Application of NLP technology to production of closed-caption TV programs in Japanese for the hearing impaired

Takahiro Wakao
Telecommunications Advancement Organization (TAO) of Japan

Terumasa Ehara
NHK Science and Technical Research Labs. / TAO

Eiji Sawamura
TAO

Yoshiharu Abe
Mitsubishi Electric Corp Information Technology R&D Center / TAO

Katsuhiko Shirai
Waseda University Department of Information and Computer Science / TAO

1 Introduction

The Telecommunications Advancement Organization (TAO) of Japan, with the support of the ministry of Posts and Telecommunications, has initiated a project in which electronically available text of TV news programs is summarized and synchronized with the speech and video automatically, then superimposed on the original programs for the benefit of the hearing impaired people in Japan. This kind of service has been provided for more than 70% of the TV programs in the United States or in Europe, however, it is available in only 10% of the TV programs in Japan. Most of the closed captions are literal transcriptions of what is being said. Reasons why the availability is low are firstly that thousands of characters are used in the Japanese language, and secondly that the closed captions are produced manually at present and it is a time-consuming and costly task.

The project started in 1996 and will end in 2001. Its annual budget is 200 million yen. The main aim of the project is to establish the technology of producing closed captions for TV programs efficiently using natural language processing and speech recognition technology.

We describe main research issues and the project schedule, and show the results of preliminary research.

2 Research Issues

Main research issues in the project are as follows:

- automatic text summarization
- automatic synchronization of text and speech
- building an efficient closed caption production system

We would like to have the following system (Figure 1) based on the research on the above issues.

Although all types of TV programs are to be handled in the project, the first priority is given to TV news programs.

The outline of each research issue is described next.

2.1 Automatic Text Summarization

For most of the TV news programs today, scripts (written text) are available before they are read out by newscasters. The Japanese news text is read at the speed of four hundred characters per minute and it is too fast, and there are too many characters when all the characters of what is said are shown on the screen (Komine et al., 1996). Thus we need to summarize the news program text and then show it on TV screen. The aim of the research on automatic text summarization is to summarize the text fully or partially automatically to a proper size in order to assist the closed caption production.

2.2 Automatic Synchronization of Text and Speech

Once the original news program text is summarized, it should be synchronized with the actual sound, or the speech of the programs. At present this is done by hand when the closed captions are produced. We would like to make use of speech recognition technology to help the task of synchronizing text with speech. Please note that what we aim at is to synchronize the original text rather than the summarized text with the speech.

2.3 Efficient Closed Caption Production System

We will create a system by integrating the summarization and synchronization techniques with techniques for superimposing characters. We also need to research on other aspects such as what the best way is to show the characters on the screen for the handicapped viewers.

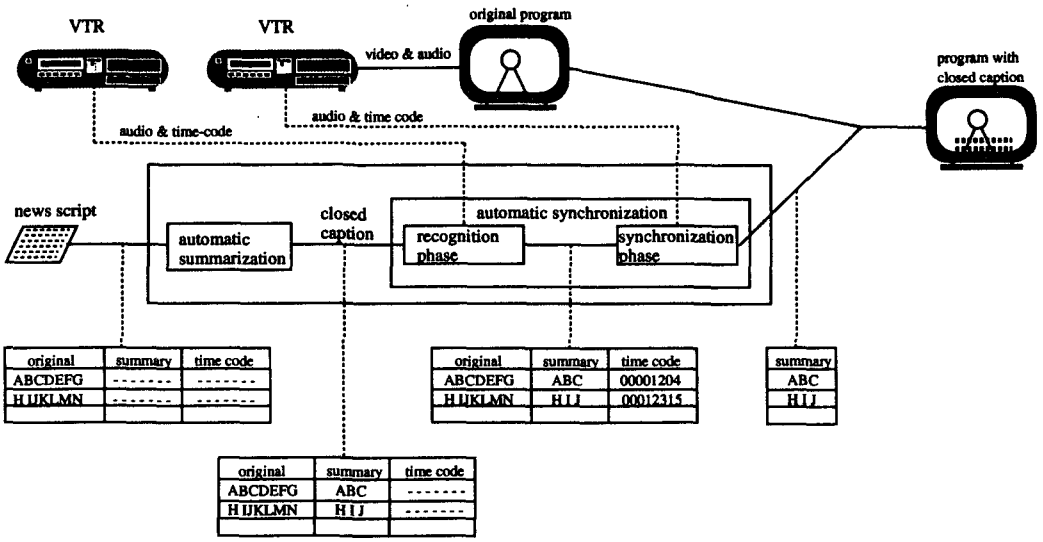


Figure 1: System Outline

3 Project Schedule

The project is divided into two stages; the first three years and the rest, two years. We conduct research on the above issues and create a prototype system in the first stage. In addition, the prototype system is to be used to produce closed captions, and the capability and functions of the system will be evaluated. We will improve the prototype system in the second stage.

In 1996 and 1997, the following research has been conducted and will be continued.

- Automatic text summarization
  - method for dividing a sentence into smaller sections
  - key word extraction
  - method for connecting sentence sections
- Automatic synchronization of text and speech
  - transcription, speech model integration system
  - maximum likelihood matching system
  - speech database
- Efficient closed caption production system
  - integrated simulation system for closed caption production

4 Preliminary Research Results

We have conducted preliminary research for automatic text summarization and synchronization of text and speech, and the results are as follows.

4.1 Automatic Text Summarization

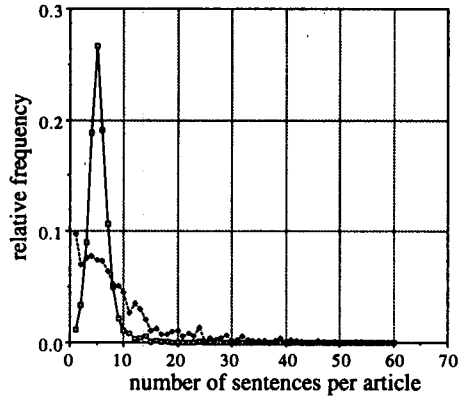
Text summarization research in the past may be grouped into three approaches. The first is to generate summarized sentences based on understanding of the text. It is desirable, however, it is not a practical method at present in order to summarize actual TV news program text.

The second is to digest the text by making use of text structures such as paragraphs. It has been applied to newspaper articles in Japanese (Yamamoto et al, 1994). In this approach important parts of the text which are to be kept in the summarization, are determined by their locations, i.e. where they appear in the text. For example, if nouns or proper nouns appear in the headline, they are considered as 'important' and may be used as measures of finding out how important the other parts of the text are. As we describe later, TV news text is different from newspaper articles in that it does not have obvious structures, *i.e.* the TV news text has fewer sentences and usually only one paragraph without titles or headlines. Thus the second approach is not suitable for the TV news text.

The third is to detect important (or relevant) words (segments in the case of Japanese), and determine which section of the text is important, and then put them together to have 'summarization' of the text. This is probably most robust among the three approaches and we are using the third approach currently (for summary of various summarization techniques, please see (Paice, 1990)).

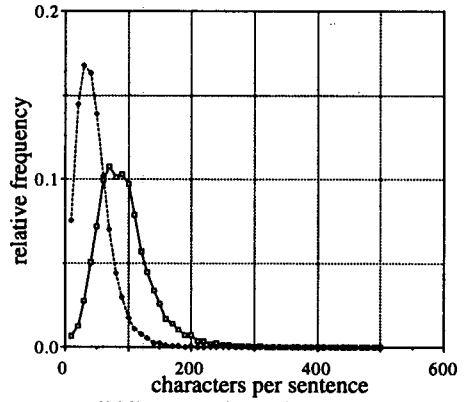
To illustrate the difference between TV news program text and newspaper articles, we compared one

thousand randomly selected articles from both domains. The results are shown in Fig 2 and Fig 3.



solid line: TV, dotted line: newspaper

Figure 2: Number of sentences per article



solid line: TV, dotted line: newspaper

Figure 3: Number of characters per sentence

Fig 2 and Fig 3 show that in comparison with newspaper text, the TV news program text has the following features:

- Fewer sentences per text
- Longer sentences

If we summarize TV news program text by selecting 'sentences' from the text, it will be 'rough' summarization. On the other hand, if we can divide long sentences into smaller sections and thus increase the number of 'sentences (sections)' in the text, then we may have better summarization (Kim and Ehara, 1994).

As a method of summarization, we are using the third approach. To find important words in the text, high-frequency key word method and TF-IDF (Term Frequency - Inverse Document Frequency) method have been adopted, and the two methods are evaluated automatically on a large-scale in our preliminary research. We used ten thousand (10000) TV news texts between 1992 and 1995 (2500 texts each year) for the evaluation. One of the features of the TV news texts is that the first sentence is the most important. We conducted the evaluation by taking advantage of the feature.

Key words used in the high-frequency key word method are content words which appear more than twice in a given text (Luhn, 1957),(Edmundson, 1969). To determine the importance of a sentence, we counted the number of the key words in the sentence and then it is divided by the number of the words (including function and content words). In the TF-IDF method, first the weight of each word is computed by multiplying its frequency in the text (TF) and its IDF in a given text collection. The importance of the sentence is thus computed by summing up all the weights of the words in the sentence and divided by the number of the words (Spark Jones, 1972), (Salton, 1971).

The evaluation details are as follows. First, the importance of each sentence is calculated by the high-frequency key word or TF-IDF method. Then the sentence are ranked according to their importance. We computed the accuracy of the method by looking at whether the first sentence is ranked the first, or ranked either the first or the second.

The evaluation results are shown in Table 1. The high-frequency key word method produced better results than TF-IDF method did.

Method	First (%)	First_or_Second (%)
High-frequency key word	68.86	88.95
TF-IDF	54.02	80.67

Table 1: Sentence Extraction Accuracy

4.2 Automatic Synchronization of Text and Speech

As the next step, we need to synchronize the text and the speech. First, the written TV news text is changed into the stream of phonetic transcriptions, and then synchronization is done by detecting the time points of the text sections and their corresponding speech sections. At the same time, we have started to create news speech database. In 1996, we collected the speech data by simulating news programs, *i.e.* the TV news texts were read and recorded in a studio rather than actual TV news programs on the air were recorded. We collected seven and half hours of recordings of twenty people

(both male and female). We plan to record actual programs as 'real' data in addition to the simulation recording in 1997. The real data will be taken from both radio and TV news programs.

Preliminary research on detection of synchronization points is conducted by using the data we have created. A speech model is produced by using three hours (four male and four female persons) of recording as training data. For each speaker, a two-loop, four-mixture-distribution phonetic HMM was learned. Based on the HMMs, key-word pair models were obtained from the phonetic transcription. The key-word pair model is shown in Fig 4. The model consists of two strings of words (keywords1 and keywords2) before and after the synchronization point (point B).

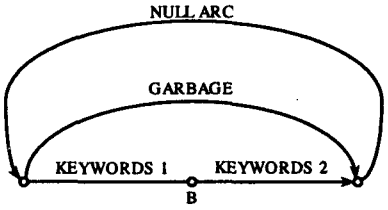


Figure 4: key-word pair model

When the speech is fed to the model, the non-synchronizing input data travel through the garbage arc while the synchronizing data go through the keywords. It means that the likelihood at point B increases. Thus if we observe the likelihood at point B and it goes over a certain threshold, we decide it is the synchronization point for the input data. Twenty-one key-word pairs were taken from the data which was not used in the training, and selected for evaluation. We fed one male and one female speech to the model in the evaluation. The result is shown in Table 2.

As we decrease the threshold, the detection rate increase, however, the false alarm rate increases rapidly.

5 Conclusion

We have described a national project in which 'speech' of TV programs is changed into captions, and superimposed to the original programs for the benefit of the hearing impaired people in Japan. We also showed results of preliminary research on TV news text summarization, and synchronization of text and speech. We continue to integrate the natural language processing and speech processing technology for efficient closed caption production system, and put it to a practical use as soon as possible.

Threshold	Detection Rate (%)	False Alarm Rate (FA/KW/Hour)
-10	69.05	2.78
-20	76.19	9.17
-30	85.71	39.72
-40	90.48	131.93
-50	92.86	409.97
-60	97.62	975.75
-70	97.62	1774.30
-80	97.62	2867.55
-90	97.62	4118.56
-100	97.62	5403.45

Table 2: Synchronization Detection

References

Komine, K., Hoshino, H., Isono, H., Uchida, T., Iwahana, Y. 1996 Cognitive Experiments of News Captioning for Hearing Impaired Persons Technical Report of IEICE (The Institute of Electronics, Information and Communication Engineers), HCS96-23, in Japanese, pages 7–12.

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

H.P. Edmundson 1969 New Methods in Automatic Extracting. In *Journal of the ACM*, 16(2), pages 264–285.

Chris D. Paice 1990 Constructing literature abstracts by computer: techniques and prospects. In *Information Processing & Management* 26(1), pages 171–186. Pergamon Press plc.

Yeun-Bae Kim, Terumasa Ehara. 1994. An Automatic Sentence Breaking and Subject Supplement Method for J/E Machine Translation *Information Processing Society of Japan, Ronbun-shi*, Vol 35, No. 6. In Japanese.

Gerard Salton 1971 (Ed) The Smart Retrieval System - Experiments in Automatic Document Retrieval, Englewood Cliffs, NJ: Prentice Hall Inc.

Karen Spark Jones 1972 A statistical interpretation of term specificity and its application in retrieval In *Journal of Documentation*, 28(1), pages 11–21.

Kazuhide Yamamoto, Shigeru Masuyama, Shozo Naito 1994 GREEN: An Experimental System Generating Summary of Japanese Editorials by Combining Multiple Discourse Characteristics NL-99-3, Information Processing Society of Japan. In Japanese.