SUMMARIZATION OF NEWS SPEECH WITH UNKNOWN TOPIC BOUNDARY

S.Takao, T.Haru and Y.Ariki

1-5, Yokotani, Oe-cho, Seta, Otsu-shi, Shiga-ken, Japan
e-mail:tail@arikilab.elec.ryukoku.ac.jp

ABSTRACT

TV viewers want to grasp the contents of the news program in a short time due to the increasing number of news channels. Conventional summarization methods based on extraction of the important sentences from each topic included in the news speech is insufficient because the important sentences can not always be extracted from each topic due to unknown topic boundary. To solve this problem, in this paper, we propose a summarization method of TV news program by segmenting the news speech into topics and then extracting the important sentence from each topic.

1. INTRODUCTION

TV news programs are now broadcast from all over the world owing to the broadcast digitization so that TV viewers require to grasp the contents of the news programs in a short time. In this situation, a lot of studies have been done in a field of sentence or speech summarization [1]-[3]. They summarize news speech by extracting the important sentences from news speech. However, it is difficult to produce reasonable summarization by their methods, because the news speech is continuous and the topic boundary is unknown. Some topics may be lost when extracting important sentences, if the extracted important sentences of the topics have less value.

To solve this problem, in this paper, we propose a method which segments the news speech into topics at first and then extracts the important sentences from each topic. Therefore, our method is composed of automatic topic segmentation of continuous news speech and topic summarization. In the topic segmentation, we propose a new method based on passage similarity and word similarity in a word space which can be constructed from the topic sections in the test data itself. In the important sentence extraction, we propose a new term weighting method called "mutual information incorporating TF-IDF".

2. SUMMARIZATION BY EXTRACTING IMPORTANT SENTENCES

2.1. Summarization Method

In technical paper [1], the news speech is summarized as follows. At first, the news speech is transcribed and the importance degree of each word in the speech transcription is computed. Then the importance degree of each sentence is computed based on the importance degree of the words. Finally, the news speech is summarized by extracting some sentences with high importance degree.

When the news speech is transcribed, some words may be inserted, deleted and mis-recognized. We call here the inserted words, deleted words and mis-recognized words as error words. These error words cause the extraction error of important sentences and as a result cause poor summarization. To solve this problem, the importance degree of the error words must be lowered. In order to make this computation feasible, mutual information incorporating TF-IDF are employed as a term weighting method in this paper.

2.2. Mutual Information Incorporating TF-IDF

There are two types of error words. One type occurs in any passage sections. Here the passage section indicates the consecutive several sentences. The other type occurs in particular passage sections. The error words occurring in any passage sections show low mutual information or TF-IDF so that they can be easily excluded by the value.

On the other hand, the term frequency of the error words may be high or low when appear in particular passage sections. Therefore the error words occurring in particular passage sections with low frequency show high IDF and low TF so that they can be excluded by using TF-IDF.

The error words occurring in particular passage sections with high frequency show high TF and IDF so that those can not be excluded from the normal words by TF-IDF. Therefore, mutual information becomes effective because it is independent of term frequency.

However, mutual information can not exclude the error words with low frequency in particular passage sections, because it is independent of term frequency. Consequently,

these three types of the error words can be excluded by integrating TF-IDF with mutual information. In this paper, mutual information incorporating TF-IDF is carried out as term weighting method as shown in Eq.(1). In Eq.(1), t_j, w_i and $i()$ denote a passage section, a word and information amount respectively.

$$\begin{aligned} & i(t_j; w_i) \times TF \cdot IDF \\ &= (i(t_j) - i(t_j|w_i)) \cdot TF(w_i, t_j) \cdot IDF(w_i) \\ &= \left(\log \frac{P(t_j, w_i)}{P(t_j)P(w_i)}\right) \cdot TF(w_i, t_j) \cdot IDF(w_i) \end{aligned} \quad (1)$$

2.3. Sentence Weighting Method

The importance degree of the sentences is computed based on the importance degree and the number of occurrences of the words included in a sentence as shown in Eq.(2). Here the number of occurrences of word w_i is denoted as x_i and the importance degree of word w_i is denoted as I_i. The total number of word occurrences in a sentence is denoted as N.

$$S = \frac{1}{N} \sum_i x_i \times I_i \quad (2)$$

3. SUMMARIZATION BY TOPIC SEGMENTATION

3.1. Problem of Topic Deletion

In the conventional method [1], the important sentence can not be extracted from each topic in the news speech transcription, because the news speech is continuous and the topic boundary is unknown. For instance, let the news speech have three topics and the number of their sentences is eight in topic A, seven in topic B and two in topic C respectively. If two sentences are extracted from topic A and one sentence from topic B with high importance degree, instead of three sentences from three topics, then the summarization of topic C is deleted. Therefore, a user can not grasp the contents of topic C in the news speech summarization. To solve this problem, we propose a method which segments the news speech into each topic and extracts the important sentences from each topic. The segmentation of the news speech into each topic is called topic segmentation in this paper.

3.2. Problem of Summarization Rate

When a user views the summarized news, he will require summarization rate to which the news speech duration is adjusted as he likes. In the summarization method using important sentence extraction, the summarization rate is decided by the number of sentences extracted. However, there is the problem that some topics may be deleted in this method. To solve this problem, we propose a method to decide the summarization rate without deleting any topics. In our proposal method, the news speech transcription is segmented into each topic at first. Then the clustering method which are described later is applied to all topics included in the given news speech to produce topic clusters. Finally an important sentence is extracted from each topic cluster. Therefore, the number of topic clusters which are produced by the clustering method decides the summarization rate.

3.3. Summarization Procedure

The transcription is required to summarize the news speech. Correct transcription can be produced if human transcribes the news speech manually. However, we purpose in this paper automatic transcription and summarization of the news speech so that we carried out automatic speech transcription to the broadcast news. The procedure of news summarization using topic segmentation is shown in Fig. 1 and is described as follows.

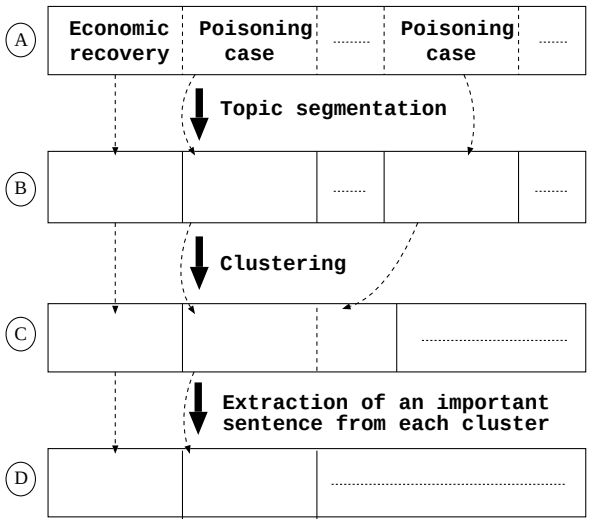

Figure 1: Summarization procedure

1. Speech transcription is carried out for Japanese Broadcast News.
2. The topic segmentation detects topic boundaries in the news speech transcription. (The dotted lines shown in Fig. 1 A are changed into the solid lines in Fig. 1 B.)
3. The clustering method produces topic clusters from all the topics. (From B to C.)
4. An important sentence is extracted from each cluster. (From C to D.)

In this study, we used the topic segmentation system described in the technical paper [5] for topic segmentation. The clustering is performed based on a merging method employing minimum distance.

3.4. Clustering Method

In producing topic clusters, each topic section is regarded as an individual cluster at first. Then two clusters are merged repeatedly until the distance between the clusters becomes shorter than the threshold. The distance D_{AB} between two clusters A and B is shown in Eq.(3).
$S(X_k, X_l)$ in Eq.(3) is the similarity between the topic vectors X_k and X_l which are produced using the words included in the topic t_k and t_l respectively. The similarity is computed based on the word distance $WD(w_i, w_j)$ as well as the number of overlapping words [4] as shown in Eq.(4). Here x_{ik} is the number of the word w_i in topic section t_k. The word distance is shown in Eq.(5). In Eq.(5), $TF(w_i, t_m)$ is the frequency of the word w_i in topic section t_m and $IDF(w_i)$ is the inverse document frequency. $i(t_m; w_i)$ indicates the mutual information between topic section t_m and word w_i and M is the number of all the topic sections.

$$D_{AB} = \min_{k \in A, l \in B} \frac{1}{S(X_k, X_l)} \quad (3)$$

$$S(X_k, X_l) = \sum_i \sum_j x_{ik} \cdot x_{jl} \times \frac{1}{WD(w_i, w_j) + 1} \quad (4)$$

$$\begin{aligned} WD(w_i, w_j) = \frac{1}{M} \sum_{m=1}^{M} & ((TF(w_i, t_m) - TF(w_j, t_m))^2 \\ & + (IDF(w_i) - IDF(w_j))^2 \\ & + (i(t_m; w_i) - i(t_m; w_j))^2)^{\frac{1}{2}} \end{aligned} \quad (5)$$

4. SPEECH TRANSCRIPTION

4.1. Experimental Condition

We carried out automatic speech transcribing for the broadcast NHK news programs using a language model and an acoustical model. The language model is the word bigram constructed from RWC text database which was produced by morphologically analyzing the MAINICHI Japanese newspaper of 45 months from 1991 to 1994. The number of the words in the dictionary is 20,000. The word bigram was back-off smoothed after cutting off at 1 word.

Speaker independent cross-word triphone HMMs were constructed. They were trained using 21,782 sentences spoken by 137 Japanese males. These speech data is taken from the database of acoustical society of Japan. The acoustic parameters are 39 MFCCs with 12 Mel cepstrum, log energy and their first and second order derivatives. Cepstrum mean normalization was applied to each sentence to remove the difference of input circumstances. Table1 shows the experimental conditions for acoustic analysis (AA) and HMM.

In the transcribing experiment, we used HTK (HMM Toolkit) as the decoder which can perform Viterbi decoding with beam search using above mentioned language model and acoustic model.

Table 1: Acoustic Analysis(AA) and HMM

A	Sampling frequency	12kHz
A	High-pass filter	$1 - 0.97z^{-1}$
A	Feature parameter	MFCC,Pow,, (39th)
A	Frame length	20ms
A	Frame shift	5ms
A	Window type	Hamming window
H	Learning method	Concatenated training
H	Type	Left to right continuous HMM
M	Number of states	5 states with 3 loops
M	Number of mixtures	8

4.2. Transcription Result

Speech transcribing was carried out for the NHK spoken news programs broadcast in 1998. The total number of time duration and sentences were 3.0 hours and 587 respectively. The transcription result is shown in Table 2. In the table, the "Corr" indicates the correctness defined by Eq.(6). On the other hand, the "Acc" indicates the accuracy defined by Eq.(7).

The reason why the transcription result is a little lower is explained as follows. The language model was constructed from the MAINICHI Japanese newspaper published from 1991 to 1994. On the other hand, the test data was NHK spoken news broadcast in 1998. This time difference mainly seems to have caused the lower transcription result. Further more, most of the topics included in the MAINICHI Japanese newspaper are concerned with economy and politics, and other various topics are equally included in the test data. This unevenness of topic distribution also seems to have caused the lower transcription result. This transcription result was used for spoken document retrieval.

$$Percent\ Correct = \frac{N - D - S}{N} \times 100\% \quad (6)$$

$$Percent\ Accuracy = \frac{N - D - S - I}{N} \times 100\% \quad (7)$$

$$\begin{aligned} N &: Total\ number\ of\ labels \\ S &: Number\ of\ substitution\ errors \\ D &: Number\ of\ deletion\ errors \\ I &: Number\ of\ insertion\ errors \end{aligned}$$

Table 2: Transcription result(%)

	Corr	Acc
19980820-12:00NHK	77.83	75.57
19980820-23:00NHK	77.42	75.43
19980824-12:00NHK	76.46	73.74
19980824-19:00NHK	75.89	72.74
19980825-12:00NHK	77.81	73.94
19980826-12:00NHK	78.91	76.24
Total	77.53	74.70

5. EXPERIMENTAL RESULTS

Summarization experiments were carried out for the test data shown in Table 2. We compared two methods: conventional one based on important sentence extraction and the proposed one based on important sentence extraction after topic segmentation. The summarization rate is evaluated by Eq.(8) for each summarization method. Here, SR, BS and AS are summarization rate, the number of sentences in the data before summarization and the number of sentences included in the data after summarization.

TV viewers can grasp the contents of TV news in a short time if the summarization rate is high, possibly decreasing the number of topics. Therefore, deletion rate and overlapping rate of the topics have to be evaluated for the summarized data. The deletion rate is defined as the ratio of the number of deleted topics in the test data after summarization to the total number of topics as shown in Eq.(9). The overlapping rate is defined as the ratio of the number of sentences belonging to the same topic in the summarized data to the total number of sentences included in the test data as shown in Eq.(10).

If the deletion rate is low, TV viewers can grasp many kinds of topics. On the other hand, if the overlapping rate is low, TV viewers can grasp the summarized contents in a short time. Consequently, we integrated them into balance average as shown in Eq.(11). The experimental results are shown in Table 3. In the table, "Important" indicates the method based on important sentence extraction and "Integration" indicates the method based on important sentence extraction after topic segmentation.

$$SR = \frac{BS - AS}{BS} \quad (8)$$

$$Deletion = \frac{Number\ of\ topics\ deleted}{Total\ number\ of\ topics} \quad (9)$$

$$Overlap = \frac{Number\ of\ sentences\ overlapping}{Total\ number\ of\ sentences} \quad (10)$$

$$Average = \frac{2 \times Deletion \times Overlap}{Deletion + Overlap} \quad (11)$$

Table 3: Experimental results(%)

	Important	Integration
Summarization rate	91	90
Deletion rate	47	40
Overlapping rate	47	36
Balance average	47	38

In Table 3, the method using important sentence extraction after topic segmentation has 9% (-(38%-47%) superiority to the method using only important sentence extraction at balance average under the condition that the both summarization methods show almost the same summarization rate. The main reason of this superiority of the proposed method is attributed to the effectiveness of the topic clustering to reduce the overlap rate and the topic segmentation to reduce the deletion rate.

6. CONCLUSION

In this paper, we proposed the summarization technique to news speech under unknown topic boundaries. The experimental results showed that our proposed method is superior to the conventional method based on important sentence extraction. In the proposed method, important sentences are extracted from each topic cluster after topic segmentation. Therefore, it is difficult to summarize the content of topics daily changing. Hereafter, we will study on a method to extract the changing content of topics.

7. REFERENCES

[1] F.Ren and Y.Sadanaga: "An Automatic Extraction of Important Sentences Using Statistical Information and Structural Feature", **NL98**-125, pp.71-78, 1998-05.

[2] C.Hori and S.Furui: "Improvements in automatic speech summarization and evaluation methods", **ICSLP98**, 1998-10.

[3] K.Koumpis and S.Renals: "Transcription and Summarization of Voicemail Speech", **ICSLP98**, 1998-10.

[4] S.Takao, J.Ogata, and Y.Ariki, "Study on New Term Weighting Method and New Vector Space Model Based on Word Space in Spoken Document Retrieval", RIAO00, Volume I, pp.116-131, 2000-04.

[5] S.Takao, J.Ogata and Y.Ariki: "Topic Segmentation of News Speech Using Word Similarity", **ACM00**, pp.442-444, 2000-11.