Yet another summarization system with two modules using empirical knowledge

Kiyonori Ohtake†, Daigo Okamoto, Mitsuru Kodama, Shigeru Masuyama
Department of Knowledge-based Information Engineering,
Toyohashi University of Technology,
Toyohashi 441-8580, Japan
kohtake@slt.atr.co.jp, {okamoto, kodama}@smlab.tutkie.tut.ac.jp,
masuyama@tutkie.tut.ac.jp

Abstract

We previously proposed a summarization system, GREEN, for Japanese newspaper editorials. However, GREEN is not suitable for summarizing ordinal newspaper articles which are different from newspaper editorials. To participate in subtasks A-1 and A-2 of TSC (text Summarization Challenge) in NTCIR-2, we developed a new summarization system from scratch which copes with both ordinal articles and editorials in a Japanese newspaper.

The new summarization system resulted in good evaluations: the mean value of all evaluations held the foremost place among ten systems in subtask A-1 and nine systems in subtask A-2, respectively.

Keywords: *Summarization System, Deletion of modifiers, Extracting sentences, Abstracting*

1 Introduction

We previously proposed a summarization system GREEN for Japanese newspaper editorials. It chooses sentences stating opinions as important based on dependency structure, and summarizes by sentences reduction and deletion of noun modifier, etc. However, GREEN is not suitable for summarizing ordinal newspaper articles, because the news-report-styled newspaper articles are different from newspaper editorials.

Thus, we developed, from scratch, a new summarization system which copes with both ordinal articles and editorials in Japanese newspaper articles to participate in subtasks A-1 and A-2 of TSC in NTCIR-2.

The new summarization system is designed to avoid omission of important information, namely to make informative summarization. Therefore, the aim of the system is different from that of GREEN which aims to make natural summaries.

The system was evaluated by extraction of important sentences in subtask A-1 and is evaluated by comparing with human-made summaries in subtask A-2.

The system was composed of two components, an extract-type summarizer for subtask A-1 and an abstract-type summarizer for subtask A-2.

The extract-type summarizer chooses important sentences due to a level of importance attached for each sentence. And it outputs the selected sentences.

Some features on surface information decide the level of importance for each sentence. Moreover, different weight are attached according to whether the input is a newspaper article or a newspaper editorial.

The abstract-type summarizer summarizes sentences by deleting multiple modifiers for nouns and illustrations and by paraphrasing. For this purpose, it employs a parser KNP.

Our new summarization method focuses on multiple modifiers to make natural summary. Mikami et al.[4] also proposed a method which summarizes each sentence by deleting noun modifiers. However, it sometimes deletes some modifier whose removal causes loss of important information. Consequently, we make our method prudent so that the system does not delete important information.

In the field of automatic summarization, there are some researches which only use surface information[5]. Yamasaki et al.[11] and Wakao et al.[9] proposed a method of paraphrasing for TV news manuscript. And Kodama et al.[3] proposed extraction of summarization knowledge from direct quotations.

The system also employs a table to paraphrase some expressions to more concise expressions. Moreover, the system adopts a method of kodama et al.[3] and try to eliminate the direct quotation.

†Currently with ATR Spoken Language Translation Research Laboratories.

End expression of a sentence	「～したい」(want)「～ほしい」(want) 「～と思う」(think)「～と考える」(consider) 「望まれる」(hope)「～かもしれない」(may)
Terms	「大切」(importance)「必要」(need)「期待」(expectation)「残念」(regrettable) 「注目」(attention)「課題」(subject)「べき」(should)「はず」(should)

Table 1. Example of dictionary of opinion sentences

2 System configuration

This system is implemented on Vine Linux 2.0 using Perl. The system comsists of two components: an extract-type summarizer and an abstract-type summarizer.

In the extract-type summarizer, some features, main terms, high frequency words, location information in a paragraph, etc., decide the weight of each sentence, and the predetermined number of sentences are selected from the sentences with the longest weight.

In the abstract-type summarizer, the system selects sentences to suit the predetermined number of characters based on the weights of sentences by the extract-type summarizer. And this part summarizes each sentence with the KNP.

3 Extract-type summarizer

The extract-type summarizer is composed of five components illustrated in Fig. 1.

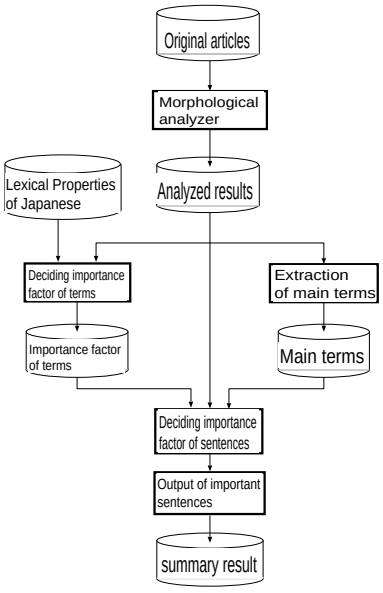

Figure 1. The outline of the extract-type summarizer

3.1 Morphological analyzer

In the morphological analyzer, the system employs a morphological analyzer, JUMAN[1]. An analyzed result is used for determining importance factor of words, extraction of main terms and determining importance factor of sentences.

3.2 Determining importance factor of terms

We consider that high frequency words in an articles are strongly related to the author's opinion. Thus, these words are important. However, these words are unimportant if they appear in too many other articles. Thus, the importance factor of each term is decided by the following expression.

$$\text{Importance factor} = \frac{\text{Word frequency in articles}}{\log(\text{Word frequency in Lexical Properties of Japanese})}$$

This formula is based on the idea of $tf \cdot idf$, where the word frequency of Lexical Properties of Japanese[1] is used instead of idf to reflect the commonness of the term.

The Lexical Properties of Japanese contains frequency counts for terms and characters which appeared in all articles in 14 years (1985 - 1998) of the "Asahi" newspaper. 340,000 words are extracted by morphological analysis, and the frequencies of occurrence for each word and character are counted.

3.3 Extraction of main terms

We define main terms as nouns, which are implicated in the theme of the article. We can assume that a headline of an article is an ultimate summarization of the article. In addition, a head line includes main terms for the article.

Thus, we regard all nouns in the headline as the main terms. In addition, we regard undefined terms at morphological analysis written in KATAKANA or alphabet as nouns.

In GREEN, proper nouns are considered main terms, and is also defined by employing a thesaurus, KADOKAWA RUIGO SHIN JITEN[6]. However, we

[1]http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

do not employ any thesaurus in the system, because we consider that all nouns in the headline are relevant to the theme of the article. In addition, we suppose that each relevance level of the theme is in proportion to the importance factor of terms. Moreover, we suppose that if a sentence contains some nouns included in other sentences previously judged important, the sentence would be as important as other important sentences.

The system extracts the main terms from different location depending on whether newspaper articles or editorials are treated. We consider that the first paragraph of newspaper articles contains important terms, because an author tends to write important terms in the first paragraph. We suppose that the first sentence in a newspaper article and the last opinion sentence in an editorial contain a lot of important terms. Here, an opinion sentence will be defined in subsection 3.4.6.

This is because, the first sentence in a newspaper article offers new information to readers having few knowledge about premise information, and the last opinion sentence in an editorial expresses a conclusion. Thus, for the newspaper articles and editorials, a noun contained at the location below is defined as a main term.

- the first paragraph in newspaper articles
- the first sentence and the last opinion sentence in editorials

3.4 Deciding importance factor of sentences

The system decides importance factor of each sentence by employing the following features: the existence of main terms, the importance factor of terms, a structure of articles, location information in a paragraph, etc. Each parameter is computed by summing up values decided empilically. However, newspaper articles are different from newspaper editorials in some features, i.e., sentences which state facts are more important and sentences at the beginning of an article are important.

3.4.1 Main terms

If a main term in a sentence is a subject term(having the nominative case), we attach heavy weights on the sentence.

3.4.2 High frequency word in articles

We suppose that terms having high frequency in an article are relevant to insistence of the author. Thus, sentences containing high frequency terms are important. Therefore, the system gives weights to sentences if they have high frequency terms.

3.4.3 Structure of articles

In news paper articles, the first sentence of an article is important[2]. In addition, more important information tend to be written at the beginning of the article. Thus, we should attach weights to paragraphs in and around the beginning of the article.

Meanwhile, the last paragraph in a newspaper article contains information that readers consider to be interest. The system attaches weights to the last paragraph.

3.4.4 Location information in a paragraph

The first sentence of each paragraph offers new information to readers who do not have premise information. The last sentence of each paragraph is written with special intention to conclude the paragraph.Thus, the system attach large weights to the first sentence and the last sentence of each paragraph.

3.4.5 Unimportant sentence

There are some unimportant sentences in articles such as titles of paragraphs, supplementary explanations, etc. These sentences tend to contain the signs(e.g., ◇, ＝, etc.) peculiar to the newspapers. In addition, direct quotations can be understood by pre and post contexts.

We defined sentences containing the signs and the direct quotations as unimportant sentences and attached small weights on it.

3.4.6 Opinion sentence and phenomenon sentence

Sentences in articles are classified into two groups[10]. A sentence in the first group tells one's opinions and that in the second group tells facts. An opinion sentence is defined as a sentence which expresses author's insistence, opinions or hopes. A phenomenon sentence is defined as a sentence which expresses accidents, facts or phenomena.

In editorials, opinion sentences tend to be important. Thus, we adopted different weighting for editorials from newspaper articles.

To extract the opinion sentences, we should pay attention to expressions of the end of sentences, i.e.,「～が必要である (need)」,「～すべきである (should)」. The system extracts opinion sentences by matching to pattern of 55 rules, based on a method of GREEN. The pattern table defines dictionary of opinion sentences.

The dictionary for extracting opinion sentences is illustrated in Table 1.

3.4.7 Weighting importance factor

The importance factor of each sentence is computed by summing up points that are shown in Tables 2, 3

Condition	Point
Subjects are main terms	Importance factor of terms × 10
Containing main terms	Importance factor of terms × 2
Containing high frequency word	Importance factor of terms × 1
First sentences of each paragraph	20
Last sentences of each paragraph	10
Unimportant sentences	Importance factor of sentence :1/10

Table 2. Weighting in common with newspaper articles and editorials

Condition	Point
First paragraph	100
Second paragraph	50
Third paragraph	20
Last paragraph	importance of sentence × 10
Except	0

Table 3. Weighting for newspaper articles

and 4(but only an unimportant sentence is divided by points attached to it).

These values of weighting parameters are decided by the heuristics based on the result of DRYRUN. Thus, we can say that the parameters are suitable for evaluations in NTCIR-2.

3.5 Output of important sentences

To output important sentences, the system fixes an order of priority based on the importance factor of each sentence, and selects the sentence to the predetermined number of sentences.

Deciding preference ranking In deciding preference ranking, the system fix an order of the priority in order of the importance factor of each sentence. But we decide that the ranking of first sentence in articles is the first without the importance factor.

Output sentences The system selects sentences in accordance with established priorities and outputs them.

Condition	Points
First paragraph	20
Opinion sentences	10
Except	0

Table 4. weighting for editorials

4 Abstract-type summarizer

The abstract-type summarizer is composed of two components, a component of summarizing a sentence and a component of selecting sentences. An outline of the process in the abstract-type summarizer is shown in Fig. 2.

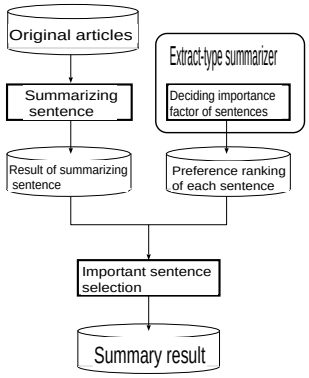

Figure 2. The outline of the abstract-type summarizer

4.1 Overview of summarizing sentence

In summarizing sentence, the system summarizes each sentence by six methods based on syntax analysis by a parser, KNP.

- Deletion of supplementive explanation
- Deletion of expression of direct quotation
- Deletion of multiple modifiers for nouns
- Deletion of illustration
- Paraphrasing
- Deletion of the head of sentences conjunction

These processings are based on the following concepts.

- The system summarizes each sentence by deleting parts of prolixity and does not process plural sentences as a unit.

- The system aims at not deleting an important information, and retains the naturalness.

Moreover, the system employs a parser KNP without using the verb dictionary of IPAL[8]. This is because, the number of verbs in the dictionary of IPAL does not reach a practical level for the purpose of the own system, and that verbs in IPAL have different effect on an accuracy of analysis from verbs not in IPAL if we use KNP with IPAL option.

4.1.1 Deletion of supplementive information

There are many supplementive expressions that in circle parentheses are FURIGANA or abbreviation, etc. in articles.

Thus, the system deletes supplementive expressions in circle parentheses or can be distinguished from other portion by mark(＝, ＜＞, etc.) .

4.1.2 Processing of expression of direct quotation employing heuristics

An expression of direct quotation is unimportant, and we can understand a content without it by seeing pre and post contexts. The first sentence in direct quotation explains facts, the other sentences states opinions.

Thus, we attach importance to paragraphs at the utterance of opinions. If two or more sentences are included in a direct quotation, the system deletes the first sentence. However, if the first sentence is connected to the second sentence by demonstrative, the system does not delete the first sentence, as doing so makes resulting summaries unnaturally.

Example 1

「うちの事務次官と小沢さんの二人三脚で消費税率アップをもくろんでいるとマスコミで取り上げられたからね。 今度の大臣との関係に関心がないといったらウソになりますよ」("It was taken up in mass communications when the rate rise of a consumption tax was planned with the two person tripod of our Administrative Vice-Minister and Mr. Ozawa. So, it will become a lie if I say that I'm not interested in the relation with this minister.")

↓

「今度の大臣との関係に関心がないといったらウソになりますよ」("It will become a lie if I say that I'm not interested in the relation with this minister.")

In addition, there is another case that the next expression of a direct quotation is a summary of the quotation[3]. In this case, the sentence expresses a statement that the same opinions are expressed by both clauses and sentences.

When the statement matches some pattern made by heuristics, we suppose that the sentences are retained natural, even if it deletes all parts of direct quotations.

Example 2

前副社長側は撤回の上申書を検察に提出しているが、検察側は「捜査段階で事実を認めていた」と主張して、タイミングを計って証拠申請する構え。(Although the former vice president side has submitted the revocatory written statement to criminal investigation, prosecutors claimed "the fact was accepted in the criminal-investigation stage", and they are planning to make the evidence application with precise timing.)

↓

前副社長側は撤回の上申書を検察に提出しているが、検察側はタイミングを計って証拠申請する構え。(Although the former vice president side has submitted the revocatory written statement to criminal investigation, prosecutors are planning to make the evidence application with precise timing.)

4.1.3 Deletion of multiple modifiers for nouns

Two or more adnominal forms, which modify one noun, are defined as a multiple modifiers. In particular, two parts which modify one noun is defined as double modifiers. In this system, since most multiple modifiers are double-modifiers, only double-modifiers are treated. Therefore, the case when three or more modify a noun, is not treated.

When there is a double modifier, even if one of the two clauses is deleted, a meaning is assumed not changing seriously in many cases, and one of adnominal forms is deleted.

However, it is dependent on the kind of each part of speech which composes a clause where the former and latter adnominal part are deleted.

Thus, we employ rules for the elimination of an adnominal part of double modifiers. Each rule consists of three components, former adnominal part, latter adnominal part and modified.

In the system, if the rules created by the heuristics matched each part, a latter adnominal part is deleted. Otherwise, a former adnominal part is deleted.

In addition, rather than the verbs, deleting an adjective preferentially etc., we consider that naturalness is not spoiled as important.

The rules are illustrated in Table 5.

The example of deletion of multiple modifiers for nouns is shown below.

Example 3

政党が軽い存在となり、政党に対する 慢性的な 不信が渦巻いている。(A political party serves as a light existence and the chronic distrust to a political party is whirling.) ↓

政党が軽い存在となり、政党に対する不信が渦巻いている。(A political party serves as a light existence and the distrust to a political party is whirling.)

Former modifier	Latter modifier	Modified
動詞連体形 (participial adjective form of verbs)	イ形容詞 (*i*-adjective)	名詞 (noun)
動詞連体形 (participial adjective) form of verbs)	ナ形容詞 (*na*-adjective)	名詞 (noun)
～の (*no*)	数量 (quantity)	-
～という (*toiu*)	～の (*no*)	～の (*no*)

"-" means do not care.

Table 5. Example of rules which deletes a back modifier element

We suppose that the parser KNP have a possibility of mistaking analysis, and we cope with the mistaken result.

Processing of syntax over a thematic part In syntactic analysis, a modifier part does not accept the relation over a thematic part. Because, such sentence of a relation is easy to mistake during the analysis by KNP.

In addition, the thematic part is defined as a clause ended by the thematic particle or the collection particle.

Exception handling by the pattern match When each element of double modifiers is the structure which KNP tends to mistake, exception handling using fine rules is needed.

When double modifier elements matched the rules created by the heuristics, the system does not delete the double modifier elements.

The example of a rule which does not delete the double modifier elements is shown in Table 6.

Former modifier	Latter modifier	Modified
～の (*no*)	～の (*no*)	-
-	-	～との (*tono*)
-	～な (*na*)	こと (*koto*)
～のは (*noha*)	～という (*toiu*)	-
動詞 (verb)	～ための (*tameno*)	-
動詞 (verb)	イ形容詞 (*i*-adjective)	-
イ形容詞 (*i*-adjective)	～の (*no*)	-
～が (*ga*)	～れる (*reru*)	-
指示詞 (demonstrative)	～の (*no*)	-

"-" means do not care.

Table 6. Example of rules which does not delete

Processing of the structure with the possibility of multiple modifiers Multiple modifiers may not be correctly detected due to the analysis error of KNP.

In this system, when in the sentence of the structure of "adnominal clause - noun +の - modified", irrespective of the syntactic-analysis result of KNP, it is regarded as multiple modifiers.

However, if the result by elimination of the adnominal clause is unreadable, we would do nothing. This is judged from the noun in "noun + の". If the noun is abstract, the attributive function of "noun + の" will not affect the modified. Thus, if the noun is abstract, we will do nothing. Whether a noun is abstract or not is judged from a thesaurus, Goi-Taikei[7]. To obtain semantic codes from the Goi-Taikei, we employed the ALTJAWS Ver.2.0: a morphological analyzing library for Japanese.

4.1.4 Deletion of illustration

Illustration is considered to be a modifier in a broad sense, and it is assumed that a sentence sense is not changed by its removal.

This system deletes clauses of "～などの (*nadono*)" and clauses of "～などで (*nadode*)" which modify a verb.

Example 4

現場で取り押さえられ、強盗殺人未遂容疑などの現行犯で逮捕された中学３年生の少年は、「けん銃を奪おうと思った」と供述しているという。(The boy, the third-year student in a junior high school, was captured on the spot, and arrested in the act of burglar attempted murder suspicion etc. It is said that he stated "I thought that a handgun would be taken.")

↓

現場で取り押さえられ、現行犯で逮捕された中学３年生の少年は、「けん銃を奪おうと思った」と供述しているという。(The boy, the third-year student in a junior high school, was captured on the spot, and arrested. It is said that he stated "I thought that a handgun would be taken.")

4.1.5 Paraphrasing

A tedious expression in an article can be summarized by paraphrasing of making a briefer expression.

This system paraphrases based on 96 rules created with the heuristics focusing on expressing briefly in the end of a sentence.

However, the rules are newly created aiming at the customizing for a newspaper article, as the rule created by paraphrasing of Wakao et al.[9], Yamasaki et al.[11] were customized for TV news articles.

The Paraphrasing rules are illustrated in Table 7.

	All articles		Only newspapers		Only editorials	
	Recall(%)	Precision(%)	Recall(%)	Precision(%)	Recall(%)	Precision(%)
10%	33.7	33.7	47.8	47.8	19.6	19.6
30%	45.1	45.1	49.2	49.2	41.1	41.1
50%	61.2	61.2	63.4	63.4	59.0	59.0
ave	46.7	46.7	53.4	53.4	39.9	39.9

Table 8. The evaluation result in subtask A-1

Before paraphrasing	After paraphrasing
入るだろう。	入る。
(will enter)	(enters)
変えることにもなる。	変える。
(also become changing)	(changes)
声明した。	声明。
(announced that)	(declaration)
狙いでもある。	狙い。
(also an aim)	(aim)
決められないでいる。	決められず。
(have not been decided)	(not decided)
決まらないようだ。	決まらない。
(seems that it is not decided)	(not decided)
ことである。	こと。
(*koto – dearu*)	(*koto*)
だからである。	だからだ。
(*dakara – dearu*)	(*dakara – da*)
になるだろう。	に。
(*ni – narudarou*)	(*ni*)

Table 7. Example of paraphrasing rules

4.1.6 Deletion of the head of sentences conjunction

The system does not consider relation among sentences into consideration, in order to extract important sentence. Therefore, the conjunction at the head of a sentence does not act its primary role in produced summary.

Thus, all the conjunctions of the head of each sentence are deleted in this system.

4.2 Important sentence selection

In an important sentence selection, the sentence is chosen based on each sentence summarized in the sentence reduction and preference ranking of each sentence. Here, the preference ranking of each sentence is the value calculated in the preference ranking of the extract-type summarizer.

The algorithm of the important sentence selection is shown as follows.

Algorithm of the important sentence selection

Step 1. Place sentences in the order of importance by the extract-type summarizer.

Step 2. Adopt the summarization method to all sentences except paraphrasing.

Step 3. Select sentences in order of importance to make a summary until the total length of summary exceeds the predetermined length of summary.

Step 4. Adopt the summarization method other than paraphrasing to all sentences again.

Step 5. If the total length of the summary is less than the predetermined one, the summary would be outputted and processing would be terminated.

Step 6. Apply the paraphrasing by tables.

Step 7. If the total length of the summary is less than the predetermined one, the summary would be outputted and processing would be terminated.

Step 8. Eliminate the most unimportant sentence, which was selected as the last sentence at step 3., from the summary, and search a suitable sentence, whose adoption satisfies the length constraint in order of importance from sentences which have not been selected yet.

Step 9. If adoption of a sentence causes violation of the length constraint, i.e., the total length of selected sentences exceed the predetermined length, we would adopt the paraphrasing to the sentence, and the summary would be outputted and processing would be terminated.

Step 10. If addition of a sentence satisfies the length constraint, the summary would be outputted and processing would be terminated.

5 Evaluation

We participated in subtasks A-1 and A-2 among the tasks of TSC(Text Summarization Challenge) in NTCIR-2, and evaluations on the extract-type summarizer and the abstract-type summarizer were performed.

5.1 Extract-type summarizer

Subtask A-1 evaluated a summary on the basis of the coincidence between the important sentences which man chose.

The following two formulas were used as evaluation measures.

$$\text{Recall} = \frac{\text{The number of texts for which the subjects judged correctly as relevant}}{\text{The total number of relevant texts}}$$

$$\text{Precision} = \frac{\text{The number of texts for which the subjects judged correctly as relevant}}{\text{The total number of texts judged as relevant by subjects}}$$

These values are calculated for every rate of a summary(10%,30%,50%).

The averaged results and the results on limitation of newspaper articles and editorials, respectively, are shown in Table 8. Consequently, it turns out that the precision of important sentence extraction of an editorial is low as a whole, compared with newspaper articles, and in particular, the precision over the editorial in the case of summarization to 10% is extremely low.

It is mentioned that the parameter about the opinion sentences is small and description is insufficient in the dictionary of the opinion sentences as a cause.

By expanding the dictionary of opinion sentences, the precision of the summary to 10% with an editorial is improved to 24.3%, and the whole average is improved to 47.4%

However, even if it makes the parameter about the opinion sentences heavy and it adopts the opinion sentences compulsorily like GREEN, the improvement in the precision beyond this is difficult. This means that the opinion sentences are not necessarily important in editorials. Thus, it is necessary to classify opinion sentences more finely to the following two classes: the sentence which expresses the opinion of an author and that states the opinion on future development. In addition, it is also necessary to give the different weighting.

Moreover, there was an error by the forcible adoption of the first sentence about editorials. The first sentence in editorials does not necessarily become the whole outline, which depends on the author. Therefore, it is necessary to adopt the first sentence after judging whether it shows the whole outline.

As for the newspaper articles, a good result was obtained compared with the editorials. However, when seen for every article, there was the articles with low precision. Although important sentence extraction to the newspaper articles of such important composition that it is close to the beginning, it may perform inadequate extraction to other articles. Therefore, the articles are not simply classified into the editorials and the newspaper articles, the measure of classification according to the composition of the articles finely is required.

5.2 Abstract-type summarizer

For subtask A-2, two kinds of evaluations, subjective evaluation and content-based evaluation, were performed.

In the subjectivity evaluation, the evaluator (one person) read the summaries texts by the system. Then, evaluate and score them in terms of how readable they are, and how well the content of the text is described in the summary. The scores are one of 1, 2, 3, and 4 where 1 is the best and 4 is the worst, i.e., the lower score, the better evaluation is.

In the content-based evaluation, morphological analysis was done to the system results and human summaries, and only content words(morpheme), which are nouns, verbs, adjectives, and undefined words, were selected. Moreover, term weights in each summary were calculated by $tf \cdot idf$ measure. Then the distance between the document vector of human summary and a system result were computed by the cosine of the angle between vectors, and we observed how close the two summaries based the content word.

Moreover, the following are two kinds of correct answers as a summary: the summary in which human did free creation (following, FREE), and the summary which human created by important part extraction (following, PART). Each evaluation was performed by specifying the predetermined number of words that it becomes rate of summary 20%, and 40%.

The result of the subjective evaluation is shown in Table 9

	Evaluation value
Readability 20%	2.53
Contents evaluation 20%	2.93
Readability 40%	2.73
Contents evaluation 40%	2.77

Table 9. Subjectivity evaluation

	Evaluation value
FREE20%	0.472670
FREE40%	0.648264
PART20%	0.513655
PART40%	0.660800

Table 10. Content-based evaluation

In the subjective evaluation, although the summary to 20% was more readable than the summary to 40%, the content-based evaluation exhibits a little worse result.

As for the content-based evaluation, the summary to 20% was not similar to a correct answer as a summary from the summary to 40%, and it was with the bad result so that the rate of a summary was low like subjectivity evaluation.

This is because the sentence is chosen in the abstract-type summarizer on the basis of the preference ranking searched for in the extract-type summarizer. Therefore, the result in extract-type summarizer influences subjective evaluation in abstract-type summarizer.

Moreover, precision of FREE is lower than that of PART. The advanced summary is required in order to raise the precision to FREE. In order to improve precision to FREE, an advanced summary method like the summary used in case human does a free summary is required.

5.2.1 Summarization rate

The contribution to summarization rate of the deletion (2129 characters) of redundant part performed in the abstract-type summarizer to the sentences (22812 characters) extracted by the extract-type summarizer is about 9%.

In this system, in order to prevent lack of important information, bold deletion was not performed but a prudent summary was performed. This strategy is successful and resulted also in good evaluation.

5.2.2 Comparison of the effect of each method

The abstract-type summarizer consists of six methods. As a result of investigating the number of deletion of characters by each method independently is shown in Table 11.

There are many deletion characters by the deletion of supplementive explanation among six methods. This is because circle parenthesis expression is used abundantly in Mainichi Newspapers. Moreover, the deletion of supplementive explanation is very effective because of the additional information like abbreviated name or the FURIGANA of a difficult Chinese character.

The processing of expression of direct quotation was the result on the whole being hard to use. This is because importance tends to be attached to direct quotation expression and each sentence is hard to adopt in articles.

The processing of expression of direct quotation is classified into deletion of a first sentence, and deletion of the whole direct quotation expression. However, most frequently used was deletion of a first sentence. In order that this method may delete the all sentences in a parenthesis, it has many deletion characters and tends to maintain naturalness. Therefore, it is an effective method.

On the other hand, deletion of the whole direct quotation expression is used only twice in 30 articles, and the result is hard to be called effective. It becomes a subject to consider how the patterns are increased.

Although the deletion of a double modifiers deleted many characters at once, it had the case where naturalness was spoiled by excessive deletion. Therefore, it is necessary to decide and control a rule more carefully.

Moreover, in this system, in order to cope with errors of a syntactic-analyzer, the structure of the sentence which is easy to carry out an analysis error took the object of safe serious consideration of not deleting. But it will be a difficult subject to consider how to cope with it from now on.

Deletion of illustration is used little number of times although its deletion is successful as it maintains naturalness. This reason is that importance information is not attached to the illustration.

Although only a few characters are deleted at each application of paraphrasing, since it is applicable to many sentences, it is an effective method in a summary. However, since the table for paraphrasing is created manually, in order to make it applicable to more sentences, it is necessary to extend a table automatically.

5.3 On naturalness

In GREEN, cohesion analysis was performed on the basis of the abbreviation of a key word or a subject etc., and the sentence with the above sentence and cohesion took the method of adopting the previous sentence. This is for preventing a summary result from becoming unnatural.

Since this system considered that extraction of the important information rather than the naturalness of a sentence, the method, used in GREEN, to retain cohesions on all selected sentences was not taken. Therefore, it is inferior to GREEN from the viewpoint of naturalness, and may summarize unnaturally.

For example, if the sentence where the object is described to be is not adopted when the demonstrative is contained in the sentence, it becomes unnatural as a sentence. Moreover, when a theme changes in an article and only the extracted sentence is read, there is also a problem that a relation of sentences becomes ambiguous in the portion of relation of a theme.

How to solve such a problem is a future subject. However, in the subjectivity evaluation, the evaluation value of readability was about 2.6(average value of 20% and 40%).

This is because minimum disposals, such as deletion of the conjunction in the first of sentences and the forcible adoption of the first sentence, are performed.

Method	Number of deletion characters	Rate(%)	Number of times to use	Number of mean deletion characters
Deletion of supplementive explanation	662	31.1	335	1.97
Processing of expression of direct quotation	298	1.4	8	37.25
Deletion of multiple modifier for nouns	729	34.2	61	11.95
Deletion of illustration	126	5.9	6	21.00
Paraphrasing	262	12.3	74	3.54
Deletion of the head of sentences conjunction	52	2.4	15	3.47

Table 11. Comparison by the methods

Moreover, the fact that the middle sentence in which a demonstrative tends to occur frequently is hard to adopt is also one of the reasons of weighting to the first sentence and the last sentence of each paragraph.

6 Conclusion

The summary system customized for both newspaper articles and editorials was introduced. We participated in subtasks A-1 and A-2 of TSC in NTCIR-2. The evaluated results on the system were the best in average among all participants for both subtasks A-1 and A-2.

To improve the naturalness by employing a method to retain cohesion on a produced summary is left for future subjects.

Acknowledgements

The authors are grateful to Nippon Telegraph and Telephone Corporation for permitting use of ALT-JAWS Ver.2.0, a morphological analyzing library for Japanese, to obtain semantic codes from Goi-taikei.

References

[1] S. Amano and T. Kondo, editors. *the Lexical Properties of Japanese*. SANSEIDO Co., Ltd., 1999.

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

[3] M. Kodama, A. Kataoka, S. Masuyama, and K. Yamamoto. The automatic extraction of summary knowledge using direct quotation expression. In *Proceedings of The Sixth Annual Meeting of The Association for Natural Language Processing*, pages 241–244, 2000. (in Japanese).

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

[5] M. Okumura and H. Nanba. Automated text summarization. *Journal of Natural Language Processing*, 6(6):1–25, 1999. (in Japanese).

[6] S. Ôno and M. Hamanishi. *Kadokawa Ruigo Shinjiten (Kadokawa New Thesaurus)*. Kadokawa Publishing Co., 1981.

[7] S. Ikehara and M. Miyazaki and S. Shirai and A. Yokoo and H. Nakaiwa and K. Ogura and Y. Oyama and Y. Hayashi, editors. *Goi-taikei*. Iwanami Publishing, CD-ROM edition, 1999.

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

[9] T. Wakao, T. Ehara, and K. Shirai. The technique of the summary looked at by the title of a television news program. In *IPSJ SIG Notes NL-122-13*, pages 83–89, 1997. (in Japanese).

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

[11] K. Yamasaki, M. Mikami, S. Masuyama, and S. Nakayama. The title generation summary by paraphrasing for hearing-impaired person. In *Proceedings of The Fourth Annual Meeting of The Association for Natural Language Processing*, pages 646–649, 1998. (in Japanese).