Analysis and modeling of manual summarization of Japanese broadcast news

Hideki Tanaka, Tadashi Kumano, Masamichi Nishiwaki and Takayuki Itoh
Science and Techinical Research Laboratories of *NHK*
1-10-11, Kinuta, Setagaya-ku
Tokyo, 157-8510, Japan
{tanaka.h-ja,kumano.t-eq,nishiwaki.m-hk,itou.t-gq}@nhk.or.jp

Abstract

We describe our analysis and modeling of the summarization process of Japanese broadcast news. We have studied the entire manual summarization process of the Japan Broadcasting Corporation (NHK). The staff of NHK has been making manual summarizations of news text on a daily basis since December 2000. We interviewed these professional abstractors and obtained a considerable amount of news summaries. We matched the summary with the original text, investigated the news text structure, and thereby analyzed the manual summarization process. We then developed a summarization model on which we intend to build a summarization system.

1 Introduction

Automatic text summarization research has a long history that dates back to the late 50's (Mani and Maybury, 1999). It started mainly with the purpose of information gathering or assimilation, and most of the research has dealt with extracting the important parts of the texts. The summaries obtained with these techniques, so called extracts, have been used for judging the importance of the texts.

We have started research on automatic summarization for the purpose of information dissemination, namely summarization of news texts for broadcast news. Recently, we have studied the entire manual summarization process of the Japan Broadcasting Corporation (NHK).

NHK has been making manual summarizations of news text on a daily basis since December 2000, when it started satellite digital broadcasting. The summarized text has been used for the data service of the digital broadcasting and on Web pages accessible by mobile phones.

We interviewed NHK's professional abstractors and analyzed a considerable amount of news summaries. We matched these summaries with the original news and studied the summarization process based on the results of our analysis and interviews.

In this paper, we report on what we found during the interviews with the abstractors and the results of the automatic text alignment between summaries and the original news together with the word position matching. We also propose a summarization model for an automatic or semi-automatic summarization system.

2 The manual summarization process

Most of the radio and TV news services of NHK are based on a "general news manuscript." We call such manuscripts *the original news* in this paper. The original news is manually summarized into *summary news* that are made available to the public through Web pages and digital broadcasting, as mentioned in section 1.

We asked professional abstractors about the summarization environment and process and in so doing discovered the following.

- Abstractor

The original news is written by NHK reporters, and the text is summarized by different writers, i.e., professional abstractors. Most professional abstractors are retired reporters who have expertise in writing news.

- Compression rate and time allowance

The original news is compressed to a maximum length of 105 Japanese characters. We will

show in section 4 that the average compression rate is about 22.5%. The upper bound is decided from the display design of the data service of digital TV broadcasting. The abstractors must work quickly because the summary news must be broadcast promptly.

- Techniques

The abstractors use only information contained in the original news. They scan the original news quickly and repeatedly, not to understand the full content, but to select the parts to be used in the summary news. The abstractors' special reading tendency has been reported in (Mani, 2001), and we can say the same tendency was observed in our Japanese abstractors. The abstractors focus on the lead (the opening part) of the original news. They sometimes use the end part of the original news.

3 Corpus construction

We planned the summary news corpus as a resource to investigate the manual summarization process and to look into the possibility of an automatic summarization system for broadcast news. We obtained 18,777 pieces of summary news from NHK. Although each piece is a summary of a particular original news text, the link between the summary and the original news is not available.

We matched the summary and original news and constructed a corpus. There have been several attempts to construct <summary text, original text> corpora (Marcu, 1999; Jing and McKeown, 1999). We decided to use the method proposed by Jing and McKeown (1999) for the reasons given below.

As our abstractors mentioned that they used only information available in the original news, we hypothesize that the summary and the original news share many surface words. This indicates that the surface-word-based matching methods such as (Marcu, 1999; Jing and McKeown, 1999) will be effective.

In particular, the word position matching realized in (Jing and McKeown, 1999) seems especially useful. We thought that we might be able to observe the summarization process precisely by tracing the word position links, and we employed their work with a little modification.

As a result, our corpus takes the form of the triple: <summary, original, word position correspondence>.

3.1 Matching algorithm

Jing and McKeown (1999) treated a word matching problem between a summary and its text, which they called the summary decomposition problem. They employed a statistical model (briefly described below) and obtained good results when they tested their method with the Ziff-Davis corpus. In the following explanation, we use the notion of summary and text instead of summary news and original news for simplicity.

(1) The word position in a summary is represented by <*I*>.

(2) The word position in the text is represented by a pair of the sentence position (*S*) and the word position in a sentence (*W*) as in <*S, W*>.

(3) Each summary word is checked as to whether it appears in the text. If it appears, all of the positions in the text are stored in the form of <*S,W*> to form a position trellis.

(4) Scan the *n* summary words from left to right and find the path on the trellis that maximizes the score of formula (1).

$$P = \prod_{i=1}^{n-1} P(I_{i+1} = (S_2, W_2) \mid I_i = (S_1, W_1)) \quad (1)$$

This formula is the repeated product of the probability that the two adjacent words in a summary (I_i and I_{i+1}) appear at positions (S_1, W_1) and (S_2, W_2) in the text, respectively. This quantity represents the goodness of the summary and the text word matching. As a result, the path on the trellis with the maximum probability gives the overall most likely word position match.

Jing and McKeown (1999) assigned six-grade heuristic values to the probability. The highest probability of 1.0 was given when two adjacent words in a summary appear at adjacent positions in the same sentence of the text. The lowest probability of 0.5 was given when two adjacent words in a summary appear in different sentences in the text with a certain distance or greater. We fixed the distance at two sentences, considering the average sentence count of the original news texts.

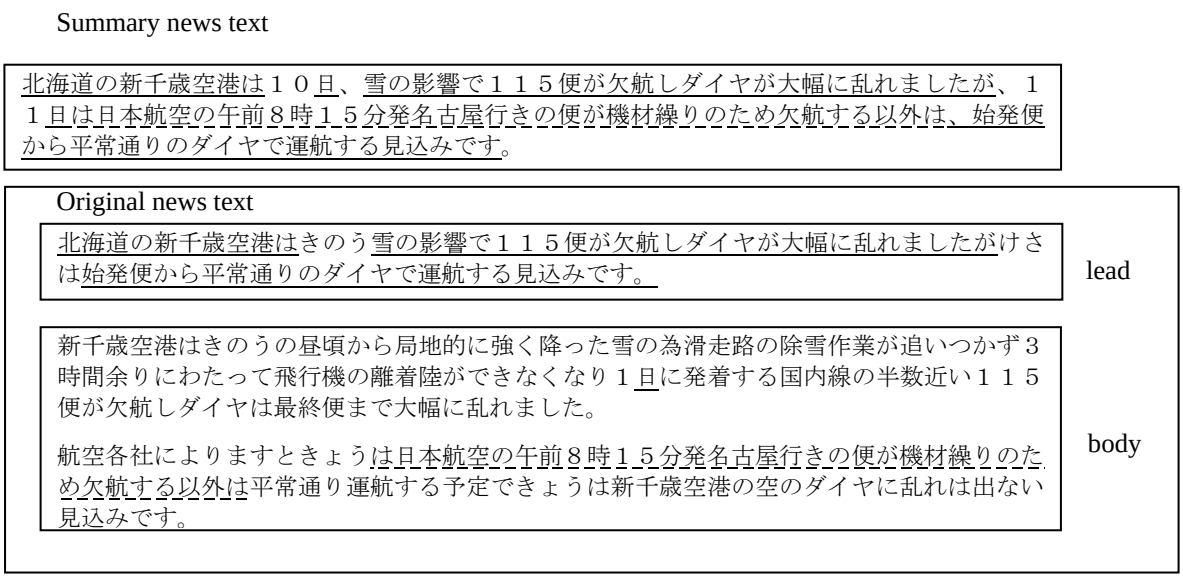

Figure 1. Summary and original news text matching.

Jing and McKeown's algorithm (1999) is designed to treat a fixed summary and text pair and needs some modification to be applied to our two-fold problem of finding the original news of a given summary news from a large collection of news together with the word position matching.

Their method has a special treatment for a summary word that does not appear in the text. It assumes that such a word does not exist in the summary and therefore skips the trellis at this word with a probability of 1. This unfavorably biases news text that contains fewer matching words. To alleviate this problem, we experimentally found that the probability score of 0.55 works well for such a case (This score was the second smallest of the original six-grade score).

We developed a word match browser to precisely check the words of the summary and original news.

3.2 Summary and original news matching

We matched 18,777 summary news texts from November 2003 to June 2004 against the news database, which mostly covers the original news of the period. We followed the procedures below.

- Numerical expression normalization

  Numerical expressions in the original news are written in Chinese numerals (Kanji) and those of the summary news are written in Arabic numerals. We normalized the Chinese numerals into Arabic numerals.

- Morphological analysis

  The summary and original news were morphologically analyzed. We used morphemes as a matching unit. In this paper, we will use morphemes and words interchangeably.

- Search span

  Each summary news was matched against the news written in the three-day period before the summary was written. This period was chosen experimentally.

4 Results and observation

We randomly checked the news matching results and found more than 90% were correct. Some of the summaries were exceptionally long, and we consider that such noisy data was the main reason for incorrect matching. Figure 1 shows a matching example. The underlined (line and broken line) sentences show the word position match.

The word matching is not easy to evaluate because we do not have the correct matching answer. Although there are some problems in the matching, most of the results seem to be good enough for approximate analysis. The following discussion assumes that the word matching is correct.

4.1 Compression rate

Table 1 shows the basic statistics of the summary and its corresponding original news.

We can see that the average compression rate is 22.5% in terms of characters. The average summary news length (109.9 characters per news text) was longer than what we were told (105, see section 2).

We then checked the length of the typical summary texts. We found that the cumulative relative frequency of the summary text with the sentence count from 1 to 4 was 0.99 and was quite dominant. We checked the average length of these summaries and obtained 105.4, which is close to what we were told. We guess that noisy "long summaries" skewed the figure.

	Original	Summary
text counts	18,777	
Ave. sent. count/text	5.13	1.63
Ave. text length (char.)	487.7	109.9
Ave. first line length (char.)	94.9	81.3

Table 1. Basic statistics of summary and original news

4.2 Word match ratio

We measured how many of the summary words came from original news. As our matching result contains word-to-word correspondence, we calculated the ratio of the matched words in a summary text. Table 2 shows a part of the result. It shows the relative frequency of the summary news in which 100% of the words came from the original news reached 0.265 and those that had more than 90% reached 0.970.

Word match ratio	Rel. summary freq.
100%	0.265
More than 90%	0.970 (cumulative)

Table 2. Word match ratio

This strongly suggests that most of the summary news is the "extract" (Mani, 2001), which is written using only vocabulary appearing in the original news. This result is in accord with what the abstractors told us.

4.3 Summary word employment in the original news sentences

The previous section indicated that our summary likely belongs to the extract type. Where in the original news do these words come from? We next measured the word employment ratio of each sentence in the original news and the result is presented in Figure 2.

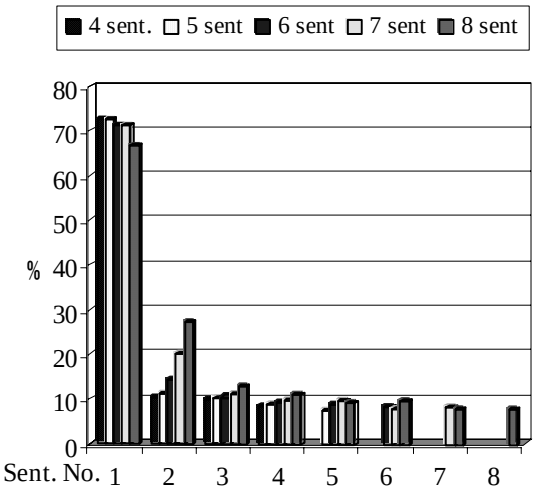

Figure 2. Summary word employment ratio of original news

In this graph, the original news is categorized into five cases according to its sentence count from 4 to 8[1] and the average word employment ratio is shown for each sentence.

Of this figure, the following observations can be made:

- Bias toward the first sentence

In all five cases, the first sentence recorded the highest word employment ratio. The percentages of the second and third sentences increase when the news contains many sentences. The opening part of the news text is called the lead. We will discuss its role in the next section.

- No clear favorite for the final sentence

There was no employment ratio rise for the closing sentences in any case even though our abstractors indicated they often use information in the last sentence. This inconsistency may be due to the word match error. Final sentences actually have an important role in news, as we will see in the next section.

5 Summarization model

In the previous section, we found a quite high word overlap between a summary and the opening part of the original news text. We checked with our word match browser the similarity of the summary news and lead sentences, and found that most of the summary sentences

[1] These news texts cover the 88 % of the total news texts.

take exactly the same syntactic pattern of the opening sentence. Based on this observation and what we found in the interviews, we devised a news text summarization model. The model can explain our abstractors' behavior, and we are planning to develop an automatic or semi-automatic summarization system with it. We will explain the typical news text structure and present our model.

5.1 News text structure

Most of our news texts are written with a three-part structure, i.e., lead, body and supplement. Figure 1 shows the two-fold structure of the lead and the body. Each part has the following characteristics.

- Lead

The most important information is briefly described in the opening part of a news text. This part is called the lead. Proper nouns are often avoided in favor of more abstract expressions such as "a young man" or "a big insurance company." The lead is usually written in one or two sentences.

- Body

The lead is detailed in the body. The 5W1H information is mainly elaborated, and the proper names that were vaguely mentioned in the lead appear here. The statements of people involved in the news sometimes appear here. The repetitive structure of the lead and the body is rooted in the nature of radio news; listeners cannot go back to the previous part if they missed the information.

- Supplement

Necessary information that has not been covered in the lead and the body is placed here. Take for an example of weather news about a typhoon. A caution from the Meteorological agency is sometimes added after the typhoon's movement has been described.

5.2 Model

We found that most of the summary news is written based on the lead sentences. They are then shortened or partly modified with the expressions in the body to make them more informative and self-contained.

The essential operation, we consider, lies in the editing of the lead sentences under the summary length constraint. Based on the observation, we have proposed a two-step summarization model of reading and editing. The summary in Figure 1 is constructed with the lead sentence with the insertion of a phrase in the body.

- Reading phase

(1) Identify the lead, the body and the supplement sentences in the original news.

(2) Analysis

Find the correspondences between the parts in the lead and those in the body. We can regard this process as a co-reference resolution.

- Summary editing phase

(3) Set the lead sentence as the base sentence of the summary.

(4) Apply the following operations until the base sentence length is close enough to the predefined length *N*.

(4-1) Delete parts in the base sentence.

(4-2) Substitute parts in the base sentence with the corresponding parts in the body with the results of (2).

(4-2') Add a body part to the base sentence. We may view this as a null part substituted by a body part.

(4-3) Add supplement sentences.

The supplement is often included in a summary; this part contains different information from the other parts.

5.3 Related works and discussion

Our two-step model essentially belongs to the same category as the works of (Mani et al., 1999) and (Jing and McKeown, 2000). Mani et al. (1999) proposed a summarization system based on the "draft and revision." Jing and McKeown (2000) proposed a system based on "extraction and cut-and-paste generation." Our abstractors performed the same cut-and-paste operations that Jing and McKeown noted in their work, and we think that our two-step model will be a reasonable starting point for our subsequent research. Below are some of our observations.

The lead sentences play a central role in our model since they serve as the base of the final summary. Their identification can be achieved with the same techniques as used for the important sentence extraction. In our case, the sentence position information plays an important role as was shown by Kato and Uratani (2000). We consider the identification of the body and the supplement part together with the lead will be beneficial for the co-reference resolution.

The co-reference resolution problem between the lead and the body should be treated in a more general way than usual. We found that our problem ranges from the word level, the correspondence between named entities and their abstract paraphrases, to the sentence level, an entire statement of a person and its short paraphrase. We are now investigating the types of co-reference that we have to cover.

We found that the deletion of lead parts did not occur very often in our summary, unlike the case of Jing and McKeown (2000). One reason is that most of our leads were short enough[2] to be included in the summary and therefore the substitution operation became conspicuous. This usually increased the length of summary but contributed to making it more lively and informative.

A supplement part was often included in the summary. We consider that this feature corresponds to the abstractors' comments on employment of the final sentence, which was not clearly detected in our statistical investigation described in section 4.3. We are now investigating the conditions for including the supplement.

We have so far listed the basic operations of editing through the manual checking of samples, and we are currently analyzing the operations with more examples. We will then study automatic selection of the optimum operation sequence to achieve the most informative and natural summary.

6 Conclusions

We have described the manual summary process of NHK's broadcast news and experiments on automatic text alignment between news summaries and the original news together with the word position matching. Through a statistical analysis of the results and interviews with abstractors, we found that the abstractors summarize news by taking advantage of its structure. Based on this observation, we proposed a summarization model that consists of a reading and editing phase. We are now designing an automatic or semi automatic summarization system employing the model.

[2] The present summary length constraint is 105 characters. Meanwhile, the average length of the first sentence (typically the lead) of a news text is 94.5 as is shown in table 1.

Acknowledgement

The authors would like to thank Mr. Isao Goto and Dr. Naoto Kato of ATR for valuable discussion and Mr. Riuzo Waki of Eugene Software Inc. for implementing our ideas.

References

Jing, Hongyan and Kathleen R. McKeown. 1999. The Decomposition of Human-Written Summary Sentences. *The 22nd Annual International ACM SIGIR Conference*, pages 129-136, Berkeley.

Jing, Hongyan and Kathleen R. McKeown. 2000. Cut and Paste Based Text Summarization. *The 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 178-185, Seattle.

Kato, Naoto and Noriyoshi Uratani. 2000. Important Sentence Selection for Broadcast News (in Japanese), *The 6th Annual convention of the Association for Natural Language Processing*, pages 237-240, Kanazawa, Japan

Mani, Inderjeet and Mark T. Maybury. 1999. *Advances in Automatic Summarization*, The MIT press, Cambridge, Massachusetts

Mani, Inderjeet, Barbara Gates and Eric Bloedorn. 1999. Improving Summaries by Revising them, *The 37th Annual Meeting of the Association for Computational Linguisics*, pages 558-565, Maryland.

Mani, Inderjeet. 2001. *Automatic Summarization*. John Benjamins, Amsterdam/Philadelphia.

Marcu, Daniel. 1999. The automatic construction of large-scale corpora for summarization research. *The 22nd Annual International ACM SIGIR Conference*, pages 137-144, Berkeley.