion of such diversity of the importance, most of researches prepare the target extractions of the texts that are assumed to be important to design or to evaluate a summarization system.

In most of the summarization systems, the importance of sentences or phrases is determined with the features of the original text. The following features are examples of such features that summarization system always adopt and are essentially multiple clues for the importance.

- Frequency of keyword appearance in an article.
- Title or headline of an article
- Key expressions that appear in an article
- Position of a sentence in an article or in a paragraph

Those features become richer corresponding to a continuum of increasing complexity in text processing techniques. For example, the improvement of the importance measurement of words has drawn upon traditional information retrieval indexing methods to incorporate knowledge of a text.

There are adaptations that have employed an automated method to combine these feature sets through the machine learning techniques such as [18, 1]. The advantage of such approaches is that once 'good extractions' are provided, researches can concentrate on inquiry into clues and on improving the machine learning techniques.

## 1.3 Approaches Based on Text Structure Analysis

Although it is intuitively understandable why the features that we showed in the section 1.2 work well in the extraction, such features do not rely on the model of text understanding. For example, we know that the first sentence of a text usually shows the important content, but it is not derived from a model of text understanding or from a result of text analysis.

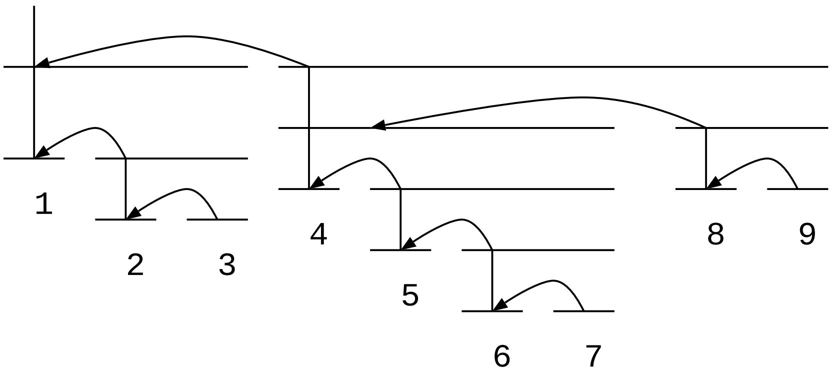


Figure 1.1. Representation of a text structure in RST

Motivated by the lack of such sophisticated models, much of the work focus on the text structure to identify the important candidates. Discourse model can orgnaize individual discourse features. Ono and his colleagues consider the way to select the sentences based on the representation of the rhetorical structure of texts [30, 34]. Rhetorical Structure Theory is a theory to represent how the text consists their sentences. Figure 1.1 shows an example of text structure. Each short horizontal line with a number represents a sentence in the text. Such sentences are connected with each other with arcs and form the tree structure.

Marcu [23, 22] also challenges automated summarization using such representation. He extracts sentences by the positions in the tree structure and gives a few candidate methods to extract important sentences.

A series of such development of automated summarization systems influences us in many ways. However, there is no concrete way to make a summary through a representation of text structure. Moreover, what representation in various discourse model is reliable for generating a summary is not clear. In this dissertation, we discuss this problem in Chapter 3.

## 1.4 Summarization based on Investigation of human-written Summaries

There are two reasons why there are gaps between human-written summaries and ones by a system. One is that the definition of "good summaries" is very ambiguous as we described in Section 1.2. The other is the output of such a summarization system have not focused on generation process of readable summaries.

According to the progress of the researches, summarization systems came to have the following two general phases.

- Extraction phase: select the significant sentences that show the main information of the text.
- Revision phase: revise the extracted sentences into simpler ones.

Although the central issue of the summarization system was the improvement of the extraction phase, some works have started focusing on revising phase. In the revision phase, however, the problem of defining the good/ readable/ high quality summaries arises again. In order to know the properties of good summaries, which should be explained in terms of the model of text structure, more investigation into the human-written summaries is needed.

Researches such as [20, 29] propose models to revise extracted sentences. Such works decompose a process how people revise extracted sentences into several operations from the investigations of human written summaries.

Jing and McKeown, among others, propose a model, namely a cut and paste base text summarization [12]. They divide the summarization process into six operations, which are derived from their manual investigation on human-generated summaries. Those operations can be used alone, sequentially, or simultaneously to transform extracted sentences.

In this dissertation, we discuss text summarization from that point of view. Figure 1.2 shows the overview of our project. In the figure, the solid arrows show the flow of information from an original text to the summary. Our summarization model decomposes revising process into 3 operations. On the other hand, the dotted arrows reprsent a plan how we acquire the knowledge to the operations.

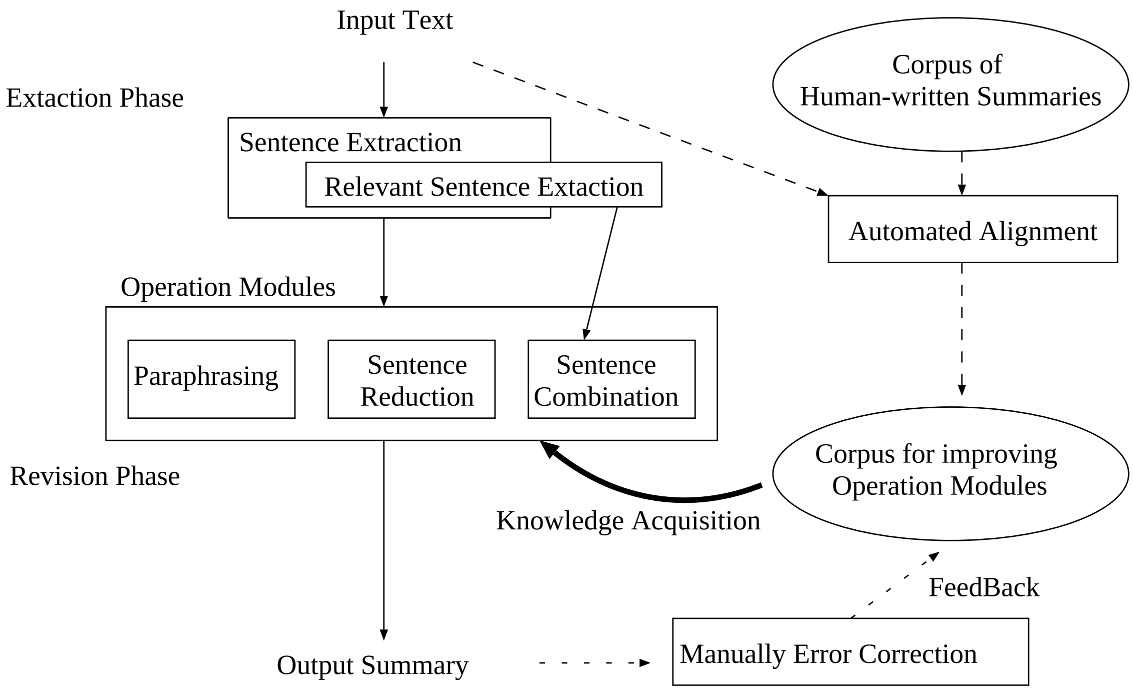


Figure 1.2. Overview of Our Project

## 1.5 Outline

The target data in this dissertation is Japanese texts. We investigate operations in manually generated summaries from the viewpoint of discourse analysis. In this section, we introduce the outline of this dissertation referring the overview of project in Figure 1.2.

In Chapter 2, we will briefly describe the notions of text analysis and the operations that related works assumed.

We, then, describe two investigations in Chapter 3. One is to estimate the consistency of human analysis on text structure and to investigate clues for forming such consistent structure. The other is to investigate operations how human do to produce summary sentences. Those investigation contribute why we focus on operations in summary generation and relate with the Relevant Sentence Extraction and Sentence Combination in Figure 1.2.

In Chapter 4, we further investigate the operations in summary generation based on dependency structure. The results of the Chapter 4 contribute to find the properties of the operations and techniques used there strongly relate with the arrows with a broken line in Figure 1.2.

In Chapter 5, we describe an experimental implementation of one of the operations to generate summary sentences. The process corresponds to the broad arrow in Figure 1.2.

Finally, Chapter 6 is the conclusion and shows some implications for future research.

# Chapter 2

# Related Work

## 2.1 Notions of Text Analysis

### 2.1.1 Cohesion

A text is not just a random collection of sentences. It possesses coherence and thematic structure, with which the content is expressed in a way that is easy for humans to understand. A computer database can accept updates and facts in random order, but a human reader finds information much easier to assimilate if it is presented in a well-structured manner.

One way of classifying text model used in text summarization is in term of the linguistic distinction between cohesion and coherence proposed by Halliday and Hasan [6]. In their book, the property that makes up a text is called texture. Readers can tell whether or not a series of sentences exhibits texture. In the following pair of sentences, sentence (a) exhibits texture and sentence (b) does not.

a) Wash and core six cooking apples. Put them into a fireproof dish.

b) Wash and core six cooking apples. The prices of computers drop regularly.

Cohesion is one of the elements of a text which contributes to its texture. Halliday and Hasan identify the following cohesive relations which make a sequence of sentences be a text.

- Reference

  References are like pointers. Rather than repeating a phrase in the text, a writer or speaker may use a a pronoun instead of the original phrase. Halliday and Hasan distinguish two main types of reference. Exospheric references are to the entities in the world of the discourse and endophoric references are to the positions of the text itself.

- Substitution

  Substitution and reference are similar, but differ in that substitution occurs prior to semantic interpretation while reference occurs after interpretation. That is, a substitute acts merely as a pointer to a region of text which refers to an entity in the world of the text or the discourse, while a reference refers directly to an entity without the mediation of the original referring phrase.

- Ellipsis

  Ellipsis can be viewed as a special case of the substitution. It avoids obvious repetitions and substitutions by omitting such entities.

- Conjunction

  Conjunction holds between elements of a text when they are ordered temporally, one causes the other, when they describe a contrast or when one elaborates on the other.

- Lexical cohesion

  Lexical cohesion holds between two words in a text which are either of the same type or are semantically related in a particular way. Halliday and Hasan propose 5 semantic relations that constitute lexical cohesion. However, their definition of the lexical cohesion is ambiguous.

Although some criticizes that Halliday and Hasan's categories overlap to some degree, the notion of cohesion is recognised as a fundamental notion of text analysis.

On the other hand, coherence is defined as element that conveys the better interpretation of a text, however, Halliday and Hasan did not propose a concrete model for the coherence.

### 2.1.2 Coherence Representation: Rhetorical Structure

Cohesion is basically related to the linguistically realized cues as shown in the previous section, but coherence is considered to be a more abstract notions.

A number of theories to discuss the coherence structure in text have been proposed in the literature. More concretely, there is a model to produce a tree whose leaf nodes are messages and whose internal nodes specify the following information [9].

- how sentences are grouped together thematically
- the order in which sentences (or groups of messages) should appear in the text.
- which groups of sentences correspond to text structure such as paragraphs and sections
- the discourse relations which hold between sentences or group of them

In the researches we described in the Introduction, Rhetorical Structure Theory (RST) [21] is used as a model to represent such a text structure. RST is one of the well-known models for text structure representation and is mainly used to represent coherence of texts. With RST we can decompose a text into sub-parts forming a hierarchical structure. Every sub-part has a relationship to another sub-part with one of the relation types (rhetorical relations). These relations form an overall coherence structure of the text.

For example, the representation of the following simple text is shown in Figure 2.1.

a) I like to collect old Fender Guitars.

b) My favourite instrument is a 1951 Stratocaster.

c) However, my wife does not like the guitars.

In the simple text, the sentence b connect with the sentence a in the meaning of the text. And the sentence c prefer to connect with the sentence a rather than

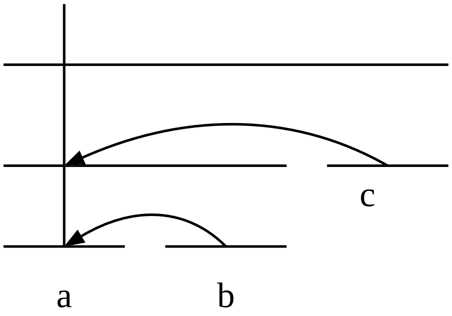


Figure 2.1. Representation for a simple text in RST

the sentence b. Such a meaning of the text is represented by the arcs in the Figure 2.1. The connection that is represented by an arc does not only denote the pair which has relation, but also holds one of the semantic types. In this case, the relation $b \rightarrow a$ holds "Elaboration", $c \rightarrow a$ does "Contrast".

The notion of rhetorical relation is a key concept in RST. Rhetorical relations specify the relationships that hold between messages or groups of message. In particular, RST claims that a small number of defined rhetorical relationships, such as Motivation, Contrast, and Elaboration, can be used to explain the relationships that hold within an extremely wide range of texts.

Marcu [25] proposes a comprehensive corpus analysis of cue phrases and develops new algorithms that identify discourse usages of cue phrases, divide sentences into clauses, and generate valid rhetorical structure trees.

### 2.1.3 Discourse Segment

The notion "discourse segment" must play an important role in identifying the coherence structure of a text, because it provides the basic subpart that corresponds with the main / sub topic of the text.

Texts frequently exhibits varying degrees of cohesion in different sections. For example, the start of a text cannot be cohesive with preceding sections, nor can the end exhibit cohesion with succeeding sections. In the middle of a text, however, the quantity of cohesion can vary greatly.

Morris and Hirst [27] introduce a method for finding such segments in text based on lexical cohesion. They called the manually built lists of related words lexical chains, which identify cohesions.

Hearst develops an algorithm that automatically assigns multiple topic categories to texts, based on the posterior probability of the topic given its surrounding words [7], which is a segmentation method relying on similarity between blocks of text based on vocabulary overlap. She also proposes a method to detect the discourse segments in the text using a variant of the TextTiling approach by herself [28].

However, there remains an open question for such algorithms: to what degree of segment leads to better information to text processing. Although Hearst found good agreement between the segments by the method and the segments identified by human judges, the length of the segment is section or chapter level in a long document.

## 2.2 Discourse Clues in Japanese

### 2.2.1 Clues for Cohesion and Coherence

Gross and Sidner [5] that explain some words in discourse are used to indicate changes in the discourse structure rather than to convey the information about the subject being discussed. Such words are called as cue words. Other researchers have examined the relationship between particular words and phrases and discourse structure as well.

Cue words play an important role in the discourse segmentation work by Hirschberg, Litman and their colleagues [8]. They present a number of cue words that indicate changes in discourse structure which they gleaned form various sources.

Passonneau and Litman [32] present an algorithm using cue words, pauses, and other surface cues to determine whether empirically validated discourse segment boundaries of a test corpus correlate with these linguistic devices. Passoneau and Litman claim that, although their algorithms do not perform as well as humans, the results suggested human performance could be achieved with additional knowledge. This implies that their algorithms are weak as an approach to discourse phenomena.

In Japanese, there are other well-known discourse phenomena. Many researches claim ellipses and the entity that is assigned as a topic role in the sentence play an important role in the cohesion and coherence structure of a text. In the following subsections, we introduce short summaries for such Japanese discourse cues [37, 17, 36].

### 2.2.2 Sentential Topic

The term "topic" has been used in either of the two apparently similar but actually very different meanings: what a passage or paragraph is about and what a sentence is about. The latter is what we deal with here.

Each sentence in any language has a topic for which the rest of the sentence is said or written. In English, for example, a topic is usually realized as a subject, or is sometimes introduced by *as to*, *when it comes to*, etc. In other words, a topic in English has no distinctive grammatical feature for its own. In Japanese, however, a topic is marked by one of the topic markers, the most typical of which is "wa." A phrase marked by "wa" is typically the subject of the sentence, but the subject marker in Japanese is "ga" and even an adverbial can be topicalized by "wa."

花子は本を読んだ。 (Subject topicalization)
(Hanako read a book.)

昨日は太郎が欠席だった。 (Adverb topicalization)
(Yesterday Taro was absent.)

昨日よりは気分がいい。(Adverbial topicalization)
(I am feeling better than yesterday.)

In addition, by using "wa" and "ga" in a sentence, we have a so-called Double Subject Construction, in which a topic is not an element topicalized from somewhere else but is an element functioning as a topic for the proposition in the sentence.

象は鼻が長い。
(An elephant has a long nose.)

東京は物価が高い。
(As for Tokyo, the price is high.)

物価は東京が高い
(As for the price, the Tokyo's is high.)

The functions of "wa" have been discussed extensively, but the point here is that a topic is related to the perspective the writer takes when writing a sentence.

### 2.2.3 Ellipsis

Another phenomenon relating to the writer's perspective is ellipsis. As is well known, Japanese permits an ellipsis, or omission, of phrases in a sentence rather freely, and ellipsis in Japanese is said to be heavily related to the contextual information. In other words, those contextually recoverable elements can be omitted quite freely. In the following examples, ϕ denotes the omitted element, usually called a zero anaphora.

太郎はできれば ϕするでしょう。
(If possible, Taro will do it.)

もしこの岩を持ち上げられるなら、ϕ動かしてもらえませんか。
(If you can hold up this rock, please move it.)

In the first example, ϕ is an ellipsis that refers to "太郎" (Taro) while in the second example, ϕ is an ellipsis that refers to "岩" (rock). An interesting point is that in Japanese, ellipsis is not only a possible operation but also a desirable operation for decreasing redundancy and keeping writer's perspective. So if ϕ in the latter example is substituted by the original phrase or a pronoun, then the sentence becomes rather awkward because it is unnecessarily redundant, and the phrase "this rock" is too much focused on.

もしこの岩を持ち上げられるなら、この岩を/これを動かしてもらえませんか。
(If you can hold up this rock, please move this rock/it.)

On the other hand, as is easily predicted, the excessive use of ellipsis causes unrecoverable ambiguity. So the following sentence is too ambiguous, though similar sentences are found everywhere in daily conversation.

もし持ち上げられるなら、できるでしょう。
(If you can carry, you can, in the sense that if you carry this rock, you will be able to beat that man.)

Thus, in analyzing the source text, it is important to find the omitted element and specify its antecedent, while in generating the summary, it is equally important to properly omit the redundant elements.

## 2.3 Operations in Summary Generation

### 2.3.1 Improve Cohesions in Extracted Sentences

One of the famoous previous works dealing with ways to produce more cohesive extracts is Paice [31]. As Paice pointed out a problem that computer produced extracts tend to suffer from a 'lack of cohesion'. We introduce, here, some related works that consider the improvements of cohesions in extracted sentences.

Mani and his colleagues address the problem to improve extracted sentences using revising [20]. When humans write a text, they frequently revise it for refinement. This idea is the basis for their work. In order to realize such a revising system on a machine, they assume the following three types of operations.

- Elimination
  Elimination operation eliminates constituents from a sentence. This includes the elimination of sentence-initial PPs and adverbial phrases.

- Aggregation
  Aggregation operation combines constituents from two sentences, at least one of which must be a sentence in the extracts, into a new constituent which

is inserted into an extracted sentence. For example, if a relative clause is added to an extracted sentence, the addition of relevance information is one of the aggregation operations.

- Smoothing
  Smoothing operation applies to a single sentence, performing transformations for obtaining more compact, stylistically preferred sentences. For example, in the smoothing operation, a coordinate expression is removed.

They manually make revision rules to realize such operations in the automated revising system for extracted sentences. In an evolution based on Questions and Answers of human subjects, they show that revising to the extracted sentences by such rules contributes to improve the readabilities of summaries.

Nanba and Okumura [29] investigate how people revise extracts of Japanese articles to produce more readable ones. They classify the factors that causes such revisions into five categories, most of which are related to cohesion.

- lack of conjunctive expressions / presence of extraneous conjunctive expressions.
  Extracted sentences do not always adjoin each other in the original text. Because conjunctive expression of the extracted sentences is added for the cohesion in the original text, such expression is added or eliminated according to the new cohesion among the set of the extracted sentences.

- syntactic complexity
  Long sentence tends to have complex syntactic structure. In the revision, such a sentence is sometimes divided into two simpler sentences.

- redundant repetition
  In the original text, the elements such as reference expressions and eliminations avoid redundant repetitions in the sentences that are close together in the original text. Extraction breaks such cohesion so that elements of cohesion among the extracted sentences should be reconstructed.

- lack of information
  Each sentence in a text does not give enough information for the all of its

constituents within the sentence. If extraction does not extract sentences in which relevant information to the constituents of extracted ones, such information should be added in revision.

- lack of adverbial particles /presence of extraneous adverbial particles.
  In Japanese, adverbial particle ' も' emphasises that a sentence gives additional information to the entity which is marked by the particle. However, extraction does not always extract the base information for the additional information. In the case, the adverbial particle should be eliminated.

Based on the result of the investigation, they devise revision rules for each factor and partially implemented a system that revises extracts. The following four modules are implemented.

- Deletion of conjunctive expressions
  They prepare a list of 52 conjunctive expressions, and make it a rule to delete each of them whenever the extract does not include the sentence that expression is related.

- Omission of redundant expressions
  If subjects (or topical expressions marked with topical postposition 'は') of adjacent sentences in an extract are the same, the repeated expressions are considered redundant and are deleted.

- Deletion of anaphora
  They implement a rule with ad hoc heuristics to teat anaphora and ellipsis: If an anaphora appears at the beginning of a sentence in an extract, its antecedent must be in the preceding sentence. On the other hand, if that sentence was not in the extract, the anaphor was deleted.

- Supplement of omitted subjects
  If a subject in a sentence in an extract is omitted, the revision rule supplements the subject from the nearest preceding sentence whose subject is not omitted in the original text.

They evaluate the outputs of the readability by comparing judgements between the automated revised and the original extracts. As result, they report the experimental revising system can improve the readability of extracts.

### 2.3.2 Decomposition of Summary Process

Recently, a number of researchers have started to address the model of generating coherent summaries instead of revising extracted sentences.

Jing and McKeown present the cut and paste model, which is assumed as a new computational model for generating a summary [10, 13, 12]. In their model, the summarization process is divided into six operations, which derived from their manually investigation on 30 human-generated summaries. Those operations can be used alone, sequentially, or simultaneously to transform extracted sentences.

- Sentence reduction
  Removes extraneous phrases from an extracted sentence.

- Sentence combination
  Marge material from several sentences. It can be used together with sentence reduction.

- Syntactic transformation
  In both sentence reduction and sentence combination, syntactic transformations may be involved. For example, the position of the subject in a sentence may be moved from the end to the front.

- Lexical paraphrasing
  Replace phrases with their paraphrases. For instance, the summaries substituted *point out* with *note*, and *fits squarely into* with a more picturesque description *hits the head on the nail*.

- Generalization or Specification
  Replace phrases or clauses with more general or specific descriptions.

- Reordering
  Change the order of extracted sentences. For instance, an ending sentence in an original text is sometimes placed at the beginning of the summary.

Jing and McKeown report that 19% of sentences in their analyzed summaries are written from scratch instead of applying such operations. In the 6 operations above, they regard sentence combination and sentence reduction as Major components in a summarization system.

Furthermore, Jing and McKeown show some directions to developing summarization system using cut and paste techniques. Decomposition program is one of the fundamental components of such directions. They developed a decomposition program in order to automatically analyze a large quantity of human-written abstracts. The automatic decomposition helps building large corpora for studying sentence reduction and sentence combination.

Some researchers have already started to build such corpora from the pairs between summary and the original text and to acquire a piece of knowledge from the corpora [24, 14, 15]. However, the fundamental investigations on the process of the humans summary generation are not enough to build a large corpora in Japanese. We will address this problem in Chapter 4.

# Chapter 3

# Relationship between Text Structure and Summaries

## 3.1 Introduction

In the research of automated summarization, some researchers such as Ono et al. [30] and Marcu [23] use tree structure models to represent text structure and to select important sub-parts in texts. By doing this, they exploit relations among the sentences in a text.

Although text structure plays an important role in developing an automated summarization system, there is no concrete model to make a summary through a representation of text structure. Moreover, what representation is suitable for generating a summary is not clear. We investigate mainly two problems in this chapter. One is to estimate the consistency of human analysis on text structure. The other is to investigate what kind of role the text structure plays to generate a summary. First of all, we set up an experimental scheme to analyze text structure. We ask human subjects to produce summaries of texts, where the structure of the source texts is analyzed in advance. After those preparations, we examine how the alignment is done between the sentences in the source text and the sentences in its summary. By means of the structure and the alignment analysis, we confirmed that the text structure plays an important role in generating a summary. In particular, pairs of adjacent sentences that have a direct relationship in the text structure exhibit a special role in both of the stages; in the analysis of text structure and in the generation of summaries. We could regard such a relationship as the clues for sentence combination, which is one of the crucial operations for a human to produce a summary. To investigate the characteristics of the relationship further, we apply a machine learning method using some linguistic features and make it sure that the clues are to act as the trigger for sentence combination.

## 3.2 Analyzing Text Structures

### 3.2.1 Coding Scheme for Text Structures

In general, properties of a text are classified in terms of the linguistic notion of cohesion and coherence. According to Halliday and Hasan [6], the notion of cohesion is closely related with linguistic cues such as anaphora, ellipses, conjunctions, lexical relations, etc. Those linguistic cues contribute to create semantic connectedness in a text. Compared to cohesion, coherence is related to more abstract semantic structure of a text. Although a number of models for text structure analysis have been proposed, there is not yet a specific model that can provide us with useful information for generating summaries.

In the previous researches we described in the introduction, Rhetorical Structure Theory (RST) [21] is used to represent the text structure. RST is one of the well-known models for text structure representation and is mainly used to represent coherence of texts. With RST we can decompose a text into sub-parts forming a hierarchical structure. Every sub-part has a relationship to another sub-part with one of the relation types (rhetorical relations). These relations form an overall coherence structure of the text.

We propose a coding scheme with which a human subject analyzes text structure in order to investigate how well the representation of text structure works for summarization. In the coding scheme, we modify RST in two ways as described below to help a subject to code the text structure.

First, since the original RST proposes more than 20 types of rhetorical relations, a human subject often finds difficulty in selecting a proper relation between sentences. Furthermore, as Moore and Pollack [26] point out, a pair of sentences sometimes inherently has a multiple-analysis. To avoid such complexities we introduce two fundamental relations in terms of the degree of importance between a pair of sentences. One is the relation where there is unclear distinction in degree of importance between the pair. The other is the relation where the pair has relative degree of importance. We assume this simplification does not contradict with other's work in automated text summarization using text structure.

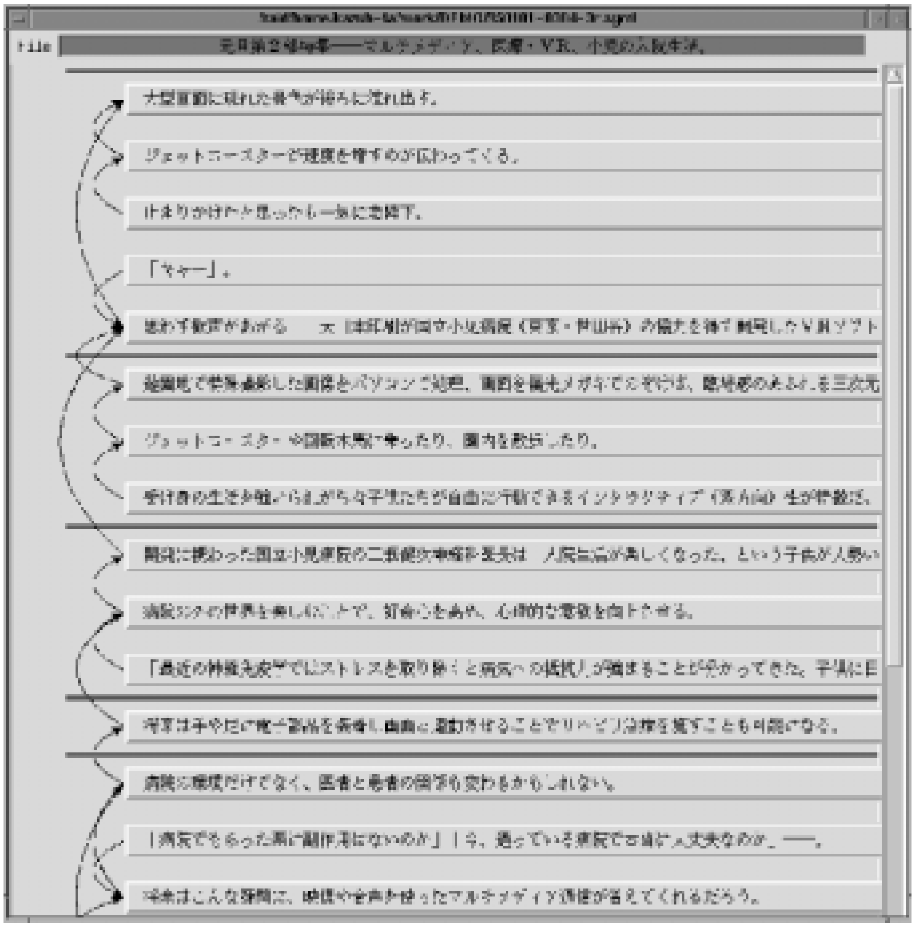
Second, we define a sentence as the elementary unit of text structure. However, in the original RST, a more fine-grained fragment (which usually corresponds to a clause) is considered as the elementary unit. Since the coherence between sentences is our main interest in this chapter, we currently assume a sentence as the elementary unit.

We developed an annotating tool for text structure analysis as shown in Figure 3.1. It helps the subject to annotate the texts based on simplified version of RST described above. On the screen-shot, boxes correspond to the sentences in the annotating text and arrows stand for coherence relations between sentences. In practice, a subject simply selects a tag through a graphical user interface(GUI) of the tool. For each sentence $S_t$ in the annotating text, a subject acts as follows:
1) selects the most relevant sentence $S_t$ while $S_t$ must be more important than or equal to $S_s$ with regard to the meaning of the whole text. ($S_s$ is the root of the text structure if $S_s$ does not have such a sentence in the text. The root must be only a single sentence in a text.)
2) selects the relation type for the pair of sentences ($S_t$ and $S_s$).

### 3.2.2 Evaluation of the Analyzed Text Structures

We let three human subjects analyze text structures using our coding scheme described in the previous section. Note that the analysis is done by a different group of subjects from ones that produce summaries. We use 32 Japanese report articles from Nihon-keizai-shinbun (Japanese financial newspaper) in 1995. The total number of sentences in the articles is 500.

To observe the overall tendency among agreement of subjects' analysis, we use the Kappa coefficient, which is used to assess agreement in the area of behavioral science. Carletta [2] introduces and discusses the use of it as an agreement measure in the discourse analysis. The Kappa coefficient measures pairwise agreement



元日第2部b面——マルチメディア、医療・VR、小児の入院生活。

大型画面に現れた景色が後ろに流れ出す。

ジェットコースターが速度を増すのが伝わってくる。

止まりかけたと思ったら一気に急降下。

「キャー」。

思わず歓声があがる　　文 日本IBMが国立小児病院（東京・世田谷）の協力を得て開発したVRソフト

遊園地で実際撮影した画像をパソコンで処理、画面を偏光メガネでのぞけば、臨場感のあふれる三次元

ジェットコースターや回転木馬に乗ったり、園内を散歩したり。

受け身の生活を強いられがちな子供たちが自由に行動できるインタラクティブ（双方向）性が特徴だ。

開発に携わった国立小児病院の二宮耕次神経科医長は「入院生活が楽しくなった、という子供が大勢い

病院の外の世界を楽しむことで、好奇心を高め、心理的な意欲を向上させる。

「最近の神経免疫学ではストレスを取り除くと病気への抵抗力が強まることが分かってきた。子供に目

将来は手や足に電子部品を装着し自由に運動させることでリハビリ治療を施すことも可能になる。

病院の環境だけでなく、医者と患者の関係も変わるかもしれない。

「病院でもらった薬に副作用はないのか」「今、通っている病院で本当に大丈夫なのか」——。

将来はこんな疑問に、映像や音声を使ったマルチメディア通信が答えてくれるだろう。

Figure 3.1. Tool for annotating a text structure

among a set of coders making category judgments and is defined as:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of frequency that the subjects agree and $P(E)$ is the proportion of frequency that they agree by chance. Intuitively, the measure shows the degree of agreement adjusted by the agreement by chance.

In our experiment of text structure analysis by human subjects, P(A) stands for the agreement ratio of selected sentences. [1] We obtain Kappa coefficient of 0.58 (P(A) = 0.63, P(E)=0.11). In the evaluation of agreement using Kappa coefficient, there is a guideline that is used to evaluate the degree of reliability of the agreement in a coding scheme [3]:

- 0 < Kappa < 0.2 is regarded as “slight” agreement
- 0.21 to 0.40 as “fair”
- 0.41 to 0.60 as “moderate”
- 0.61 to 0.80 as “substantial”
- and 0.81 to 1.0 as “near perfect”

According to the guideline, our coding scheme is evaluated as “moderate” level, close to the “substantial” level with the overall agreement.

Then, we observe the tendency of the assignments of relation types between pairs of sentences. We found that agreement rate of the assignment for the relation between adjacent pairs of sentences is much higher than that of the other pairs. The difference of tendency among subjects’ assignment is observed when it is evaluated in terms of the distance between dependent sentences as shown in Table 3.1. The distance of relation is defined as the relative distance between the dependent sentences. If a sentence relates with the preceding sentence, the distance of relation is 1. We define the ratio as follows:

[1] Unfortunately, we have not discovered the role of two relation types that we distinguished in the text structure on the stage of summary generation. Since we need to investigate the problem more in the further work, we concentrate on the selection of sentence pairs on the analysis of the text structure in this chapter.

Table 3.1. Agreement ratio against Distance

|  | at least 1 | majority |  |
|---|---|---|---|
| n | subj. selected | selected | ratio(n) |
| 1 | 352 | 293 | 0.832 |
| 2 | 113 | 48 | 0.425 |
| 3 | 53 | 21 | 0.396 |
| 4 | 36 | 14 | 0.389 |
| n≥5 | 85 | 29 | 0.341 |

$$ratio(n) = \frac{\begin{array}{l}\text{the number of pairs that have distance } n \\ \ \ \text{(agreed by the majority of subjects)}\end{array}}{\begin{array}{l}\text{the number of pairs that have distance } n \\ \ \ \text{(assigned by at least one subject)}\end{array}}$$

From Table 3.1, compared with other distances the relations of distance = 1 agree more frequently. This indicates that even a human has difficulty in judging the relation between sentences when the distance is two or more. On the other hand, when two related sentences are adjacent, the judgment by humans is significantly more accurate.

Figure 3.2 shows an example of text structure in the representation of RST. In the figure, a number corresponds to the position of a sentence and an arrow corresponds to a relationship between sentences. To summarize the result using the representation, the relationships between adjacent sentences such as 2 → 1 agrees more often by human subjects compared with the relations like 4 → 1 and 8 → 4.

In the rest of this chapter, we regard the coherence structure that are determined by the majority of the subjects as the text structure.

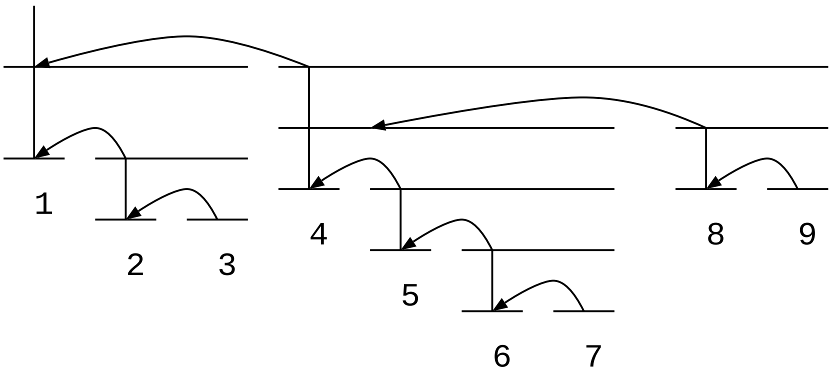


Figure 3.2. Representation of a text structure in RST

## 3.3 Analysis on Human-generated Summaries

### 3.3.1 Making Summaries

We ask two Japanese native speakers to summarize 20 texts. They are educated in literature, but are not professional summarizers. We ask them to generate summaries up to the length of 40% of the source texts. We pick up the 20 source texts to be summarized from the set of texts of which structure have been analyzed beforehand according to the experimental analyzing scheme described in section 3.2.1. The group of human subjects who made summaries are different from the group of the subjects who analyzed the text structures. Since we assume that human can generate summaries without explicit information of text structure and word importance, we removed paragraph breaks and the titles from the texts in the experiment.

In the summarization task, we give two instructions:

- The summary should keep the overall story of the source text and author's main opinion.

- If the summarizer uses proper nouns in the summary, the form of the proper nouns should be kept as the originals.

Basic information about the source texts and their summaries are shown in Table 3.2. In the table and the rest of this chapter, we refer to the summaries made by

Table 3.2. Source Texts and Their Summaries

| | Source | Summaries | |
|---|---|---|---|
| (Average number) | Texts | A | B |
| # Characters | 882.6 | 342.0 | 320.4 |
| # Sentences | 16.1 | 7.2 | 6.9 |

summarizer A as *Summary A* and those of summarizer B as *Summary B*.

### 3.3.2 Operations for Generating Summary

In order to analyze the human behavior in generating summaries, it is important to know the alignment between the sentences in the human-generated summary and the corresponding source sentences that have been used to generate the sentences. There are some related work of the summarizing task. For example, Jing and McKeown [13] identified 6 operations in human summary generating process. In this chapter, we focus on the operations where a summary sentence is generated from one sentence or more and classify the operations into following two types.

1. sentence reduction: the summarized sentence is generated from exactly one source sentence.

2. sentence combination: the summarized sentence originates from two or more source sentences.

We manually perform the alignment. For each summary sentence, two human subjects select which sentences in the source text are used in summarization. Only case where two subjects agreed with the analysis of the alignment, we accept the human analysis as valid. For other cases, we accept the most similar sentence in the source text to the summary sentence by using word-based cosine similarity.

Table 3.3 shows the number of two types of operations in summary generation. Comparing the number of sentence reduction with that of sentence combination, the former has more examples than the later in both cases. However, 103 source sentences are used to create 49 combined sentences in summary A, and 58 source

Table 3.3. The number of sentence reduction and sentence combination

| Operation Type | Summaries | |
|---|---|---|
| | A | B |
| sentence reduction | 94 | 110 |
| sentence combination | 49 | 27 |

sentences are used to create 27 combined sentences in summary B. This shows that the number of the sentences used in sentence combination with respect to the total number of source sentences for summary should not be ignored.

### 3.3.3 Coherence Structure and Sentence Combination

From the alignment result, we notice an explicit tendency in the sentence combination operation on the source text. Table 3.4 shows the number of sentence pairs combine into summary sentences. In both set of summaries, the summary sentences that are generated by sentence combination are mostly the adjacent sentences. However, even if a pair of adjacent sentences in a source text is likely to combine, the pair is not always combined to the summary sentence. We assumed the clues for sentence combination is adjacency relation in the coherence structure. This assumption comes from the fact that the adjacency relationship in coherence structure is comparatively easy for humans to analyze as we described in section 2. The bracketed figures on the upper line in Table 3.4 shows the number of examples that have coherence relation. This result shows that the text structure can act as the clues for sentence combination.

On the other hand, the number of examples of sentence combination of non-adjacent sentences is not enough to draw a conclusion. We think that the structure between non-adjacent sentences involves more complex mechanism than adjacency relation.

Table 3.4. Sentence Position in Source Texts and in Text Structure

| Position of pairs | Summary A | Summary B |
|---|---:|---:|
| adjacent | 38(34) | 19(18) |
| non-adjacent | 11 | 8 |
| Total | 49 | 27 |

Table 3.5. Features for Characterizing a pair between adjacent sentences

| Feature Categories | Features | ID |
|---|---|---|
| Syntactic Features | cue words of $S_i$ | CUE |
| | predicate type of $S_{i-1}$ | PRD0 |
| | predicate type of $S_i$ | PRD1 |
| | topic marker type of $S_i$ | TPC |
| | omission of topic or subject of the $S_i$ | OMIT |
| Semantic Features | $S_i$ introduce a new proper noun<br>or $S_i$ refers the proper noun in $S_{i-1}$ | NEWT |
| | Character based similarity between $S_i$ and $S_{i-1}$ | SIM |

## 3.4 Clues for Coherence between Adjacent Pairs

As we described in Section 3, we discover that the coherence relationship between a pair of adjacent sentences plays an important role in summary generation. In this section, we investigate clues for those relationships. Our investigation consists of two steps. The first step is to see how well a machine predicts whether a pair of adjacent sentences has a coherence relationship or not. The second step is to see whether the clues of the relationships between adjacent sentences work as the clues for the sentence combinations as well.

### 3.4.1 Clues for Relationships between Adjacent Sentences

In order to investigate the clues for the relationship between pairs of adjacent sentences, we apply the machine-learning program C4.5 [33]. C4.5 is a decision tree learning program that acquires general rules from the training examples that consist of features and the target class. In our learning task, the target class is assigned using the relations between the pairs of adjacent sentences in a coherence structure described in section 2.3. Suppose $S_{i-1}$ and $S_i$ are a pair of adjacent sentences in a text. If $S_i$ has relationship to $S_{i-1}$ in the coherence structure, the pair of adjacent sentences classified as "yes", and "no", otherwise.

We arrange the information of an example into the sets of features as shown in Table 3.5. Since the relations in coherence structure are abstract and complex, we use not only syntactic information, but also information that influences the meaning. In practice, our features can be divided into two categories; syntactic features and semantic features.

Syntactic features represent the characteristics of a sentence using the result of syntactic structure analysis. We used automated word dependency structure analyzing program developed by Fujio et al.[4]. Japanese dependency structure is usually defined in terms of the relationship between phrases called 'bunsetu.' The relationships reflect the underlying syntactic structure of a sentence. Bunsetu is a segment that consists of one or more words and that includes a head word. Fujio's program outputs not only the syntactic relationship between Bunsetus but also the pos (part-of-speech) tags of the words. Since the reliability of the program is not perfect yet, we manually correct the result of the dependencies when the result has some errors. Value of the syntactic features is assigned using the analyzed dependency structure.

The value of CUE feature is assigned based on the type of conjunctive expressions at the beginning of the corresponding sentence (e.g., *For example, However, Therefore,* etc. in English). Since conjunctive expressions provide cohesive connectivity between adjacent sentences in general, we divide conjunctive expressions into 10 types according to the function of connectivity (e.g. Exemplify, Contrast, Restating, etc.). The features PRD0 and PRD1 represent the type of the predicates of $S_{i-1}$ and $S_i$. Each predicate type of the sentence is classified according to its modality or its tense. These two features are expected to indicate the writer's attitudes in the sentence. TPC feature represents the type of topic marker in the sentence. In Japanese, the grammatical role such as topic and subject is marked by particles called postpositions. We distinguish the grammatical role using the analyzed dependency structure. OMIT feature shows whether the topic or the subject of the sentence $S_i$ is present or not. The absence will show the stronger cohesion between $S_i$ and $S_{i-1}$. Some researchers such as Walker et al.[37] claim that the entity that marked by topic maker and its omission play an important role in the discourse. Note that the values of TPC and OMIT are hard to be assigned only by the pos information. Therefore, we use the analyzed dependency structure to assign the features.

On the other hand, semantic features represent the relevancy between the contents described in the two sentences. In order to represent the meaning of the sentences, we use two features as an approximate relevancy between the sentences: NEWT and SIM features. NEWT feature represents whether $S_i$ introduces new proper nouns as a topic or not. We assign four different values to this feature based on four cases of $S_i$ described bellow [2] :

1) the topic/subject includes the proper nouns that $S_{i-1}$ does not include.

2) the topic/subject includes referring expressions that refer to the proper nouns of $S_{i-1}$.

3) a part of $S_i$ (except for the topic/subject) refers to the proper noun of $S_{i-1}$.

4) $S_i$ satisfies none of the above cases.

Proper nouns and expressions referring the proper nouns in the text are manually annotated. In this experiment, we annotated only person names, place names and organization names as proper nouns and referring expressions are restricted to pronouns and nouns that refer to the proper nouns without any inference. SIM feature uses similarity between the subpart that expresses topic/subject in $S_i$ and the whole $S_{i-1}$. The similarity is calculated by character (Kanji) based cosine similarity. If topic/subject is absent in $S_i$, the value of the SIM is assigned 1.0 (i.e. SIM includes a piece of information of OMIT feature). We expect that

[2] The notation of 'topic/subject' in this chapter stands for the subpart of a sentence. The subpart contains not only the entity marked as topic or subject, but also the phrases modifying the entity.

the SIM feature gives the simple approximation for NEWT features. We, however do not claim that our features set is an exhaustive one.

We evaluate the learned decision tree using the "leave-one-out" cross validation as follows. For every example $x_i$ in the training example set, a decision tree learns from all the examples except the $x_i$ and the learned tree is evaluated by $x_i$. Table 3.6 shows the results in terms of precision, recall and accuracy that are defined by following equations.

$$
\begin{aligned}
&Precision = \\
&\qquad \frac{\text{\# examples decision-tree classified correctly}}{\text{\# examples decision-tree classified}} \\
&Recall = \\
&\qquad \frac{\text{\# examples decision-tree classified correctly}}{\text{\# examples human classified}} \\
&Accuracy = \\
&\qquad \frac{\text{\# examples decision-tree classified correctly}}{\text{\# examples}}
\end{aligned}
$$

We assume a system which always classify examples as "yes" as a baseline, whose accuracy is 0.626. Comparing with the baseline, from the obtained results, we can conclude that the features are able to predict whether a sentence has coherence relationship to the preceding sentence.

Moreover, we also evaluate the impact of each individual feature for the learned decision tree to classify the example into the target class correctly. We remove each feature in turn from the set of features listed in Table 3.5 and apply the same process to construct the decision tree with the reduced member of the features. The accuracy of each learned decision tree in shown in Figure 3.3. The figure shows when CUE, PRD1 or SIM feature is removed from the original features, the accuracy decreases significantly. The result shows that these three features have the ability to capture the characteristics of the relationship between the adjacent sentences. Although we carefully designed NEWT in the semantic features, the feature does not show good effect on the accuracy.

Table 3.6. Evaluation for Learned Decision Tree

| | target class | |
|---|---|---|
| | yes | no |
| Precision | 0.715 | 0.638 |
| Recall | 0.850 | 0.434 |
| Overall Accuracy | 0.697 | |

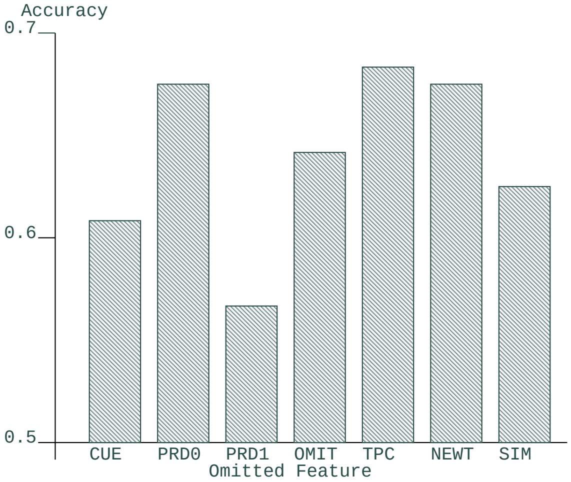


Figure 3.3. Accuracy of the decision tree for classifying the relations between adjacent sentences without each feature

### 3.4.2 Discussion on the clues for sentence combination

In this section, we want to verify that the clues for relations between adjacent sentences are also the clues for sentence combination. We conduct an experiment by using the reliability of prediction for coherence relationship between pairs of adjacent sentences. The reliability is provided by the learned decision tree and takes value from 0 to 1. The higher the reliability is, the more likely the sentence relates to the preceding adjacent sentence. In practice, we calculate the reliability as follows: A leaf node of the learned decision tree is constructed to classify the training examples. Note that the leaf node does not always classify the "real" training examples when the decision tree is pruned. We regard the confidence-ratio of classifying the training examples with the leaf node of the learned decision tree as the reliability. For example, if the all of the corresponding training examples to a particular leaf node are classified into class "yes", the reliability of the leaf node is 1.0.

We compare the average reliability for the pairs of adjacent sentences that have coherence relationship with that of the pairs of sentences in which sentence combination takes place. As shown in Table 3.7, both the average reliability and its standard deviation(stds) give similar tendencies between the pair of sentences that have coherence relationship and the pair of sentences where sentence combination occurs. Thus, the clues for sentence combination characterized by our features are quite similar to the ones for pairs of adjacent sentences with coherence relation.

In this section, we show that the strong relationship between adjacent sentences give us some hints to understand the operation of sentence combination. The relationships between adjacent sentences must represent the relevancy between the contexts or meaning of the corresponding sentences. Since the sentences produced by the operation of sentence combination are more constrained in terms of the relevancy, the investigation of the sentence combination will help us understand how to represent the coherence structure. Practically, the strength of coherence relation triggers the sentence combination operation in generating summary.

Table 3.7. Averaged Reliability for Pair of Adjacent Sentences

| type of the pair of sentences | Reliability | |
|---|---|---|
| | Ave. | stds |
| coherence relation | 0.631 | 0.301 |
| sentence combination | 0.615 | 0.295 |

## 3.5 Conclusion

In this chapter, we investigate human-generated summaries from the point of view of text structure. Even if human generates a summary without explicit discourse clues (such as paragraph breaks), the coherence structure analyzed in this paper can represent implicit clues for automated summarization. Our conclusion includes the followings.

- In analysis of coherence structure of text, even a human has difficulty in judging the relation between sentences when the related sentences stay apart from each other. On the other hand, when two related sentences are adjacent, the human judgment is far more accurate.
- Human summary generation is based on two types of operation; sentence reduction and sentence combination. An existence of relationship between the pair of adjacency sentences is a trigger to determine whether the operation of sentence combination is activated or not.
- Coherence relationships between adjacent sentences are automatically identified using features. We confirm that the features characterizing the relation can also characterize the occurrence of sentence combination.

Our work provides new perspective for automated summary generation. The perspective has three steps. At the first step, we analyze the strength of coherence relation between every adjacent pair in its source sentences. Second, we combine strongly related adjacent pairs to a new summary sentence. Finally, we reduce redundant clauses and words from the combined summary sentences. Based on this experimental results, we are motivated to develop a full-fledged summary generating system in the future.

# Chapter 4

# Analysis on Human-generated Summaries

In the previous chapter, we manually performed the alignment. For each summary sentence, two human subjects made clear which sentences in the source text were used in summarization. In this chapter,we investigate the manually generated summaries in detail. In doing so, we introduce an automated alignment method based on dependency structure analysis. The method detects not only one-to-one sentence alignment but also one-to-many sentence alignment.

## 4.1 Operations for Generating Summary

In order to analyze human behavior in generating summaries, it is important to know the alignment between the sentences in the human-generated summary and the corresponding source sentences that have been used to generate the summary sentences. There are some related work in the summarizing task. For example, Jing and McKeown [13] identify 6 operations in human summary generating process. In this dissertation, we focus on the operations with which a summary sentence is generated from one or more sentences and classify the operations into two types.

1. sentence reduction: the summarized sentence is generated from exactly one source sentence but in a reduced form

Table 4.1. Basic Information about the Summary Data

| | # Sentences |
|---|---:|
| Summarizer A | 763 |
| Summarizer B | 773 |
| Summarizer C | 931 |

2. sentence combination: the summarized sentence originates from two or more source sentences

## 4.2 Data

The target summaries in this chapter are summaries of newspaper editorials. In comparison with the summaries of articles in newspaper that we use in Chapter 3, editorials are longer in length and have richer and more complex contents. We ask three Japanese native speakers to summarize 90 editorials. The length of the summaries is supposed ti be up to the 40% of the source texts.

The conditions to summarize a text are the same as what we give in Chapter 3 as follows.

- A summary should keep the overall story of the source text and the author's main opinion.

- If the summarizer uses proper nouns in a summary, the form of the proper nouns should be kept as the originals.

Since we assume that humans can generate summaries without explicit information of text structure and word importance, we removed paragraph breaks and the titles from the texts in the experiment.

The numbers of sentences that each summarizer summarize are shown in Table 4.1. The total number of sentences in 90 editorials is 2869.

In 2467 summary sentences, only 692 sentences keep the original form. The remaining 1775 sentences have surface forms that are changed with a certain process of summary generation.

## 4.3 Automated Alignment

To investigate the manually generated summaries, we propose an automated alignment algorithm that aligns a summary expression with the corresponding original expression in the source text.

A summary sentence is not always generated by only one original sentence. As we mentioned in Chapter 3, when sentence combination is applied to generate a summary sentence, more than one original sentences needed to be combined into the summary sentence. Moreover, expressions in the summary sentences are sometimes different from the original ones. We take such complexities into consideration to design our alignment algorithm.

Figure 4.1 shows the overview of the algorithm. In the following subsections, we describe each process in the Figure 4.1 in detail.

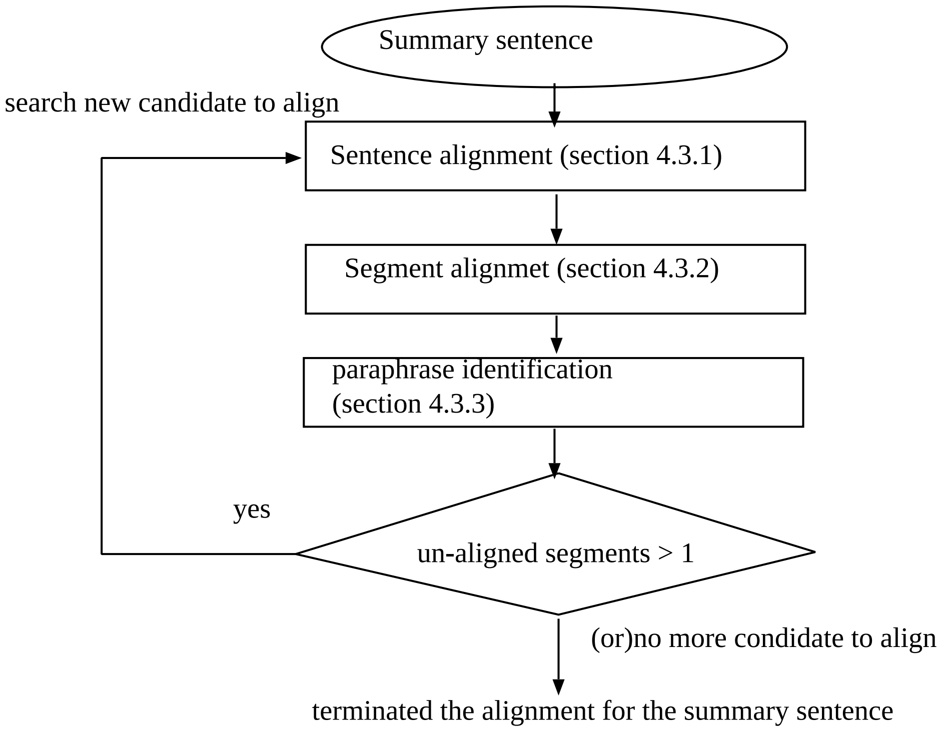

Summary sentence
search new candidate to align
Sentence alignment (section 4.3.1)
Segment alignmet (section 4.3.2)
paraphrase identification
(section 4.3.3)
yes
un-aligned segments > 1
(or)no more condidate to align
terminated the alignment for the summary sentence

Figure 4.1. Alignment Algorithm

### 4.3.1 Sentence Alignment

We assume that the operation "sentence combination" is applied in the summary generation. We take this assumption into consideration to design the alignment algorithm.

The dependency structure represents the syntactic structure of sentences. In a dependency structure, the fundamental unit is "bunsetu segment", which is the base phrases used as the basic units in the syntactic structure of Japanese sentences. In this paper, segments are referred to as *bunsetsu segments* or simply segment.

The dependency structure represents the modification relationship between segments. Each segment, except for the final segment in the sentence, modifies one of the segments in the sentence. The modification relationship of segments do not cross each other and always go from left to right in a sentence. From the definition, a representation of the dependency structure of a sentence forms a tree structure, in which a node represents a segment and the last segment in a sentence forms the root of the tree. An example of the dependency structure of a sentence is shown in Figure 4.2.

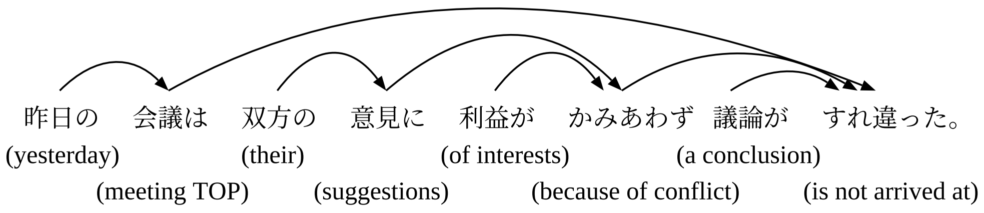


Figure 4.2. An Example of Dependency Analysis

The dependency structure has the property that hardly changes even if a word order is changed in the summary sentence. For an example, the following sentences have the same relationship between the segments:("花を" →" あげる", "彼女に" →" あげる") and describe the similar meaning while their surface orders are different.

彼女に 花を あげる。
(to her) (the flower) (give)
花を 彼女に あげる。
(the flower) (to her) (give)
彼女に なんとしてでも 花を あげる。
(to her) (at any cost) (the flower) (give)

This property works beneficially in aligning a summary sentence to the corresponding source sentences in the original text if the surface form of the summary sentence is changed.

Our algorithms is based on paths in dependency trees. In our algorithm, a path is defined as a list of nodes from a leaf to the root in a dependency tree. When a tree consists of only one node, we do not regard it as a tree. The dependency structure of a sentence is analyzed by the dependency structure parser CaboCha [16]. The example sentence in Figure 4.2 has the sets of paths as follows:

昨日の→会議は→すれ違った
双方の→意見に→かみあわず→すれ違った
利益が→かみあわず→すれ違った
議論が→すれ違った

We extract all paths in a summary sentence and the all of sentences in the source text.

Now we are to explain the method to align a summary sentence $s$ with the corresponding sentence $a$ in the source text. Assume the set of paths in $s$ to be $P_s$ and the set of all paths of the sentences of the source text to be $P_{cand}$. The most similar sentence to $s$ in the source text is decided to be the corresponding sentence $a$. The longest common subsequences (LCS) are evaluated for the all combinations of the paths between the $P_s$ and $P_{cand}$ to search the similar candidate to $s$. When a sentence has the path whose length is $j$, the sentence is defined as the corresponding sentence $a$. The following formula gives $j$.

$$j = argmax_{x \in P_s, y \in P_{cand}}(LCS(x, y))$$

In the formula, LCS(x,y) is a function that returns the length of LCS between the path x and the path y.

Table 4.2. Table to compute LCS

| p | $o_1$ | $o_2$ | $\ldots$ | $o_j$ | $\ldots$ | $o_n$ |
|---|---|---|---|---|---|---|
| $s_1$ | $c_{1,1}$ | $c_{1,2}$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |
| $s_2$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |
| $\vdots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |
| $s_i$ | $\ldots$ | $\ldots$ | $\ldots$ | $c_{i,j}$ | | |
| $\vdots$ | | | | | | |
| $s_m$ | | | | | | $c_{m,n}$ |

The LCS is calculated with DP matching. Given the comparing pair of paths $[s_1, s_2, \ldots, s_m]and[o_1, o_2, \ldots, o_n]$, a table exemplified in Table 4.2 is created in order to calculate LCS. Elements $s_i$ and $o_j$ of the path, are referred to as a segment of the path. In the table, each cell represents the cost to make the corresponding pair $(s_i, o_j)$ correspond. Each cell of the table is calculated in the order: $c_{1,1}, c_{1,2}, \ldots, c_{1,n}, c_{2,1}, \ldots$ and then the path in the table from $(s_1, o_1)$ to $(s_m, o_n)$ shows the total cost to make the LCS between the path. The cost $c_{i,j}$ is calculated with the following formula.

$$c_{i,j} = min(c_{i-1,j} + 1, c_{i,j-1} + 1, c_{i-1,j-1} + x(i,j))$$

$$x(i,j) = \begin{cases} 0 & \text{if } s_i = o_j \\ 1 & \text{if } s_i \neq o_j \end{cases}$$

In the function $x(i,j)$, the comparison between $s_i$ and $o_j$ is that of segments. The comparison between segments allows the changes of the segment as follows.

- compression of a long compound noun into a shorter form
- changes of postposition to mark its constitute as a topic
- inflectional change of verbs, adjectives and nouns that are used as a verb

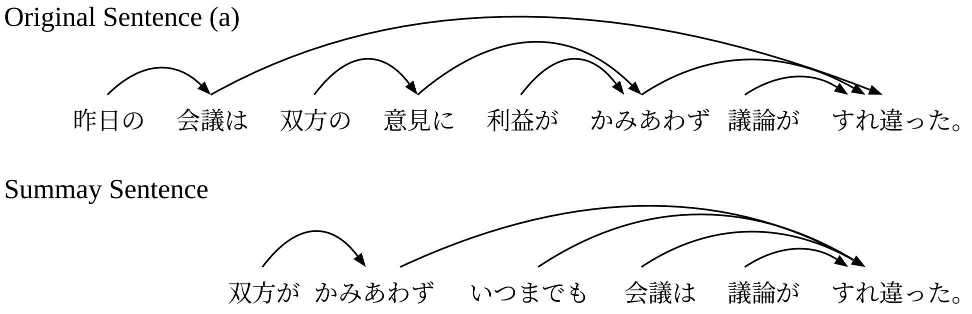


Figure 4.3. Sentence Aligned Pair

### 4.3.2 Segment Alignment

Following the sentence alignment, we apply the segment alignment. On the process of the segment alignment, we use the information of the set of LCSs which we use in the sentence alignment. The LCS that are included in the set are restricted to the one whose length is 2 or more, because we manage the source expression which has 2 or more related segments. For example, the sentence aligned pair in Figure 4.3 has the set of paths as follows:

昨日の→会議は→すれ違った
双方の→意見に→かみあわず→すれ違った
利益が→かみあわず→すれ違った
議論が→すれ違った

On the other hand, that of sentence (a) is as follows:

会議は→すれ違った
双方が→かみあわず→すれ違った
いつまでも→すれ違った
議論が→すれ違った

In these sets, we find the following set of LCSs are common in the above two sets:

会議は, すれ違った
かみあわず, すれ違った
議論が, すれ違った

We align each segment of the summary sentence with the corresponding segment in the original sentence based on the set of LCSs. When a summary segment can be aligned with that of the original, both of the segments between the summary and the original are marked as 1. The result of the segment alignment is shown in Figure 4.4. In this chapter, such mark is referred to as a edit strings. In the Figure 4.4 such edit string $0, 1$ is shown under each segment.

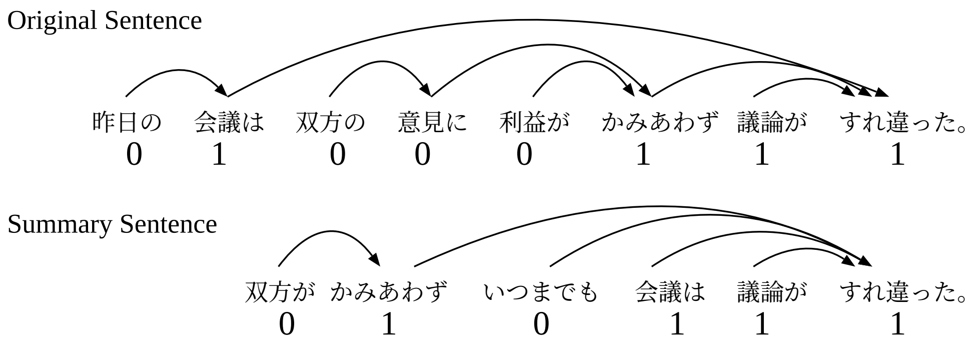

Original Sentence
昨日の 会議は 双方の 意見に 利益が かみあわず 議論が すれ違った。
0 1 0 0 0 1 1 1
Summary Sentence
双方が かみあわず いつまでも 会議は 議論が すれ違った。
0 1 0 1 1 1

Figure 4.4. Example of Segment Alignment

### 4.3.3 Automated Extracting of Paraphrasing

We collect examples of paraphrasing to capture the basic properties of paraphrasing. Figure 4.5 shows the parts which we regard as simple example of paraphrasing. In Figure 4.5, boxes represent the segments in the sentences and the shaded ones represent the segments that are aligned by the process described in Section 4.3.2. More concrete explanations for each case is described as follows:

- Pattern A: the case in which an un-aligned segment is modified by more than one aligned segments.

- Pattern B: the case in which an un-aligned segment modifies another aligned segment and one or more aligned segments modify the un-aligned segment.

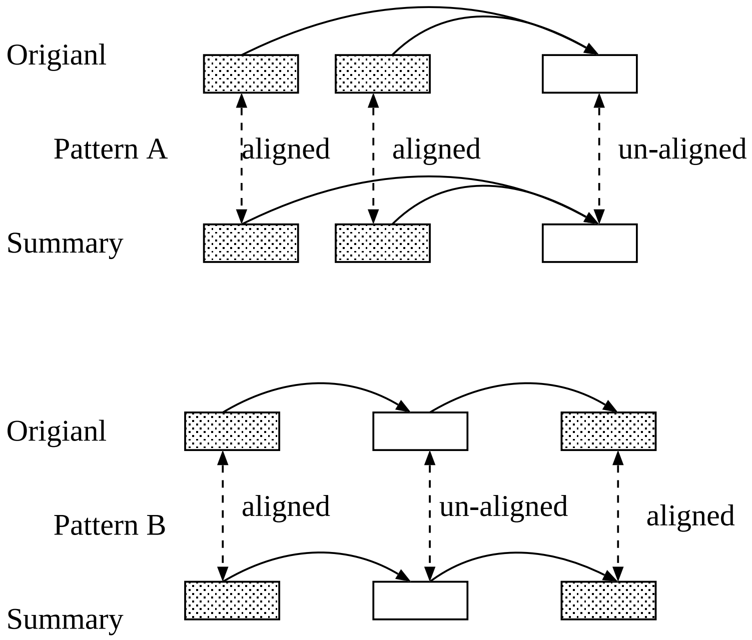


Figure 4.5. Automated Extraction of Paraphrasing

### 4.3.4 Iteration for detecting the occurrences of sentence combination

So far we have discussed, a summary sentence always corresponds to only one sentence in the source text. In order to deal with the one-to-many correspondence between a summary sentence and the source sentences, the algorithm iterates the sequence of the processes described in the section.

The following pair of sentences shows an example of the pair where the summary sentence corresponds with the original sentence in the trial of alignment. At the first iteration of the alignment, the alignment program yields the edit string '00011' for the summary sentence.

Original　「県には、前例に とらわれず、 積極的対応を 求めたい。」
(the prefecture office TOP) (precedents) (being unconcerned)
(the proper measures) (to be needed)
Summary　「借り上げに 力を 入れるなど 積極的対応を 求めたい。」
0　0　0　1　1
(the official leases) (an effort to) (such as making)
(the proper measures) (to be needed)

As shown in Figure 4.1, more than one un-aligned segments existing in the summary sentence is the condition to make the alignment process to be iterated. For the pair in the example, the sequence [借り上げに, 力を, 入れるなど] (3 segments) still remains as a candidate expression for alignment. If the source text gives a sentence as shown below, the sentence is aligned to the subsequence.

「借り上げに、力を 入れ、仮設住宅を 確保すべきだ。」
(the official leases) (an effort to) (making) (temporal residents) (should secure)

As the result of the sentence alignment, each segment in the pair of sentences is aligned as follows and the edit strings '11011' is obtained for the summary sentence.

Original1　「県には、前例にとらわれず、 積極的対応を 求めたい。」
Original2　「借り上げに、力を 入れ、 仮設住宅を 確保すべきだ。」
Summary　「借り上げに 力を 入れるなど 積極的対応を 求めたい。」
1　1　0　1　1

Note that, in the example above, the segment “入れるなど”does not align to the segment “入れ”. This is reasonable to identify the summary expressions that are not occurred in the original text.

The condition to terminate the iteration of alignment is the case when subpart of the summary sentences does not align with the expression in the original text, and when there is no subparts which has two or more un-aligned segments in the summary sentence.

Table 4.3. Results of Alignment Iterations

| # Trial | # un-Aligned Segments (l) | | | Total |
|---|---|---|---|---|
| in the iteration | 0 | l < 50% | l ≥ 50% | |
| 1 | 990 | 444 | 164 | 1598 |
| 2 | 141 | 401 | 67 | 609 |
| 3 | 25 | 118 | 3 | 146 |
| 4 | 3 | 14 | 1 | 18 |
| 5 | 0 | 1 | 0 | 1 |
| Total | 1159 | 978 | 235 | 2372 |

## 4.4 Results of alignment and un-corresponding expressions

In this section, we describe the the results of applying the alignment algorithm to the summaries that we have collected. The rows in the table are divided into the number of trails in which the alignment iteration ended. The columns in the table are classified into the ratios of un-aligned segments in the summary sentence. The ratios of the un-aligned segments in the summary sentence are classified into 3 cases as described below.

- all segments are aligned
- less than the half of the segments are still un-aligned
- half or more of segments are un-aligned

The number of the summary sentences which align to only one sentence is 990. Those aligned pairs include 672 examples in which the aligned sentence is the same as the summary sentence.

As is found in Table 4.3, most of the sentence alignment end within the third iteration. There are 978 summary sentences that have half or more aligned segments, beside the completely aligned 1159 sentences, which have no un-aligned segments. As compared with these, there are only 235 summary sentences in which half or more segments are still un-aligned. From the results, we can conclude that 81% of summary sentences have half or more aligned segments.

In 2467 summary sentences, only 95 summary sentences do not have the expressions that are based on the dependency relationship in the source text. This also supports the assumption that summary sentences are generated based on the expressions that closely relate with the dependency structure of the source text.

In the rest of this chapter, we analyze the results of the alignments by the two points of view. One is the expressions with the un-aligned segments, in other words new expressions in the summary sentences, are how a human summarizer generate such expressions. The second is how sentence combination is applied in generating the summaries.

## 4.5 New Expressions in Summary Sentences

As a preparation for analyzing the un-aligned segments, we categorize the expressions by the position where the expression appears in the summary sentence. We categorize the position into three cases: un-aligned expressions appearing at the beginning, the middle, and the end of the summary sentence. Furthermore, when the un-aligned segments appear in the middle of the summary sentence, the un-aligned segments can be divided into two cases. When the expression in the middle of the original sentence is paraphrased into another expression as in the summary B in Figure 4.6, the un-aligned segments must appear in the middle of the summary sentence. The other reason is concerning with the sentence combination as in summary A in Figure 4.6. In this section, we discuss the former examples, and the the latter cases are described in the next section.

### 4.5.1 Un-corresponding segments in the beginning of the sentence.

In general, un-aligned segments appear when a new expression is added to the summary sentence and when paraphrasing of the original expression is applied. Addition of a new expression tends to appear in the beginning and the end of the sentence.

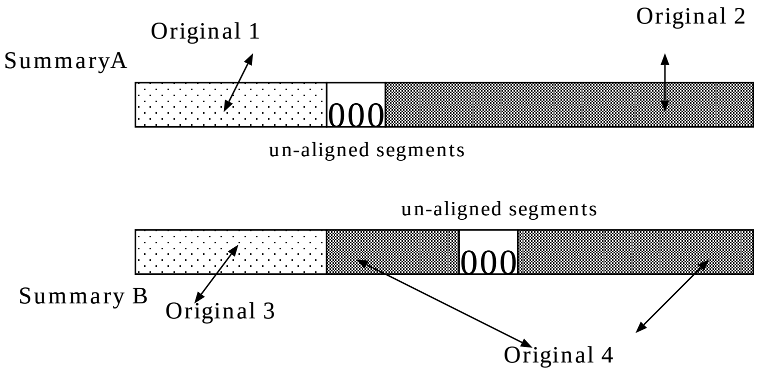


Figure 4.6. Patterns of Un-aligned Segments in the Middle of the Summary Sentence

Table 4.4. The Length of un-Aligned Segments

| | length of un-aligned Segs | | | | |
|---|---|---|---|---|---|
| Position | 1 | 2 | 3 | 4 | 5 |
| Beginning of Sent. | 369 | 157 | 72 | 47 | 50 |
| End of Sent. | 93 | 75 | 31 | 39 | 33 |

Table 4.4 shows the number of the substrings of the un-aligned segments in the beginning and the end of the summary sentences. The table shows one of the specific features that there are many un-aligned segments of length 1. Such examples that has a single un-aligned segment at the beginning of the sentence are shown as follows in which the bracketed segments represent un-aligned segments.

［捜査は、］ 大きな山を越し、…

([the inquiry TOP] the most difficult problem is cleared ...)

［自民党は、］ 領土問題に合意したが、…

([Liberal Democratic Party TOP] though the agreement on the matter was reached ...)

［また、］ 業務の多角化により…

([And,] Because of the doversofocation of the service ...)

The first two examples have a single un-alignment segment that is TOP in the sentence and third example has a conjunction. Such expressions are also known as expressions in which the state-of-the-art dependency analysis module have difficulty in deciding the relationship with other expressions. One of the reasons for the un-alignment of the beginning single segment must be errors in dependency analysis.

However, the expressions that appear in the beginning of the general summary sentences play a special role in generating summaries even if such problems of the dependency analysis modules are eliminated from the considerations.

The following pairs of sentences show typical examples of the initial expressions in the summary sentence. Each pair is aligned by our algorithm and the sentence marked by "s" shows the summary sentence and that of "o" shows the corresponding sentence. The bracketed expressions show the un-aligned segments.

(1o) 共同作業所と公的な授産施設は…

(1s) [大震災で 兵庫の] 共同作業所と公的な授産施設の…

[in the disaster] [in Hyogo Prefecture] public works spaces and employment offices are ....

(2o) 基本的には、恒久住宅の建設を進めなければならない。

(2s) [そして] 基本的には、恒久住宅の建設を進めなければならない。

[And] The construction of lasting residents is fundamentally needed.

(3o) [開業して三十年たち]、施設の劣化が目立つ。

[it used for 30 years] there are a conspicuous number of damages in the equipment.

(3s) [特に、東海道新幹線は] 施設の劣化が目立つ。

[(Especially) (Tokaido-Shinkannen TOP) ]

The pair of sentences (1o) and (1s) is an example in which the information of the event or the location is added. Information such as location, time, and the event is sometimes omitted from the middle of the text.

The pair of sentences (2o) and (2s) is an example that the conjunction “そして”is added to improve the coherence between the generated summary.

In the last pair (3o) and (3s), both the sentences have the un-aligned segments. Instead of the eliminating the unimportant information from the source sentence, the summary sentences adds the information for understanding the proposition of the summary sentence efficiently.

### 4.5.2 Un-correspond segments in the end of the sentence

The length of un-aligned segments at the end of the summary sentence is shown in the second row in Table 4.4. The following pairs are the examples of such un-aligned segments.

(4o) 辞退した ケースも [六百件以上 ある]。

(declining) (cases TOP) (be more than 600 cases)

(4s) 辞退した ケースも [多い]。

(declining) (cases TOP) (many cases exist there)

(5o) 乏しさを [弁解する 理由には できない。]

(the poverty) (is inexcusable)

(5s) 乏しさを [露呈した。]

(the poverty) (is exposed)

(6o) 有無を 早急に 明らかにしなくてはならない。

(the existence) (at once) (is needed to be investigated)

(6s) 有無を 早急に 明らかにして、 [国民の 不安を 取り除く 必要が ある。]

(the existence) (at once) (is needed to be investigated) [in order to relieve the social unrest]

The pair (4o) and (4s) show a paraphrasing of the original expression in a different expression which has the same number of the segments. In this example, the original expression "六百件以上ある。" gives the actual number of the subject in the original sentence and it is paraphrased to the expression "多い" in the summary sentence. The summary expression "多い" states that there are many occurrence of the event that the subject of the sentence mentions.

An example of the similar paraphrasing is given in the pair of (5o) and (5s). This example shows an example where the original expression is changed into a shorter expression. We understand that such a paraphrasing is one of the operations to simplify the original sentence.

However, there is an example like the pair (6o) and (6s), in which the length of the summary expression is longer than that of the original. This is a special property of the paraphrasing at the end of sentences.

### 4.5.3 Paraphrasing in the middle of sentence

We have already described an automated extraction of paraphrasing in Section 4.3.3. Although there are not so many examples that are extracted by the extraction, those examples are regarded as reliable examples of paraphrasing. Table

Table 4.5. Paraphrases Extracted Automatically

| Group | # | | Examples |
|---|---|---|---|
| G1 | exp1 | Original | 重く、 [つらい] 課題である。 |
| | | Summary | 重く [辛い] 課題である。 |
| | | | *In this case, the notation was changed from by Hiragana to by Kanji. |
| | exp2 | Original | 女性が 生涯に [産む] |
| | | Summary | 女性が 生涯に [生む] |
| | | | *An example for error of using Kanji |
| | exp3 | Original | 崩れれば、 競争も [強まる] |
| | | Summary | 崩れれば、 競争が [活発になる] |
| | | | (IF CLAUSE) (the competitions) (to be activated) |
| G2 | exp4 | Original | 電事連が [公に しないのは] |
| | | Summary | 電事連が [公開しないのは] |
| | | | (Denjiren NOM) (the reason why they discoluse TOP) |
| | exp5 | Original | 反するのではという [問題も 抱える。] |
| | | Summary | 反するのではという [問題である。] |
| | | | (such as) (is a problem) |
| | exp6 | Original | 支持率は、 [六七%にも 達している。] |
| | | Summary | 支持率は [六七%。] |
| | | | (the rate of the supporting) (reaches 67%) |

4.5 shows such examples that are automatically extracted. The total number of those is 121. They are divide into two types. One type is examples in which the summary expression can be aligned to the original expression in one-to-one basis. The other type is examples in which the original expression is paraphrased into a shorter expression in the summary sentence. We label the former group as G1 and the latter group as G2 in Table 4.5. To discuss un-aligned segments in the middle of the summary sentences, we employ such a distinction of the groups as a tool for analysis.

Before discussing paraphrase in the middle of summary sentences, we show another operation that simply adds segments to the summary sentence as shown

Table 4.6. Examples of un-aligned segments in the middle of summary sentences

| Pattern | $n = m$ | $n >$m | $n < m$ |
|---|---|---|---|
| Examples | 210 | 241 | 42 |

below.

(7o) 感染力が 弱く、 二次感染の 心配はない。
(7s) 感染力が 弱く、 [日常的接触では] 二次感染の 心配はない。
(infection NOM) (weak) [with daily connection] (the spreading of the infection) (there is no anxiety)

(8o) しかも、 今後
(8s) しかも [北朝鮮は] 今後
(moreover) [North Korea TOP] (after this)

(9o) 要望を 聞く
(9s) 要望を [幅広く] 聞いた
(the demands) [broadly] (to be complied)

The total number of such examples that add the new segments to the summary is 113. Both pairs (7o)-(7s) and (8o)-(8s) show examples where the TOP expressions are added to the summary sentence. The pair (9o)-(9s) is also an example that gives additional information.

To discuss the remaining examples, we introduce two length $m$ and $n$ as shown in Figure 4.7: $n$ refers to length of the un-aligned segments in the sources sentence, and $m$ refers to that of the summary sentence. For example, when $m < 0$ and $n = 0$ in a pair of segments, an addition of new segments is applied in the summary sentence of the pair. Table 4.6 shows the number of the occurrence such paraphrasing categorised in terms of $n, m$.

When the length of the un-aligned segments in the summary sentence is the same as that of the original sentence, the pair of the un-aligned segments is a paraphrase that is grouped as G1 in Table 4.5. Such paraphrases are shown in the following examples.

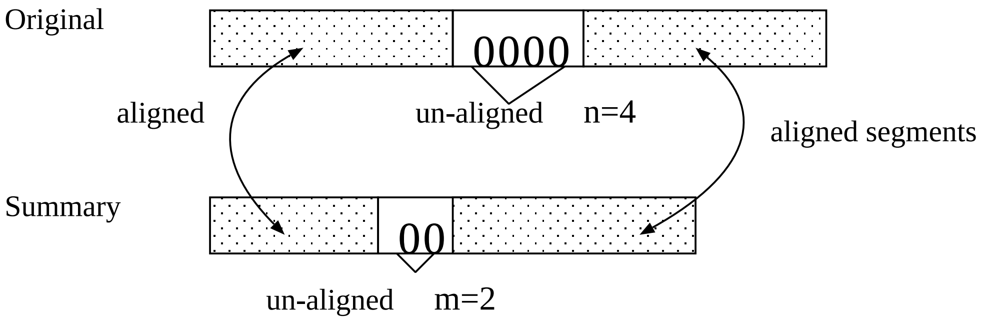


Figure 4.7. The Length of Un-aligned Segments between the Corresponding Pair

(10o) 政治家は、 [韓国の] 発展と
(the politicians TOP) [South Korean] (the development)
(10s) 政治家は、 [その] 発展と
(the politicians TOP) [PRONOUN] (the development)

(11o) 事態の [推移を] 注意して
(11s) 事態の [推移に] 注意して
(of a situation) [the transition] (with notice for)

(12o) オウム真理教関連の [施設に 一斉に] 家宅捜査を 行った。
(12s) オウム真理教関連の [施設に対して 全国一斉の] 家宅捜索を 行った。
(Related with Aum) [(establishments OBJ) (at once)] (domiciliary search) (did)

The pairs of (10o)-(10s) and (11o)-(11s) show simple examples that are very similar to the examples (exp1),(exp2), and (exp3) in Table 4.5. The pair (12o) and (12s) shows the case where simple paraphrase occurs twice in adjacent segments.

On the other hand, the following examples show those of the group where the original expression is changed into shorter expression in the summary sentence. The major pattern of such examples is that only one un-aligned segment appears between the aligned segments (101 out of 241 examples). Such paraphrases are shown as follows:

(13o) してもらって [支援の ネットワークを] 広げては

(13s) してもらって [支援を] 広げては

[the supporting network] → [the supporting]

(14o) 社会党・さきがけが、[侵略行為・侵略戦争や 植民地支配の] 反省の

(14s) 社会党・さきがけが [侵略行為などの] 反省の

[the acts and the aggressive war and the colonial rule] → [such as the acts of aggression]

(15o) 青島知事の [こうした 方針や 行動に] 反発した

(15s) 青島知事の [施政方針に] 反発した

[those policies and the behaviour] → [those policies]

(16o) 実施する [「措置予定」は 七百件を 超えたが、] 大半は

(to apply) [(the plan) (700 cases) (over)] (the major cases TOP)

(16s) 実施する [「措置予定」の] 大半は

(to apply) [of the plan] (the major cases TOP)

(17o) 当局が [実力行使を 行うと 警告、 厳戒態勢で] 臨み、

(the authorities) [(force) (use) (to warn) (in the strict watch) ](to face)

(17s) 当局が [民主化グループに 強行姿勢で] 臨んでいる。

(the authorities) [(democratisation group) (with decisive attitude)](to face)

The examples (13o)-(13s), (14o)-(14s), and (15o)-(15s) show examples that have the common case marker between the original segments and the paraphrased ones. In semantic of those paraphrases, original segments are abstracted or represented by showing an example.

The example (16o)-(16s) show an example in which the original verb is phrases paraphrased into the noun phrase. Such paraphrases are characterized by the occurrence of "A の B", where "の" is a postposition and A and B are noun phrases.

(17o)-(17s) is an example of complex paraphrasing. In this case, the addition of segment "民主化グループに" and the elimination of segments "実力行使を行うと警告" are applied, and segment "厳戒態勢" is paraphrased into segment "強行姿勢".

There are some exceptions that the summary sentence has an expanded expression of the original one. The pair (18o) and (18s) is just the opposite paraphrase to those in Group G2 of Table 4.5. On the other hand the pairs (19o)-(19s) and (20o)-(20s) may relate with the addition of segments rather than with simple paraphrases.

(18o) 将来のために [工夫が] 必要だ。
(for the future) [an idea] (is needed)

(18s) 将来のために [工夫する ことが] 必要だ。
(for the future) [(an idea) (thinking out)] (is needed)

(19o) 警察組織の [トップが] 銃撃された。
(of the police office) [(the head)] (was shot)

(19s) 警察組織の [トップ、 国松警察庁長官が] 銃撃された。
(of the police office) [(the head) (Kunimatu)] (was shot)

(20o) 学校は [ようやく] 隔週休みとなる
(schools TOP) [(at last)] (adopt an every other week holiday)

(20s) [今 ようやく] 学校は 隔週休みとなる
[(now)(at last)] (schools TOP) (adopt an every other week holiday)

## 4.6 Sentence Combination

### 4.6.1 Sentences in which Sentence Combination is applied

In Chapter 3, we found most of the sentence combination applied between adjacent sentences. In order to examine that in the longer and more complex summaries, we analyze the distribution of the positions of the pair of sentences that sentence combination is applied. There are 1054 examples that sentence combination is applied. Table 4.7 shows the distribution of the examples. In Table 4.7, the distance is corresponds how far the pair of sentences that sentence combination is applied is apart from each other. If a summary sentence comes from an adjacent pair of sentences, the distance is 1.

As a result of Table 4.7, the most examples of sentence combination are applied between adjacent sentences. Furthermore, we found that the number of examples

Table 4.7. The position of sentence that Sentence Combination is applied

| distance $n$ | # examples |
|---:|---:|
| 1 | 601 |
| 2 | 197 |
| 3 | 84 |
| 4 | 49 |
| 5 | 29 |
| 6 | 23 |
| 7 | 11 |
| 8 | 10 |
| 9 | 7 |
| n $\geq$ 10 | 43 |
| Total | 1054 |

of sentence combination decreases according as the distance between a pair of sentence becomes apart from.

### 4.6.2 Types of Sentence Combination

The examples of the sentence combination are divided into examples with and without un-aligned segments. In order to characterise the properties of the sentence combination, we first investigate 169 sentences in which the sentence combination is applied without un-aligned segments (Those examples are in the sentences that have no un-aligned segments in Table 4.3). Those sentences contain 232 examples. From the analysis of the examples, they are further classified into 4 types of connections as shown in Table 4.8.

Type A is the combination where one sentence is connected with another sentence to describe two or more proposition in a single sentence. Most of the connection between the propositions have a rhetorical relationship such as background, reason and result. This shows that generating a summary is not only an extraction of important sentences but also contains additions of the relevant information to the extracted sentences.

Table 4.8. Types of Sentence Combination

| Type | | Example |
|---|---|---|
| A | Original 1 | 大学側も 柔軟に 共同研究に 応じるべきであろう。<br>University should adopt joint researches flexibly. |
| | Original 2 | 特に、 各地の 大学は 地方ニーズに 応える よう 配慮してほしい。<br>Especially, the local universities must reflect what the area demands. |
| | Summary | 大学側も 共同研究に 応じるべきで、 特に、 各地の 大学は 地方ニーズに 応える よう 配慮してほしい。<br>University should adopt joint researches and the local universities must reflect what the area demands. |
| B | Original 1 | 北京での 交渉は、関連問題で 最後の 詰めを 残している。<br>The negotiation in Beijing is in a last stage. |
| | Original 2 | 北朝鮮側は、 コメの 支援問題以外の 政治問題は 今回の 会談では 話し合いたくない立場を 譲っていないようだ。<br>The North Korea side keeps the attitude that does not want to discuss the political issues other than the rice aid. |
| | Summary | 北京での 交渉は、北朝鮮側も、 コメの 支援問題以外の 政治問題は 今回の 会談では 話し合いたくない立場を 譲っていない。<br>As talking about the negotiation in Beijing, the North Korea side keeps the attitude that does not want to discuss the political issues other than the rice aid. |
| C | Original 1 | 新党結成を めぐり 社会党は 大混乱で、 政権基盤は 大きく 揺らいでいる。<br>The movement of establishing new political party from the Socialist Party is confusing the political fundation of the prime minister. |
| | Original 2 | 首相は 演説どころではなかった。<br>The prime minister did not concentrate on the speech. |
| | Summary | 新党結成を めぐり 社会党は 大混乱で、 政権基盤が 揺らいでいる 首相にとっては 演説どころではなかった。<br>The prime minister whose political fundation was being confused did not concentrate on the speech. |
| D | Original 1 | 県は恒久住宅の整備で手一杯で、仮設住宅の居住者への配慮が足りない。<br>The prefecture office is very busy with the constructions of lasting residences and a few supports are provided for the people that live in the temporary houses. |
| | Original 2 | 今後、就労問題などに もっと目を向けた精神面での支援が必要だ。<br>Henceforth, the supports in the life side such as the aid for employment are needed. |
| | Summary | 仮設住宅の居住者に目を向けた生活面での支援が必要だ。<br>The supports for the people that live in the temporary houses is needed. |

Type B is the combination where a sentence is combined with another sentence and plays the topic segments of the summary sentence. This combination can be divided into two cases: the topic segments of the summary sentence is also the topic of the original sentence or not. The following pair of sentences is an example of the latter case, in which the bracketed expression shows the segments that play a topic role in the summary sentence.

Original 1 協定の効果を確実にするのが <紛争処理案だ。>

It is the dispute-settlement system to ensure the effect of an agreement.

Original 2 紛争処理機構では、処理を調停にとどめる案と、拘束力を持つ処理を行う案がある。

The dispute-settlement mechanism can offer a mediation plan or can execute a process to mediate the dispute.

Summary <紛争処理案は、>処理を調停にとどめる案と、拘束力を持つ案がある。

(The phrase ’the dispute-settlement system’ in the original sentence 1 is combined as the topic of the summary sentence.)

Type C is the combination where a sentence changes its form by modifying a segment in the other sentence in the summary sentence.

Type D is the combination where a part of sentence is used as an argument of the predicate of the summary sentence and the other segments of the summary sentence come from another sentence. In Japanese, an argument of a predicate is indicated by a postposition according to the semantic role between the argument and the predicate.

Table 4.9 shows the numbers of examples of those types of sentence combination. The numbers of examples without un-aligned segments are shown in the first row in Table 4.9 and those with un-aligned segments in the second. The tendency of un-aligned segments with the sentence combination is similar to that of the un-aligned segments appear in the beginning of the sentence.

As a result of Table 4.9, major type of sentence combination is Type A. It shows the rhetorical relation between propositions of the pair of sentences to combine into a summary sentence.

Table 4.9. Occurrence of each combination Types

| # un-aligned | Combination Type | | | | Others | Total |
|---|---|---|---|---|---|---|
| segments | A | B | C | D | | |
| 0 | 176 | 25 | 19 | 10 | 2 | 232 |
| rather than 0 | 319 | 46 | 56 | 11 | 41 | 473 |

## 4.7 Conclusion

In this chapter, we investigate operations in summary generation. In order to align a summary expression with the corresponding original expression in the source text, we introduce an automated algorithm based on the dependency structure of sentences. Our algorithm detects not only one-to-one sentence alignments, but also one-to-many sentence alignments. We apply the algorithm to human made summaries, and analyse the results of the alignments. As a result of the analysis, we find most summary expressions keep their dependency relation in the original sentences and confirm one of the operations called ”sentence combination”, in which more than one source sentence are used to generate a summary sentence, plays an important role in summary generation. Furthermore, we characterise operations and paraphrasing that cover most summary generation.

# Chapter 5

# Implementation of Sentence Reduction

## 5.1 Introduction

In this chapter, we describe an experimental implementation one of the operations in summary generation, sentence reduction. As we described in Introduction, most of the automated summarization systems adopt the following two general phases.

- Extraction phase: To select the significant sentences that show the main information of the article.
- Revision phase: To revise the extracted sentences into simpler ones.

In this dissertation, we decompose the process in the revision phase into 3 operations as shown in Figure 5.1. Among them, sentence reduction, which eliminates unimportant segments from sentences, is the major operation for generating more readable and simpler summaries.

The right hand side of Figure 5.1 shows an overview of knowledge acquisition for realizing such an operation from human-written summaries. For the experimental implementation, we apply support vector machines (SVM), based on machine learning method. The training data are extracted automatically from the pairs of the original texts and their corresponding summaries.

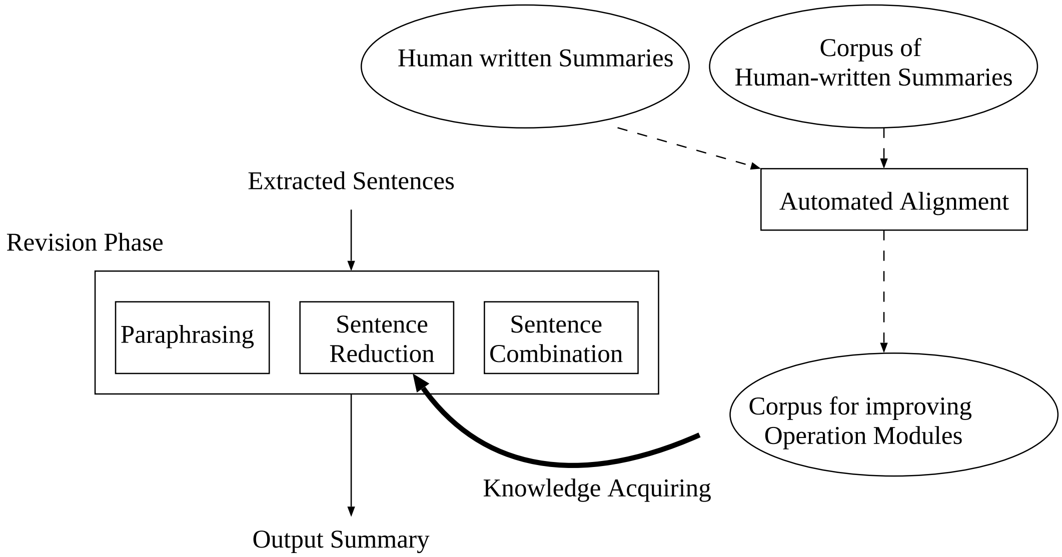


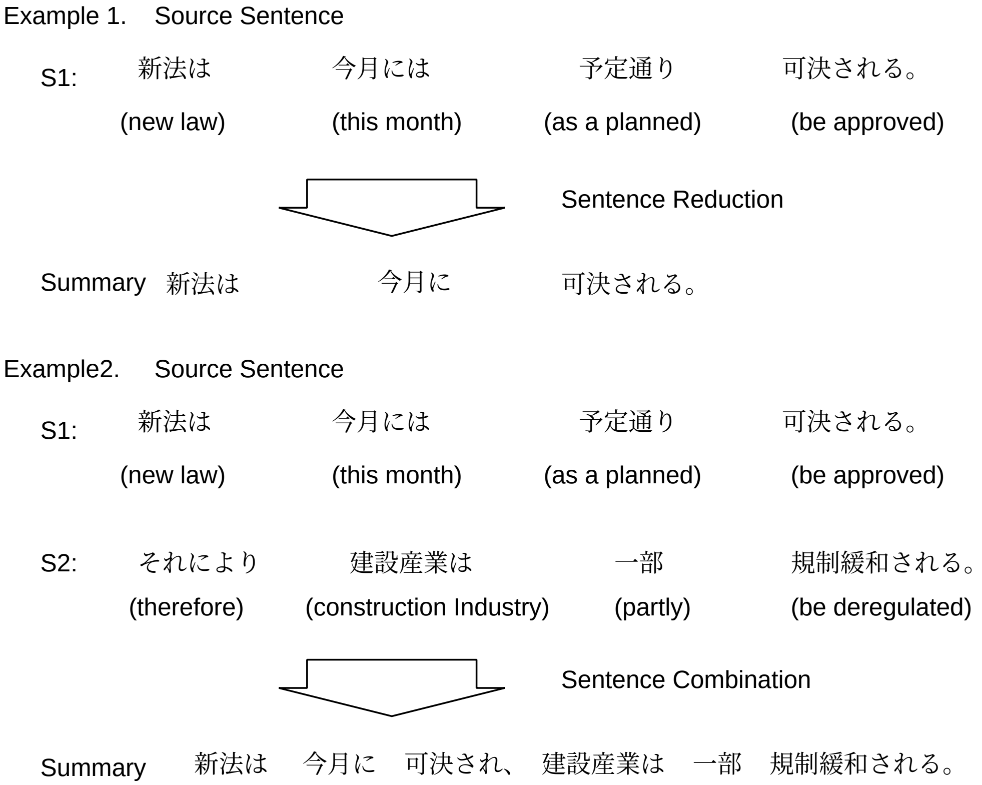
Figure 5.1. Acquisition of Knowledge for Revising Operations

According to Chapter 4, we assume the units in the sentence to be eliminated are *bunsetu segments*, which are very basic phrases used as the basic units in the syntactic structure of Japanese sentences. In this chapter, segments are referred to as *bunsetu segments*.

This chapter is organized as follows. Section 5.2 describes the target operation that deletes segments in a sentence while keeping their main meaning. Section 5.3 describes the algorithms to build the corpus data for learning. Section 5.4 describes the experiment using support vector machines (SVM) for acquiring the rules to realize the target operation.

## 5.2 Target Operation in Revision Phase

Several previous works point out that there are various operations in the revision phase. For example, Jing and McKeown [13] introduce 6 types of operations to revise sentences. Moreover the revision operations that humans perform in summarization are very complicated. From our previous work, however, we consider that machines are capable of emulating some of the operations on the revision phase.



Example 1.　Source Sentence

S1:

| 新法は | 今月には | 予定通り | 可決される。 |
|---|---|---|---|
| (new law) | (this month) | (as a planned) | (be approved) |

Sentence Reduction

Summary　新法は　今月に　可決される。

Example2.　Source Sentence

S1:

| 新法は | 今月には | 予定通り | 可決される。 |
|---|---|---|---|
| (new law) | (this month) | (as a planned) | (be approved) |

S2:

| それにより | 建設産業は | 一部 | 規制緩和される。 |
|---|---|---|---|
| (therefore) | (construction Industry) | (partly) | (be deregulated) |

Sentence Combination

Summary　新法は　今月に　可決され、 建設産業は　一部　規制緩和される。

Figure 5.2. Examples of Summary Operations

Prior to constructing the system, we have investigated the revising operation that human performs. From the investigation, we decide to focus on the following two operations as the most fundamental operations to revise sentences. Examples of each summary operation is shown in Figure 5.2.

- Sentence reduction: A summarized sentence is generated from exactly one source sentence.
- Sentence combination: A summarized sentence originates from two or more source sentences.

According to our investigation, the operation of the sentence reduction that makes a sentence into a simpler one is used when the summary sentence is generated with sentence combination. In this chapter, we concentrate on discussing sentence reduction. Our implementation of the operation is to eliminate the unimportant segments from the source sentences.

## 5.3 Construction of Aligned Data

Recent research on statistical natural language processing has shown that statistical learning approach is useful and can be applied to various applications, such as POS tagging, syntactic dependency analysis, etc. In the studies of automated summarization, some works acquire summarization knowledge from human made summaries, for example, Jing and McKeown [11] uses a statistical model to acquire the rules that specify how sub-parts are removed from the original sentence.

The main obstacle of Machine Learning based summarization research is the lack of adequate corpora today. Only a few small collections of texts whose units have been manually annotated in terms of textual importance are available. To circumvent this problem, some works propose an algorithm that constructs training corpora automatically [23, 10]. Their algorithm takes a set of the document and the corresponding summary as input, and extracts the sentences that are used to produce the summary. Such corresponding pairs of the sentences are called aligned pairs.

Table 5.1. Paraphrasing of the words in segments

| POS | Allowed Operation |
|---|---|
| Noun | Make the form shorter |
| Verb | Change the inflectional form |
| Adjective | Change the inflectional form |
| Adverb | Prohibit any paraphrase |
| Conjunction | Prohibit any paraphrase |
| The others | Change the inflectional form |

As shown in Figure 5.1, the automated alignment plays an important role to acquire knowledge for sentence reduction. We apply the following alignment algorithm to each sentence $S_i$ in the source document.

1. A sentence is extracted from the human made summary. We use the character-based cosine distance between the pair of sentences as the degree of similarity. Let the extracted sentence be the candidate $c$ for the aligned summary sentence with $S_i$.

2. Let the set of the commonly used segments between $c$ and $S_i$ be $C$. The paraphrase of segments is allowed if the pair of segments satisfies the assumption shown in Table 5.1.

3. If the size of $C$ is 1 or 2, $c$ is withdrawn from the candidate. If $C$ does not contain the last segment of $S_i$, $c$ is withdrawn from the candidate. Otherwise, the pair $< S_i, c >$ is added to the set of training examples.

## 5.4 Machine Learning Method

Since the elimination operation that humans perform in summarization is quite complex, it costs too much to model the operation manually. In this section, we describe an experiment to confirm that a machine learning method is useful to model that type of human operation.

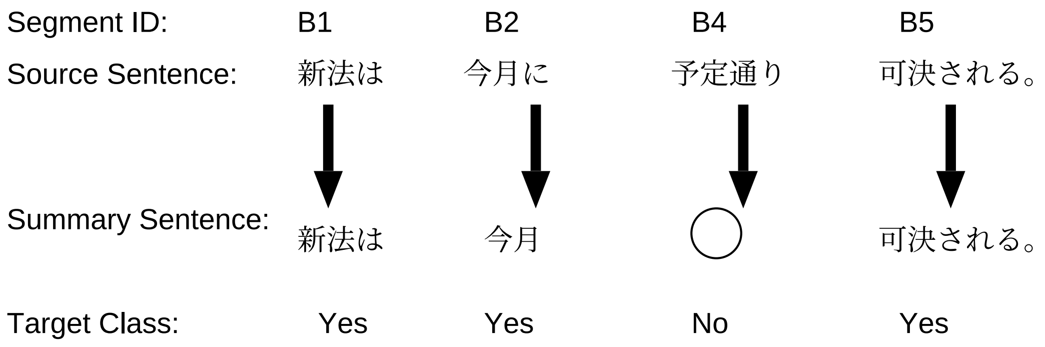

Segment ID: B1 B2 B4 B5
Source Sentence: 新法は 今月に 予定通り 可決される。
Summary Sentence: 新法は 今月 ○ 可決される。
Target Class: Yes Yes No Yes

Figure 5.3. Examples of Aligned Pair

### 5.4.1 Data

In our approach, the training data are generated from human written summaries. We have 90 newspaper articles and the corresponding summaries of these articles that are compiled by three human subjects with the following instructions. (270 summaries in total)

- The summary should keep the overall story of the source text and the author's main opinion.
- If the summarizer uses proper nouns in the summary, the form of the proper nouns should be kept intact.
- The number of characters in a summary should be about 40% of that of the original article.

The basic information of the summaries that we collected are shown in Table 5.2.

With the algorithm described in Section 5.3, we obtained 1057 aligned pairs. Figure 5.3 shows an example of the aligned pairs. From the pairs, we let a machine "learn" the rules to automatically eliminate unimportant segments. In the set of training examples, the number of the removed segments is 2942 our of the 10856 segments in the source sentences.

Table 5.2. Basic Information about the Summary Data

| | # Sentences |
|---|---:|
| Summarizer A | 763 |
| Summarizer B | 773 |
| Summarizer C | 931 |
| Total | 2467 |

### 5.4.2 Support Vector Machines

Machine learning is a method to acquire rules from the training data. In the training data, each example is represented by a tuple $< f, t >$: $f$ represents the features of the example and $t$ represents the class that the examples belong to. When the learning is completed, the acquired rules take unknown examples as an input and predict the class that each of the example belongs to.

Support Vector Machines (SVM) [35] are a learning system based on recent advances in statistical learning theory. SVM delivers a state-of-the-art performance in real-world applications such as text categorization, hand-written character recognition, image classification, and bioinformatics.

Although the aim of this chapter is not to describe Support Vector algorithms, one thing should be noted; the learning ability of SVMs is independent of the dimensionality of the feature space. SVM measures the complexity of hypothesis (rules) based on the margin with which they separate the training data, not on the number of features. This means that it can be generalized even in the presence of very many features, if the training data is separable in the hypothesis space.

### 5.4.3 Features

The characteristics of the segments are arranged into a vector of features. We represent the feature-set with a vector $f$ as follows. The vector $f$ is an n-dimensional vector, in which $a_i$ stands for the $i$-th feature.

$$f :< a_1, a_2, \cdots, a_n >$$

SVM is capable of managing high dimensional vectors. In our system, the size of the dimensions of the feature vector that characterizes the segments exceeds 2000. Instead of listing the entire features, we describe the following 5 bases to be used to define the features [1].

1. **Semantic features**

   The features that characterize the meaning of the target segment are assigned according to the independent words in the segment. Suppose the target segment is $b_i$.

   If the word is an inflectional word, such as a verb, the individual forms of the word are used as the semantic feature. For example, if the segment is "Setumei-sita" (*explained* in English), the word "Setumei" is used as the semantic feature.

   If the word is a noun, sub-category names of nouns such as person name, place name, numerals, etc. is used as the semantic features.

2. **Syntactic features**

   The syntactic features are characterised with the inflectional forms of a word or the attached function words in $b_i$. For examples, if the segment is "Daitouryou-ga" (*President* :AGENT in English), the particle "ga" marks the element as an AGENT, and we use the marker as the syntactic feature.

3. **Position in the sentence**

   It is important to clarify in which the position $b_i$ appears in a sentence. Since the results of dependency structure analysis of the sentence forms a tree structure, we use the position of $b_i$ in the tree structure such as leaf, root, to assign this type of the features. We also set a feature with which the occurrence of some punctuation after $b_i$ is indicated.

4. **Context features**

   As the interpretation of nouns depends on the context, it is very difficult to represent the characteristics of the meaning of nouns in a certain context. In the area of Information Retrieval, term frequency in the document is

[1] Prior to present characteristics of the target segment $b_i$ in the sentence as feature vector $f$, we use the Japanese dependency structure analyzer CaboCha to determine the syntactic structure of each source sentence in the training examples.

sometimes used to represent the significance of the word in the context. We use the term frequency of nouns to represent the characteristics of the segment in the context.

5. **Inter-segment relationship**
   A segment in the sentence has at least two directions of relationship to the others: modifying another and being modified by another. For the former type of relationship, what segment $b_i$ modifies is represented by the semantic and syntactic features of the segment that is modified by $b_i$. For the latter, the number of segments that modify $b_i$ is used as a feature.

## 5.5 Results of the Experiment

Using the method introduced in Section 5.4, we apply the SVM-based learning to acquire rules for sentence reduction. In the experiment, we use the 1057 aligned pairs that are selected by using the algorithm described in Section 5.5.3. 10856 segments in 1057 source sentences are represented by the features vector and the target class: if the source segment appears in the summary sentence, the target class is “Yes.” Otherwise the target class is “No.” For example, in the training data generated from the aligned pair in Figure 5.3, each segment is represented by the tuple of $< f, target\ class >$ as follows.

$B_1 :< f_1, Yes >$
$B_2 :< f_2, Yes >$
$B_3 :< f_3, No >$
$B_4 :< f_4, Yes >$

After the SVM learning is completed, we evaluate the results. 10-fold cross-validation is used to see the overall performance. To evaluate the degree of correctness of the output of trained SVM, we use Reduction Ratio and Accuracy defined as follows:

Table 5.3. Reduction Ratio and Accuracy of the SVM Learning

| | |
|---|---|
| Reduction Ratio | 76.5 % |
| Accuracy | 77.4 % |

$$\text{Reduction Ratio} = \frac{\text{The number of the 'yes' predictions by SVMs}}{\text{The number of segments in the test data}}$$

$$\text{Accuracy} = \frac{\text{The number of SVMs' predictions that agree with the class of test data}}{\text{The number of segments in the test data}}$$

The results of the cross-validation are shown in Table 5.3. Compared with a simple rule that always returns class 'yes' , which will be acquire 72.8 % in accuracy, the result has an advantage to such a simple rule.

In order to evaluate the SVMs' learning in more detail, we apply an objective evaluation to the generated sentences. The evaluation data is a new set of newspaper articles, and we extract randomly 100 sentences in which sentence reduction is applied by the learned SVMs. Two human subjects determined if the 100 sentences are natural while the context around the evaluating sentence is not given. As a result, 82 % of sentences were judged as natural by both subjects.

Table 5.4 shows some of the examples that both of subjects judged as natural. In Table 5.4, segments in each Japanese sentence are separated by a space and the bracketed segments represent the segments that the program module eliminates for sentence reduction. From the investigation of such examples, we confirm that the methods acquire useful rules such as to eliminate a parenthetical, sentence initial conjunctive expressions and certain adverbial phrases, etc. Those acquired elimination rules agree with what human frequently make in the related works.

However, there are 18 examples that one or more subjects judged as un-natural. In 18 examples, 10 examples were judged as un-natural by both of subjects. The examples (7) and (8) in Table 5.5 show the examples that half of subjects judged as un-natural. The example that both of the subjects judged as un-natural shown in examples (9) to (13) of Table 5.5. Such un-natural sentences, especially examples (7) and (8), include the example that seems to be a natural sentence when a context is given. This shows more sophisticated information

Table 5.4. Examples of Generated Summary Sentences 1

| Both Subjects judged as Natural | |
|---|---|
| (1) | ところが、 ＜いったん＞ ＯＫが 出たのに 「一校 許可すれば 申し込みが 殺到し、 芝生が 傷む」などの 理由で その後 断られていた ことが 三十日、 分かった。<br>But on the 30th, it became cleared that the permission [once] given has been cancelled afterwards, with a reason: If we gave permission to one school, then we would have a rush of orders and the lawn would be damaged. |
| (2) | ＜それ以降、＞ 県警は 二つの 「アジト」を 二十四時間監視下に 置いた。<br>[Then,] the police put these two 'hideouts' under observation of 24 hours. |
| (3) | ＜しかし、＞ 東京、 大阪の 春の 知事選に みられた 有権者の 政党不信は ＜なお＞ 強く、 政界液状化は とどまりそうに ない。<br>[But] like a gubernatorial election in the spring at Tokyo and Osaka, the distrust of voters against political parties are [still] firm, and the malaise for the politics doesn't seem to stop. |
| (4) | ＜逆に、＞ 秋田城跡の 遺構から 見つかった 寄生虫の 卵は、 藤原京や 平城京で 見つかった 遺構の 分析結果と ＜ピタリと＞ 一致。<br>[On the contrary], spawns of parasites found at the ruins of Akita-jo Castle coincided [perfectly] with the result of analysis of ruins found at Fujiwara-kyo and Heijo-kyo. |
| (5) | 捜索は、 ＜東京都内の＞ ＜教団施設や＞ ＜静岡県富士宮市の＞ 富士山総本部でも 行われた。<br>The investigation was also carried out at [the establishment of the religious group ][in Tokyo] and at the central office [in Fujinomiya city]. |
| (6) | ＜また＞ オウム真理教か、という 感じを 持つ 人が 多いだろうが、 オウムを犯人にしようとして ＜何者かが＞ やった 疑いも 残る。<br>Many people may be think, "Omu-Shinrikyo did [again] !", but there remains a suspicion that [somebody] have done it, to make Omu a criminal. |

of context is needed even if sentence reduction is applied to the sentence. In further work, investigations on the human-written summary corpus and inquiries to features that represent such information are needed.

## 5.6 Conclusion

In this chapter we described an experimental implement of one of operations in text summarization. The operation "Sentence Reduction", which we focus on, selects adequate segments from the sentences for generating simpler sentences.

We applied support vector machines (SVM) for acquiring the natural method to select segments. The data for the learning is extracted from the manually written summaries. As a result, we confirm the linguistic features have capability to describe the knowledge to generate the simpler sentences from the original one.

Table 5.5. Wrong Examples of Generated Summary Sentences

| | |
|---|---|
| One Subject judged as Un-natural | |
| (7) | ＜今後、＞ 日蓮宗系の 全国 十七の 寺で ＜オウムについて＞ ＜相談に＞応じていく。<br>[From now] on they will give [advice] [about Aum], at 17 temples of Nichiren-shu in all over the country. |
| (8) | ＜金原学芸員は＞「役人は ＜せっかくの＞ 地元特産の 味に なじめず、都の食べ物を 持参したり 送らせたりして 食べていたらしい」と 話している。<br>[Mr. Kanahara] says ” it seems that public officers couldn’t get used to tastes of local [special] products, and that they ate foods made in the capital, by bringing with them or ordering to send to them.” |
| Both Subjects judged as un-natural | |
| (9) | 一九九三年 十月、 国と 都の 課長ら 十七人が 出席したはずの 港区赤坂の かっぽうは 「客の ことは 言えないが、 ＜ウチの＞ 座敷は 六畳一間だけ。 ＜一度に＞ ＜十七人なんて、＞ ＜とても＞ 入れない」と いぶかる。<br>A person of that restaurant at Akasaka, Minato-ku, where 17 chiefs from the nation and the Metropolis should have gathered in October 1993, said ”we have only one small room. It is [utterly] impossible that [17 persons] enter there [at once].” |
| (10) | 午後 十時前、 コンクリート片に 全身が 埋まり、 ＜顔だけ＞ 出ていた 男性を 近くで 発見、 事故発生から 二十八時間ぶりに 救助。<br>Before ten in the morning, a man whose whole body,[except his face], buried in concrete was found and rescued 28 hours after the accident. |
| (11) | 周囲で 歓声が 上がるが、 ＜目指す＞ 六人とは 違う。<br>There is a shout of joy around us, but this is not directed to the 6 persons [we have an eye on]. |
| (12) | Ａ棟と Ｂ棟の 境界部分の 地下で、 がれきの 中に ＜生存者らしい＞ 反応を得た。<br>In a heap of rubble in the underground among a border of building A and B, there was a reaction [which seemed likely of a survivor]. |
| (13) | ＜私は＞ 過去 五年間、 何時に ＜だれと＞ 会ったかを ＜電子手帳に＞ 入れてあるが、 都と 夜の 打ち合わせを した ことは 一度も なかった。<br>Although [I] have recorded [in my electric schedulor] all schedules of meeting [somebody] for past 5 years, on such and such a time, I’ve never had a meeting at night with the Metropolis. |

# Chapter 6

# Conclusion

In this dessertation, we investigate basic operations that humans use for summary-generation and formulate a couple of effective methods to incorporate those operations in automatic summarizations. The overview that is referred to in the Introduction of this dissertation is reproduced in Figure 6.1 again.

In Chapter 3, we investigated relations between text structure and summarization. The outcomes of the investigation motivated our research directions.

- In analysis of the coherence structure of text, even a human has difficulty in judging the relation between sentences when the related sentences appears far apart from each other. On the other hand, when two related sentences are adjacent, human judgment is far more accurate.

- Human summary generation is based on the two types of operation: sentence reduction and sentence combination. An existence of relationship between a pair of adjacent sentences is a cue to determine whether the operation of sentence combination is activated or not.

- Coherence relationships between adjacent sentences can be automatically identified using features. We confirmed that the features characterizing such relations can also characterize the occurrence of sentence combination.

According to these findings, relevant sentence extraction is needed in extraction phase. The features which can characterize the relations between adjacent sentences will contribute to such extractions.

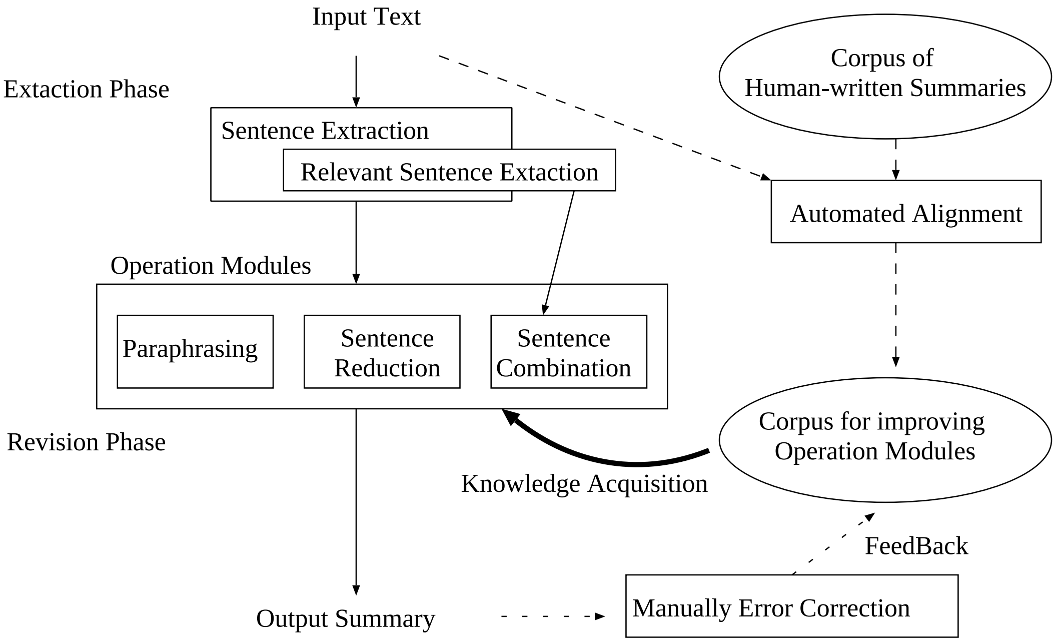


Figure 6.1. Overview of Our Project

In Chapter 4, we further investigated the operations in summary generation. In order to align a summary expression with the corresponding original expressions in source text, we introduced an automated algorithm based on the dependency structure of sentences. Our algorithm detects not only one-to-one sentence alignments, but also one-to-many sentence alignments. We applied the algorithm to human-made summaries, and analyzed the results of the alignments. After analyzing the results, we found that most summary expressions preserve the dependency structure of the original sentences. Our observations also confirmed that one of the operations (namely, ”sentence combination”) plays an important role in summary generation. Furthermore, we characterize paraphrasing that cover most of the summary generation. The alignment algorithm (Figure 6.1) helps in building training corpora which can augment other operation modules.

Finally, in this dissertation, we examined an empirical implementation of one of the operations: namely, sentence reduction. We applied support vector machines (SVMs) for acquiring knowledge for the operation. As shown in Figure 6.1, the training data were automatically generated from manually written summaries. The result shows that linguistic features with the help of SVMs can capture the knowledge required for sentence reduction.

In order to improve the program module for more sophisticated summarization, we have to take into consideration more complex operations as described in Chapter 4. Although this approach made an experiment for the simple operation that manages the elimination of segments, the corpus construction process that shown with dotted arrows in Figure 6.1 can be extended to capture other higher level summary operations. Since our approach inherently utilizes SVM to capture the knowledge from training corpora for summarization, we can extend the feature space to more fine-grained linguistic information such as rhetorical dependency between sentence or clauses, co-reference relations among entity descriptions, etc. How these combinations of the features improve the program modules of the operations will be the major focus of further work.

# Acknowledgements

I would like to express my sincere appreciation to Professor Matsumoto. His comments on my research and the manuscript have been of great value. And all of this research would not have been possible without his comments and encouragement.

I am also very grateful to my second supervisor Professor Uemura, and Professor Watanabe of Nara Institute of Science and Technology (NAIST), for their comments on my work and detailed correction of my original manuscript.

I am indebted to Dr. Miyamoto and Dr. Shimbo for constructive and fruitful discussions and advise.

Thanks are due to all members of Natural Language Processing Laboratory for their discussions and helpful comments. My time at NAIST was made more enjoyable by the friendship of each of them as well.

Finally, I owe the most gratitude to my parents, who always support and encouraged me to pursue my dreams and to strive to do my best.

# References

[1] Chinatsu Aone and Mary Ellen Okurowski. Trainable, scalable summarization using robust nlp and machine learning. In *the 17th International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics*, pages 62–66, 1998.

[2] Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, Vol.22, No.2*, pages 249–254, 1996.

[3] Jean Carletta et al. The reliability of a dialogue structure coding scheme. *Computational Linguistics, Vol.23, No.1*, pages 13–31, 1997.

[4] Masakazu Fujio and Yuji Matsumoto. Japanese dependency structure analysis based on lexicalized statistics. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing*, pages 88–96, 1998.

[5] B. Grosz and C. Sidner. Attention, intention, and the structure of discourse. *Computational Linguistics, Vol.12, No.3*, pages 175–204, 1986.

[6] M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English.* Longman, 1976.

[7] Marti Hearst. Multi-paragraph segmentation of expository text. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 9–16, New Mexico State University, Las Cruces, New Mexico, 1994.

[8] J. Hirschberg and D. Litman. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics, Vol.18, No.4*, pages 501–530, 1993.

[9] Jerry R. Hobbs, editor. *Literature and Cognition.* Center for the study of language and information, 1990.

[10] Hongyan Jing. Sentence simplification in automatic text summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'00)*, pages 310–315, 2000.

[11] Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Summarization evaluation methods: Experiments and analysis. In *Intelligent Text Summarization *Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06*, pages 51–59, Mar 1998.

[12] 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, 1999.

[13] Hongyan Jing and Kathleen R.McKeown. The decomposition of human-written summary sentences. In *SIGIR '99 *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, University of California, Berkeley*, pages 129–136, 1999.

[14] Naoto Katoh and Noriyoshi Uratani. A new approach to acquiring linguistic knowledge for locally summarizing japanese news sentences. *Journal of Natural Language Processing, Vol.6 No.7*, 1999.

[15] Kevin Knight and Daniel Marcu. Statistics-based summarization — step one: Sentence compression. In *National Conference on Artificial Intelligence (AAAI)*, pages 703–710, 2000.

[16] Taku Kudo and Yuji Matsumoto. Japanese dependency analysis based on support vector machines. *EMNLP/VLC 2000*, 2000.

[17] Susumu Kuno. *Danwa no Bunpo (In Japanese)*. Taishukan, 1978.

[18] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Research and Development in Information Retrieval*, pages 55–70, 1999.

[19] Hans P. Luhn. The automatic creation of literature abstracts. *IBM Journal, Vol.2, No.2*, pages 159–165, 1958.

[20] Inderjeet Mani, Eric Bloedorn, and Barbara Gates. Improving summaries by revising them. In *37th Annual Meeting of the Association for Computational Linguistics,Proceedings of the Conference*, pages 558–568, Jun 1999.

[21] William C. Mann and Sandra A. Thompson. *Rhetorical Structure Theory: A Theory of Text Organization*. Tech. Report ISI/RS-87-190, 1987.

[22] Daniel Marcu. Improving summarization through rhetorical parsing. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 206–215, Aug 1998.

[23] Daniel Marcu. To build text summaries of high quality, nuclearity is not sufficient, 1998.

[24] Daniel Marcu. The automatic construction of large-scale corpora for summarization research. In *SIGIR '99 *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, University of California, Berkeley*, pages 137–144, 1999.

[25] Daniel Marcu. A decision-based approach to rhetorical parsing. In *37th Annual Meeting of the Association for Computational Linguistics,Proceedings of the Conferrence*, pages 365–372, Jun 1999.

[26] Johanna D. Moore and Martha E. Pollack. A problem for rst: The need for multi-level discourse analysis. *Computational Linguistics, Vol.18, No.4*, pages 537–544, 1992.

[27] Jane Morris and Graeme Hirst. Lexical cohesion, the thesaurus, and the structure of text. *Computational Linguistics, Vol.17, No.1*, pages 21–48, 1991.

[28] Jane Morris and Graeme Hirst. Text tiling: A quantitative approach to discourse segmentation. *Technical report, University of California*, 1993.

[29] Hidetsugu Nanba and Manabu Okumura. Producing more readable extracts by revising them. *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, pages 1071–1075, 2000.

[30] Kenji Ono, Kazuo Sumita, and Seiji Miike. Abstract generation based on rhetorical structure extraction. In *COLLING-94*, pages 344–348, 1994.

[31] C. Paice. Constructing literature abstracts by computer: Techniques and prospects. In *Information Processing and Managemen 26*, pages 171–186, 1990.

[32] Rebecca J. Passonneau and Diane J. Litman. Combining multiple knowledge sources for discourse segmentation. In *Proc. of Meeting of the Association for Computational Linguistics*, pages 1133–1140, 1995.

[33] J. Ross Quinlan. *C4.5: Programs for Machine Learning.* The Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann, 1992.

[34] K. Sumita, K. Ono, and Others. A discourse structure analyzer for japanese text. In *Proc. of International Conference of Fifth Generation Computer Systems*, pages 1133–1140, 1992.

[35] Vladimir N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, 1998.

[36] Marilyn A. Walker, Aravind K. Joshi, Ellen F. Prince, and Others, editors. *Centering in Discourse.* Oxford University Press, 1998.

[37] Matilyn Walker et al. Japanese discourse and the process of centering. *Computational Linguistics, Vol.20, No.2*, pages 193–232, 1994.

# List of Publications

## Journal Papers

- Kazuhiro Takeuchi and Yuji Matsumoto, Relation between Text Structure and Linguistic Clues: An Investigation on Text Structure of Newspaper Articles, Journal of the Mathmatical Linguistics Society of Japan, Vol 22, No. 8, pp.319-334, March 2001. (In Japanese)

- Kazuhiro Takeuchi and Yuji Matsumoto, Sentence Reconstruction in Summary Generation: An Investigation using Autometed Alignment, Journal of Natural Language Processing. (Under Review) (In Japanese)

## International Conference

- Kazuhiro Takeuchi, Role of Text Structure for Summary Generation: Clues for Sentence Combination, 38th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop, pp.68-74, Hong Kong, October 2000.

- Kazuhiro Takeuchi and Yuji Matsumoto, Acquisition of Sentence Reduction Rules for Improving Quality of Text Summarization, Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001), Tokyo, pp.447-452, November 2001.

## Other Publications

- Kazuhiro Takeuchi and Yuji Matsumoto, An Empirical Analysis of Text Structure as a Basis for Automated Text Summarization, IPSG SIG NOTE, 2001-NL-147, pp. 21-28, January 2002. (In Japanese)

- Kazuhiro Takeuchi and Yuji Matsumoto, Acquisition of Sentence Simplification Rules for Improving Quality of Text Summaries, IPSG SIG NOTE, 2001-FI-63, pp.137-144, July 2001. (In Japanese)

- Kazuhiro Takeuchi, Satoshi Yamada, Kiyota Hashimoto and Yuji Matsumoto, Yoyakuni Okeru Bun KanRyakuKeikouno Tyousa, Proceedings of The 7th Annual Meeting of Japanese Cognitive Science Society, pp.82-83, June 2001. (In Japanese)

- Keiichi Yoshino, Kazuhiro Takeuchi and Yuji Matsumoto, Zero Pronoun Identification using Support Vector Machines, Proceedings of The 7th Annual Meeting of The Association for Natural Language Processing, pp.506-509, March 2001. (In Japanese)

- Tsutomu Hirao, Mamiko Hatayama, Satoshi Yamada, and Kazuhiro Takeuchi, Text Summarization based on Hanning Window and Dependency Structure Analysis, Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, National Institute of Informatics, pp.5-248–253, March 2001.

- Kazuhiro Takeuchi and Yuji Matsumoto, Role of Text Structure for Automated Summarization, ISPJ SIG NOTE, 2000-NL-138, pp.9-16, July 2000. (In Japanese)

- Kazuhiro Takeuchi and Yuji Matsumoto, A Study of Role of Text Strucutre, Proceedings of The 6th Annual Meeting of The Association for Natural Language Processing, pp.368-371, March 2000. (In Japanese)

- Kazuhiro Takeuchi and Yuji Matsumoto, An Empirical Analysis of Text Structure as a Basis for Automated Text Summarization, IPSG SIG NOTE, 99-NL-133, pp.61-68, September 1999. (In Japanese)

- Kazuhiro Takeuchi and Yuji Matsumoto, A Workbench for Discourse Structure Analysis, Proceedings of The 5th Annual Meeting of The Association for Natural Language Processing, pp.239-242, March 1999. (In Japanese)