Elimination of Multiple Modifiers in Summarization

Kiyonori Ohtake[†] **and Shigeru Masuyama**

Department of Knowledge-based Information Engineering, Toyohashi University of Technology
Tenpaku-cho, Toyohashi 441-8580, JAPAN.
Email: kohtake@slt.atr.co.jp, masuyama@tutkie.tut.ac.jp

Abstract

We propose a method that summarizes a Japanese sentence. The method aims to produce a natural and readable summary from the sentence. This method eliminates a part of multiple adnominal modifiers including adnominal clauses by employing natural language processing tools[1]: KNP (a parser), and JUMAN (a morphological analyzer). With this proposed method, we participated in subtask A-2 (for producing summaries to be compared with human-prepared summaries) of the TSC [2] (Text Summarization Challenge), one of the tasks of the second NTCIR Workshop[3], which aims to generate summaries as natural as possible. The summarization system with the proposed method resulted in good evaluations: the mean value of all evaluations held the foremost place among nine systems.

Keywords: summarization; multiple modifiers; natural summary.

1 Introduction

This study aims at generating general and natural summary from a sentence. Studies to date on summarization have focused on extracting important parts [1]. Many of them have also employed the sentence as the unit of extraction. However, each sentence requires a more natural and readable summary.

Yamamoto et al. claimed, "A summarization of high quality has to consider not only a discourse analysis for one linguistic phenomenon but also many linguistic phenomena [2]." This claim is applicable to a summarization for a sentence. In other words, we have to take account of various linguistic phenomena to make a natural and readable summary of a sentence.

We propose a method that summarizes a sentence by deleting adnominal modifiers, because such modifiers are considered syntactically unimportant. Eliminating syntactical unimportant parts seems to be a practical approach for robust and natural summarization. On the other hand, in order to detect the syntactically unimportant parts in a sentence, we have to employ a parser to analyze the dependency structure. Unfortunately, there have been few works on summarization methods that employ a parser. The reason why is that a parser with a high accuracy and high speed is unavailable.

By employing a parser, we can accurately detect adnominal parts in a sentence. However, the simple elimination of adnominal parts can cause an unnatural and unreadable summary [3]. On the other hand, we know empirically that eliminating a part of multiple adnominal modifiers retains the readability of a sentence. Considering this, we focus on multiple adnominal modifiers to summarize each sentence.

We propose a summarization method, which positively employs results of a dependency analysis by a parser, for Japanese sentences. The proposed method focuses on multiple adnominal modifiers in a sentence and eliminates a part of them.

2 Summarization Method Focusing on Multiple Modifiers

In this section, we define multiple adnominal modifiers, and propose a summarization method that summarizes a sentence by considering such modifiers.

2.1 Multiple Modifiers for Nouns

In Japanese, we can classify modifiers for nouns, namely adnominal forms, into the following ten types: *koso-ado-*

[†] Currently with ATR Spoken Language Translation Research Laboratories, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, JAPAN.

[1] Both JUMAN and KNP are available from http://pine.kuee.kyoto-u.ac.jp/nl-resource/index-e.html

[2] http://oku-gw.pi.titech.ac.jp/tsc/index-en.html

[3] http://research.nii.ac.jp/~ntcadm/workshop/work-en.html

attributes, attributes, adjectives, nominal adjectivals, noun + '*no*'s, noun + case marker + '*no*'s, noun + *toritate* (picking up)-postpositional particle + '*no*'s, adverb + '*no*'s, relative clauses, and content clauses.

If two or more adnominal forms modify the same noun, we recognize these forms as multiple adnominal modifiers. In particular, when the number of adnominal forms is two, we recognize the two modifiers as double modifiers. The overwhelming majority are double modifiers in Japanese texts. This is because, the forward elements of a noun phrase with three or more adnominal forms for example, "*shiroi* (white) *nagai* (long) *ookina* (big) *te* (hand)," tend to be continuous forms, e.g. "*shirokute* (white) *nagai* [or *nagakute*] (long) *ookina* (big) *te* (hand)." Therefore, double modifiers are the overwhelming majority of multiple modifiers. For this reason, we address double modifiers in this study.

2.2 Special Case of Double Modifiers

There is a special case of double modifiers, "adnominal clause - noun+*no* - modified." For example, "*watashi-ga kiita sakka-no hanashi* (A writer's talk which I heard)" and "*watasi-ga interview-shita sakka-no hanashi* (A writer's talk that I interviewed)" have the same part of speech sequence, but their dependency structures are different as shown in Figure 1.

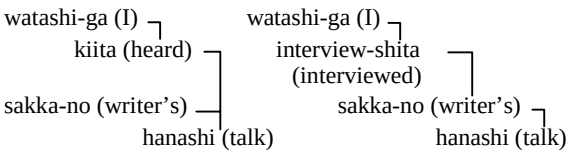

Figure 1: Different dependency structures

Parsing a sentence that includes this pattern, "adnominal clause - noun+*no* – modified", tends to fail in the dependency analysis. In other words, we cannot judge whether the sentence that includes "adnominal clause - noun+no – modified" has a double modifier or not. Accordingly, we treat this pattern as an exception of double modifiers and if we encounter this pattern, we eliminate the adnominal clause. However, if the result obtained by eliminating the adnominal clause is obviously unreadable, we do nothing. This is judged from the noun in "noun+*no*". If the noun is abstract, the attributive function of "noun+*no*" will not affect "modified," and therefore, we do nothing. Whether a noun is abstract or not is judged from a thesaurus, *Goi-Taikei - A Japanese Lexicon* [4]. To obtain semantic codes from *Goi-Taikei*, we employ ALTJAWS Ver.2.0: a morphological analyzing library for Japanese.

2.3 Summarization Method

When we eliminate such multiple adnominal parts, the question is obviously what portions should we eliminate. If we eliminate all of the multiple adnominal parts, the sentence would be unnatural, unreadable, or incomprehensible. However, eliminating only some parts can retain the naturalness of the sentence. To eliminate a part of all multiple adnominal parts, we employ a number of rules derived from knowledge on linguistics, and several heuristics.

We employ 36 rules for the elimination of adnominal parts of double modifiers. Each rule consists of four components, i.e., former adnominal part, latter adnominal part, modified, and action. Table 1 shows examples of these rules.

The basic rule of elimination is that longer or more complex parts should be unremoved, because they work more attributively than others do. That is, the deletion of such longer or more complex parts will lead to unnatural and unreadable summaries.

The heuristics are employed to avoid influences caused by parsing errors, and provide a robust summarization method.

3 Summarization System

We implemented a summarization system consisting of the following two components: a module that selects important sentences, and a module that summarizes each sentence. The latter uses the above elimination method of double modifiers. We employed a morphological analyzer, JUMAN, and a parser, KNP.

Through a formal run of subtask A-2 of the TSC (Text Summarization Challenge) in the second NTCIR (NII-NACSIS Test Collection for IR Systems) Workshop, 20% and 40% of the summaries of 30 newspaper articles were evaluated.

When participating in subtask A-2 of the TSC, it is not sufficient to only eliminate double modifiers. Considering this, we implemented five summarization methods for each sentence besides the proposed method as follows:

Elimination of supplementary explanations: Eliminate additional information expressed in parentheses, etc. For example, 公職選挙法(一九九条)違反 → 公職選挙法違反 (a violation of the election law (the 199th article) → a violation of the election law)

Elimination of the first sentence in direct quotations: Eliminate the first sentence of a direct quotation if two or more sentences compose a direct quotation. For example, そのうえ「明日の公式試合には出なくてええ．背番号も返せ」と言われたという。 → そのうえ「背番号も返せ」と言われたという。(In addition, it is said that he said, "You don't need to participate in tomorrow's regular game. Return the uniform number." → In addition, it is said that he said, "Return the uniform number.") However, this elimination is canceled if a demonstrative is included on and after the second sentence, because the demonstrative may refer to a part of the first sentence.

Elimination of direct quotations: There are some situations in which a summary of a direct quotation follows the direct quotation. These patterns are indicated by

Table 1: Examples of elimination rules for double modifiers

former	latter	Modified	Action
…*no*	…*no*	---	Do nothing
---	---	…*tono*	Do nothing
…*toiu*	…*no*	…*no*	Eliminate the latter part
adnominal clause	adjective	---	Eliminate the latter part
---	…*na*	*koto*…	Do nothing

Here, '---' means any pattern.

Kodama et al. [5], and we employ these patterns to detect direct quotations able to be eliminated. For example, 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．→ 検察側はタイミングを計って証拠申請する構え．(Prosecutors claimed, "the fact was accepted in the criminal-investigation stage", and they are planning to make the evidence application at the precise time. → Prosecutors are planning to make the evidence application at the precise time.)

Elimination of illustrations: Illustrations are considered to be modifiers in a broad sense, and it is assumed that meanings are not changed even if they are eliminated. Therefore, for example, we eliminate the "...*nado no* (such as)" in the pattern "...*nado no* + noun." We also eliminate "...*nado de*", which depends solely on a declinable word. For example, 経済や外交戦略などの専門知識はもとより，→専門知識はもとより，(... as well as special knowledge, such as economics and diplomatic strategies → ... as well as special knowledge), 既に蔵相・外相会合などで取り上げられている→既に取り上げられている (It has already been taken up by the finance minister, at the foreign minister meeting, etc. → It has already been taken up.)

Paraphrasing by table: We paraphrase conjunctions at the beginning of each sentence or expressions at the end of each sentence into shorter expressions or eliminate them. For example, …決まらないようだ．→ …決まらない．(It seems that ... is not decided. → ... is not decided.), そんな中，….　→ ….　(Under the circumstances, ….　→ ….)

We have proposed some summarization methods for the sentences above. However, these summarization methods alone cannot create the short summaries required by subtask A-2 (20% and 40% summaries). We therefore implemented another summarization method by sentence selection [5]. This method utilizes surface information of texts and some heuristics. The sentences of each text are placed in order by a sentence selection based summarizer, and they are then summarized by the summarization methods mentioned above. The overview of the processing of the summarization system (sentence selection based summarizer + sentence summarizer) is shown as follows:

1. Place sentences in the order of importance by the sentence selection based method.
2. Adopt the summarization methods for each sentence.
3. Select sentences in the order of importance to make a summary.
4. Output the summary and terminate the processing.

4 Evaluation

For subtask A-2 of the TSC, two evaluations were carried out. One was a subjective evaluation and the other was a content-based evaluation. Both evaluations were done on two different summary lengths, i.e., 20% and 40%. The targets of the summarization were 30 newspaper articles in total from Mainichi Shimbun, one of the major newspapers in Japan.

In the subjective evaluation, an evaluator read the summaries produced by each summarization system. Then, the evaluator evaluated and scored them in terms of how readable they were, and how much the system covered the important contents of the original articles. The subjective evaluation was done by rating each summary on a scale of one to four (1, 2, 3, or 4, where 1 is the best).

In the content-based evaluation, morphological analysis was done on the system results and human summaries, and only content words (morphemes), i.e., nouns, verbs, adjectives, and undefined words, were selected. Moreover, the term weights in each summary were calculated by the $tf \cdot idf$ measure. Then, the similarities between the document vectors of the human summaries and the system results were each computed by taking the cosine of the angle between the vectors.

Table 2: Results of subjective evaluation

length	type	value(average)	ranking
20%	Readability	2.53(3.16)	1
20%	Content	2.93(3.24)	1
40%	Readability	2.73(3.05)	3
40%	Content	2.77(3.12)	1

In addition, two kinds of human summarization were adopted. One was freely created (FREE), and the other was created by the extraction of important parts (PART). In the content-based evaluation, therefore, the two summaries, 20% and 40%, were compared with the results of the two kinds of human summarization, FREE and PART.

First, we show the results of the subjective evaluation in Table 2. The results show the average values of the 30 summaries by the system. Ten systems, including a base-line system, participated in this evaluation.

Second, we show the results of the content-based evaluation in Table 3. The results also show the average values of the 30 summaries by the system. A total of 11 systems, including two base-line (reference) systems, participated in this evaluation.

Table 3: Results of content-based evaluation

length	type	value(average)	ranking
20%	FREE	0.4727(0.4418)	1
40%	FREE	0.6483(0.6065)	1
20%	PART	0.5137(0.4740)	1
40%	PART	0.6608(0.6342)	1

Table 4: Comparison of six methods

method	number of times to work	number of characters deleted
(1)	61	729
(2)	335	662
(3)	7	277
(4)	1	21
(5)	6	126
(6)	89	314

5 Discussion

In the implemented system, in order to prevent the elimination of important information, and to retain the readability of summaries, prudent eliminations as opposed to bold eliminations are performed. This strategy is successful and results in good evaluations.

However, the compression ratio (ratio of summary length to source length) without the sentence selection based method for the 30 articles is about 91%.

Accordingly, the quality of summaries produced by the implemented system depends on the sentence selection based method. Despite this dependence, the subjective evaluation results show that our proposed methods work sufficiently.

We introduced six methods to summarize a sentence, (1) Elimination of double modifiers, (2) Elimination of supplementary explanations, (3) Elimination of the first sentence in direct quotations, (4) Elimination of direct quotations, (5) Elimination of illustrations, and (6) Paraphrasing by table. What portions of these methods are the most effective? To answer this question, we made a comparison among these six methods to the 30 articles in subtask A-2, and the results are shown in Table 4. The results show that the elimination of double modifiers contributes to the level of compression. Meanwhile, we found an unnatural elimination of double modifiers with the pattern: “adnominal clause - noun+*no* – modified”.

Rigid evaluations of each method are left for future work.

6 Conclusion

We proposed a method that naturally summarizes a Japanese sentence. We implemented a summarization system with this proposed method. In addition, we participated in subtask A-2 of TSC Task to evaluate the proposed method. The strategy we took, i.e., preventing the elimination of important information and retaining the readability of a summary, was found to be successful and resulted in good evaluations.

Acknowledgements

The authors are grateful to Nippon Telegraph and Telephone Corporation for permitting us the use of ALTJAWS Ver.2.0, a morphological analyzing library for Japanese, to obtain semantic codes from *Goi-Taikei*.

References

1 Mani, I. and Maybury, M. T. (eds.): Advances in Automatic Text Summarization, MIT press, 1999.

2 Yamamoto, K., Masuyama, S. and Naito, S. GREEN: An experimental system generating summary of Japanese editorials by combining multiple discourse characteristics. *Journal of Natural Language Processing*, 1995, 2(1), pp. 39 – 55, (in Japanese).

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

4 Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Oyama, Y. and Hayashi, Y.: Goi-Taikei – A Japanese Lexicon, Iwanami Publishing, CD-ROM edition, 1999, (in Japanese).

5 Kodama, M., Kataoka, A., Masuyama, S. and Yamamoto, K. Automatic extraction of summary knowledge using direct quotation expression. In *Proceedings of the 4th Annual Meeting of The Association for Natural Language Processing*. 2000, pp. 241 – 244, (in Japanese).

6 Ohtake, K., Okamoto, D., Kodama, M. and Masuyama, S. Yet another summarization system with two modules using empirical knowledge. In *Proceedings of NTCIR Workshop 2 Meeting*. 2001, http://research.nii.ac.jp/~ntcadm/publication1-en.html.