TWO-STAGE AUTOMATIC SPEECH SUMMARIZATION BY SENTENCE EXTRACTION AND COMPACTION

Tomonori Kikuchi, Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology
{kikuchi, furui}@furui.cs.titech.ac.jp

Chiori Hori

Intelligent Communication Laboratory
NTT Communication Science Laboratories
chiori@cslab.kecl.ntt.co.jp

ABSTRACT

This paper proposes a new automatic speech summarization method having two stages: important sentence extraction and sentence compaction. Relatively important sentences are extracted from the results of large-vocabulary continuous speech recognition (LVCSR) based on the amount of information and the confidence measures of constituent words. The set of extracted sentences is compressed by our sentence compaction method. Sentence compaction is performed by selecting a word set that maximizes a summarization score which comprises the amount of information and the confidence measure of each word, the linguistic likelihood of word strings, and the word concatenation probability. The selected words are concatenated to create a summary. Effectiveness of the proposed method was confirmed by testing summarization of spontaneous presentations. Optimal ratio of sentence extraction to sentence compaction changes according to the target summarization ratio and features of presentations.

1. INTRODUCTION

Speech recognition has two major applications[1] :transcribing speech documents such as presentations, lectures and broadcast news; and dialogue with computer systems. Since speech is the most natural and effective method of communication between human beings, the former application is expected to become very important in the coming IT era. Although high recognition accuracy can be easily obtained for text reading speech such as anchor speakers' broadcast news utterances, technological ability for recognizing spontaneous speech is still limited. Spontaneous speech is ill-formed and very different from written text. Spontaneous speech usually includes redundant information such as disfluencies, filled pauses, repetitions, repairs and word fragments. In addition, irrelevant information included in a transcription caused by recognition errors is usually inevitable. Therefore, an approach in which all words are transcribed is not an effective one for spontaneous speech. Instead, speech summarization for extracting important information and removing redundant and incorrect information is necessary for recognizing spontaneous speech.

Techniques for automatically summarizing written text are now actively being investigated in the field of natural language processing [2][3]. However, many of these techniques are not applicable to speech, and so techniques for speech summarization have just recently started to be investigated. We have proposed a sentence compaction-based statistical speech summarization technique, in which a set of words maximizing a summarization score indicating appropriateness of summarization is extracted from automatically transcribed speech and then concatenated to create a summary according to a target compression ratio[4][5]. The proposed technique can be applied to each sentence utterance, as well as to whole speech documents consisting of multiple utterances. This technique has been applied to Japanese, as well as English documents, and its effectiveness has been confirmed.

However, when multiple spontaneous utterances including many recognition errors and disfluencies are summarized with a high compression ratio (a small summarization ratio), the summary sometimes includes unnatural, incomplete sentences consisting of a small number of words, and it becomes difficult to read and understand. This paper proposes a new two-stage summarization method, consisting of important sentence extraction and sentence compaction, to cope with this problem. In the new method, relatively well-structured and important sentences including important information and less speech recognition errors are extracted, and then sentence compaction is applied to the set of extracted sentences.

The remainder of the paper is organized as follows. In the next section, the two-stage summarization method is described. Sections 3 and 4 provide the conditions and results of evaluation experiments for automatically summarizing spontaneous presentation utterances. The paper concludes with a general discussion and issues related to future research.

2. TWO-STAGE SUMMARIZATION METHOD

Figure 1 shows the two-stage summarization method consisting of important sentence extraction and sentence compaction. Using the speech recognition results, the score for important sentence extraction is calculated for each sentence. After removing all the filled pauses, a set of relatively important sentences is extracted, and sentence compaction using our proposed method is applied to the set of extracted sentences. The ratios of sentence extraction and compaction are controlled according to a summarization ratio given by the user.

2.1. Important sentence extraction

The important sentence extraction is performed according to the following score for each sentence, $W = w_1, w_2, \ldots, w_N$, obtained as a result of speech recognition:

$$S(W) = \frac{1}{N} \sum_{i=1}^{N} \{L(w_i) + \lambda_I I(w_i) + \lambda_C C(w_i)\} \qquad (1)$$

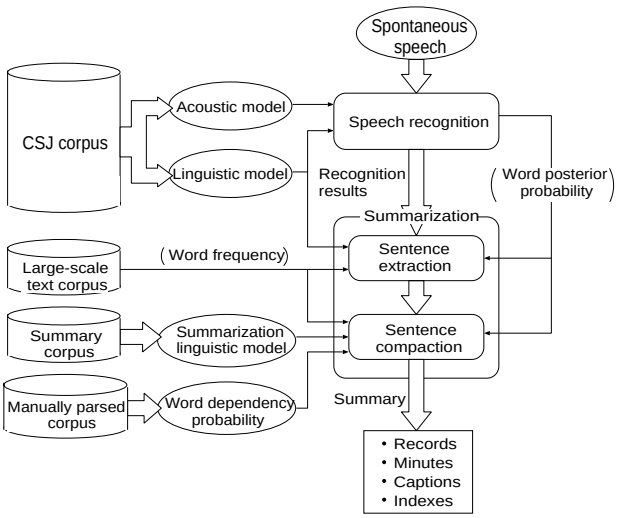

Fig. 1. Automatic speech summarization system.

where N is the number of words in the sentence W, and $L(w_i)$, $I(w_i)$ and $C(w_i)$ are the linguistic score, the significance score, and the confidence score of word w_i, respectively. The three scores are a subset of the scores originally used in our sentence compaction method and considered to be useful also as measures indicating the appropriateness of including the sentence in the summary. λ_I and λ_C are weighting factors for balancing the scores.

Details of the scores are as follows.

Linguistic score

The linguistic score $L(w_i)$ indicates the linguistic likelihood of word strings in the sentence and is measured by n-gram probability:

$$L(w_i) = \log P(w_i | \dots w_{i-1}) \tag{2}$$

In our experiment, trigram probability calculated using transcriptions of presentation utterances in the CSJ (Corpus of Spontaneous Japanese)[6] consisting of 1.5M morphemes (words) is used. This score de-weights linguistically unnatural word strings caused by recognition errors.

Significance score

The significance score $I(w_i)$ indicates the significance of each word w_i in the sentence and is measured by the amount of information. The amount of information is calculated for content words including nouns, verbs, adjectives and out-of-vocabulary (OOV) words, based on word occurrences in a corpus as shown in Eq.(3). A flat score is given to other words.

$$I(w_i) = f_i \log \frac{F_A}{F_i} \tag{3}$$

where f_i is the number of occurrences of w_i in the recognized utterances, F_i is the number of occurrences of w_i in a large-scale corpus, and F_A is the number of all content words in that corpus, that is $\sum_i F_i$.

For measuring the significance score, the number of occurrences of 120k kinds of words in a corpus consisting of transcribed presentations (1.5M words), proceedings of 60 presentations, presentation records obtained from WWW (2.1M words), NHK (Japanese broadcast company) broadcast news text (22M words), Mainichi newspaper text (87M words) and text from a speech textbook "Speech Information Processing" (51k words) is calculated. Important keywords are weighted and the words unrelated to the original content, such as recognition errors, are de-weighted by this score.

Confidence score

The confidence score $C(w_i)$ is incorporated to weight acoustically as well as linguistically reliable hypotheses. Specifically, a logarithmic value of a posterior probability for each transcribed word, that is the ratio of a word hypothesis probability to that of all other hypotheses, is calculated using a word graph obtained by a decoder and used as a confidence score.

2.2. Sentence compaction

After removing sentences having relatively low recognition accuracy and/or low significance, the remaining transcription is automatically modified into a written editorial article style to calculate the score for sentence compaction. Sentence compaction is performed using the method that we proposed in [5]. In this method, all the sentences are combined together, and the linguistic score, the significance score, the confidence score and the word concatenation score are given to each transcribed word. The word concatenation score is incorporated to weight a word concatenation between words with dependency in the transcribed sentences. The dependency is measured by a phrase structure grammar, SDCFG (Stochastic Dependency Context Free Grammar). A set of words that maximizes a weighted sum of these scores is selected according to a given compression ratio and connected to create a summary using a 2-stage dynamic programming (DP) technique. Specifically, each sentence is summarized according to all possible compression ratios, and then the best combination of summarized sentences is determined according to a target total compression ratio.

Ideally, the linguistic score should be calculated using a word concatenation model based on a large-scale summary corpus. Since such a summary corpus is not yet available, the transcribed presentations used to calculate the word trigrams for the important sentence extraction are automatically modified into a written editorial article style and used together with the proceedings of 60 presentations to calculate the trigrams.

The significance score is calculated using the same corpus as that used for calculating the score for important sentence extraction. The word dependency probability is estimated by the Inside-Outside algorithm, using a manually parsed Mainichi newspaper corpus having 4M sentences with 68M words.

3. EVALUATION EXPERIMENTS

3.1. Evaluation set

Three presentations in the CSJ by male speakers were summarized at summarization ratios of 70% and 50%. Length and mean word recognition accuracy of each presentation are shown in Table 1. They were manually segmented into sentences before recognition.

Presentation ID	Length[min]	Recognition acc.[%]
M74	12	70
M35	28	60
M31	27	65

Table 1. Evaluation set.

3.2. Summarization accuracy

To automatically evaluate the summaries, correctly transcribed presentation speech is manually summarized by nine human subjects to create targets. Variations of manual summarization results are merged into a word network as shown in Fig.2, which is considered to approximately express all possible correct summaries covering subjective variations. Word accuracy of the summary is measured in comparison with the closest word string extracted from the word network as the summarization accuracy [5].

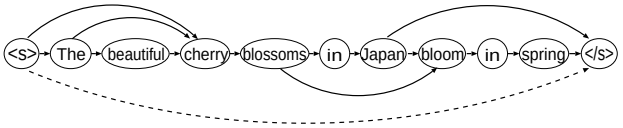

Fig. 2. Word network made by merging manual summarization results.

3.3. Evaluation conditions

Summarization was performed under the following nine conditions; single-stage summarization without applying the important sentence extraction (NOS), two-stage summarization using seven kinds of the possible combination of scores for important sentence extraction (L, I, C, L_I, I_C, C_L, L_I_C), and summarization by random word selection. The weighting factors, λ_I and λ_C, were set at optimum values for each experimental condition.

4. EVALUATION RESULTS

4.1. Summarization accuracy

Results of the evaluation experiments are shown in Figs.3 and 4. In all the automatic summarization conditions, both our previous one-stage method without sentence extraction and our new two-stage method including sentence extraction achieve better results than random word selection. In both 70% and 50% summarization conditions, the two-stage method achieves higher summarization accuracy than the one-stage method.

Comparing the three scores for sentence extraction, the significance score (I) is more effective than the linguistic score (L) and the confidence score (C). The summarization score can be increased by using the combination of two scores (L_I, I_C ,C_L) and even more by combining all three scores(L_I_C).

The two-stage method is more effective in the condition of the smaller summarization ratio (50%), that is, a higher compression ratio, than in the condition of the larger summarization ratio (70%). In the 50% summarization condition, the two-stage method is effective for all three presentations. For the presentation M31, a 5% improvement of the summarization accuracy, compared with the one-stage method, is achieved by using only the significance score, and a 6% improvement is achieved by combining all three

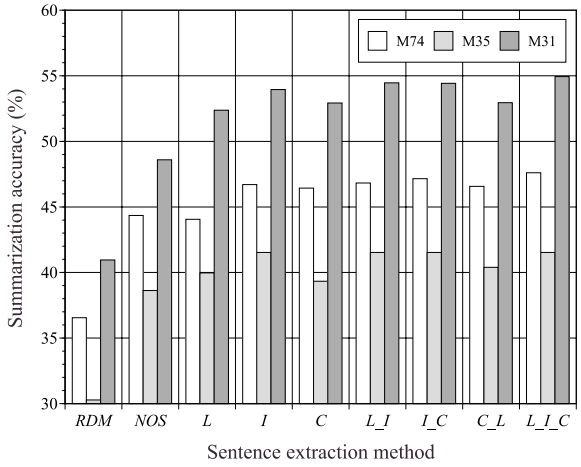

Fig. 3. Summarization at 50% summarization ratio.

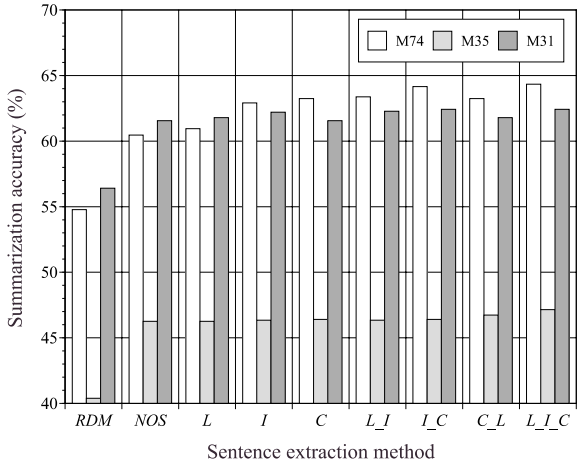

Fig. 4. Summarization at 70% summarization ratio.

scores. On the other hand, in the 70% summarization condition, the two-stage method achieves 2% improvement by using only the significant score and 4% by combining all three scores for one of the presentations, M74, and a much smaller improvement for the other presentations, M35, M31. The latter two presentations can be characterized by a relatively large number of redundant expressions, such as disfluencies, filled pauses, and repetitions. For these presentations, word deletion is more effective than sentence extraction, especially in the condition of a larger summarization ratio.

4.2. Effects of the ratio of compression by sentence extraction

Figures 5 and 6 show the summarization accuracy as a function of the ratio of compression by sentence extraction for the total summarization ratios of 50% or 70%. This result indicates that although the best summarization accuracy of each presentation can be obtained at a different ratio of compression by sentence extraction, there is a general tendency that the smaller the summarization

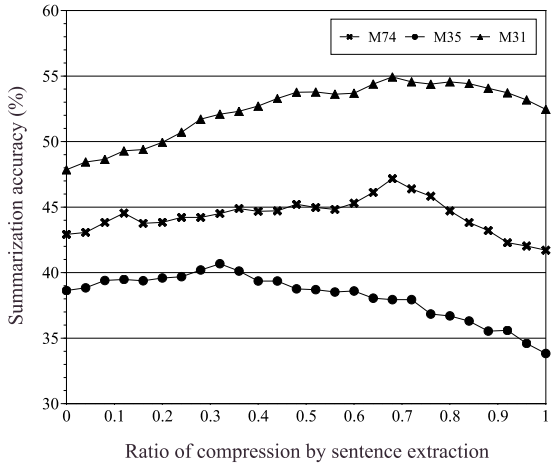

Fig. 5. Summarization accuracy as a function of the ratio of compression by sentence extraction for the total summarization ratio of 50%.

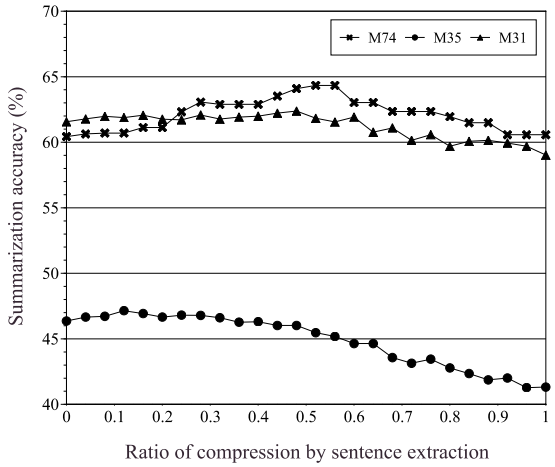

Fig. 6. Summarization accuracy as a function of the ratio of compression by sentence extraction for the total summarization ratio of 70%.

ratio becomes, the larger the optimum ratio of compression by sentence extraction becomes. That is, sentence extraction becomes more effective when the summarization ratio is getting smaller.

Comparing results at the right and left ends of the figures, summarization by word extraction, that is, sentence compaction is more effective than sentence extraction for M31 presentation, which includes a relatively large number of redundant information.

These results indicate that the optimum division of the compression ratio into the two summarization stages needs to be estimated according to the summarization ratio and features of the presentation, such as frequency of disfluencies, filled pauses and repetitions.

5. CONCLUSION

This paper has proposed a new two-stage automatic speech summarization method consisting of important sentence extraction and sentence compaction. In this method, inadequate sentences including recognition errors and less important information are automatically removed before word-based sentence compaction. It was confirmed that in spontaneous presentation speech summarization, combining sentence extraction with sentence compaction is effective; this method achieves better summarization performance than our previous one-stage method. It was also confirmed that three scores, the linguistic score, the word significance score and the word confidence score, are effective for extracting important sentences. The two-stage method is effective for avoiding one of the problems of the one-stage method, that is, the production of short unreadable and/or incomprehensible sentences. The best condition for dividing the summarization ratio into the ratios of sentence extraction and sentence compaction depends on the summarization ratio and features of presentation utterances.

Future research includes evaluation of the usefulness of other information/features for important sentence extraction, investigation of methods for automatically segmenting a presentation into sentence units for extraction and their effects on summarization accuracy, and automatic optimization of the division of compression ratio into the two summarization stages according to the summarization ratio and features of the presentation.

6. ACKNOWLEDGEMENT

The authors would like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast news database.

7. REFERENCES

[1] S. Furui, K. Iwano, C. Hori, T. Shinozaki, Y. Saito and S. Tamura, “Ubiquitous speech processing,” Proc. ICASSP2001, Salt Lake City, U.S.A., vol.1, pp.13-16 (2001-5)

[2] I. Mani and M. Maubury, “Advances in Automatic Text Summarization,” The MIT Press, (1999).

[3] K. Zech “A Literature Survey on Information Extraction and Text Summarization,” Paper for Directed Reading, (1996).

[4] C. Hori and S. Furui, “A New Approach to Automatic Speech Summarization,” To appear in the IEEE Transactions on Multimedia (2002).

[5] C. Hori, S. Furui, R. Malkin, H. Yu and A. Waibel, “Automatic speech summarization applied to English broadcast news speech,” Proc. ICASSP2002, Orlando, U.S.A., vol.1, pp.9-12 (2002-5)

[6] K. Maekawa, H. Koiso, S. Furui, H. Isahara “Spontaneous Speech Corpus of Japanese,” Proc. LREC2000, Athens, pp.947-952 (2000-5)