AUTOMATIC SPEECH SUMMARIZATION BASED ON 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

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 based on the amount of information and the confidence measures of constituent words, and the set of extracted sentences is compressed by our sentence compaction method. The sentence compaction is performed by selecting a word set that maximizes a summarization score consisting of 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 has been confirmed by summarizing a spontaneous presentation.

1. INTRODUCTION

Speech recognition has two major applications[1] :transcribing ubiquitous speech documents such as presentations, lectures and broadcast news, and dialogue with computer systems. Since speech is the most natural and effective way of communication between human beings, the former application is expected to become very important in the IT era. Although high recognition accuracy can be easily obtained for speech reading text such as anchor speakers' broadcast news utterances, it is still very difficult to recognize spontaneous speech. 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, the approach to transcribing all words does not always make sense for spontaneous speech. Instead, speech summarization to extract important information and removing redundant and incorrect information is necessary to be investigated for recognizing spontaneous speech.

Techniques of automatically summarizing written text have been actively investigated in the field of natural language processing. However, many of these techniques cannot be applicable to speech, and the 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[2][3]. The proposed technique can be applied to each sentence utterance as well as whole speech documents consisting of multiple utterances. The 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. 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 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. Section 3 provides 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 new two-stage summarization method consisting of important sentence extraction and sentence compaction. From the speech recognition results, a set of rela-

tively important sentences are 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.

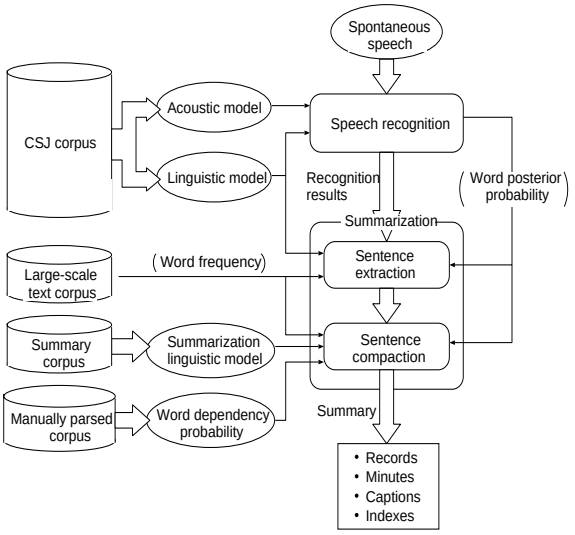

Fig. 1. Automatic speech summarization system.

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 the result of speech recognition:

$$S_s(W) = \frac{1}{N}\sum_{n=1}^{N} \{L_s(w_n) + \lambda_{I_s} I_s(w_n) + \lambda_{C_s} C_s(w_n)\} \quad (1)$$

where N is the number of words consisting the sentence W, and $L_s(w_n)$, $I_s(w_n)$ and $C_s(w_n)$ are the linguistic score, the significance score, and the confidence score of word w_n, 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_s} and λ_{C_s} are weighting factors for balancing the scores.

Details of the scores are as follows.

Linguistic score

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

$$L_s(w_i) = \log P(w_i | \ldots w_{i-1}) \quad (2)$$

In our experiment, trigram probability calculated using transcriptions of presentation utterances in the CSJ (Corpus of Spontaneous Japanese)[4] 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 $I_s(w_i)$ 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 such as nouns, verbs and adjectives by word occurrences in a corpus as shown in Eq.(3). A flat score is given to other words.

$$I_s(w_i) = f_i \log \frac{F_A}{F_i} \quad (3)$$

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 the corpus, that is $\sum_i F_i$.

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 and used for measuring the significance score. Important keywords are weighted and the words having nothing to do with original content such as recognition errors are de-weighted by this score.

Confidence score

The confidence score $C_s(w_i)$ 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, filled pauses are removed from the remaining transcription, and sentence compaction is performed using the method that we have proposed[3]. In this method, all the remaining 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 using a 2-stage dynamic programming(DP) technique. Specifically, each sentence is summarized according to all possible compression ratio, 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 to written editorial style articles and used together with the proceedings of 60 presentations to calculate the trigrams for sentence compaction.

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. Summarization experiments

One of the presentations in the CSJ by a male speaker having a length of roughly 12 minutes has been summarized at the summarization ratios of 70% and 50%. The word recognition accuracy of this presentation is 70% in average. Specification of the recognition system is as follows.

Feature extraction

Speech waveform is digitized by 16kHz sampling and 16bit quantization, and a 25-dimensional feature vector consisting of normalized logarithmic energy, 12-dimensional Mel-cepstrum and their derivatives, is extracted using a 24ms frame applied at every 10ms. The cepstral mean subtraction(CMS) is applied for each utterance.

Acoustic and linguistic models

Speaker-independent context-dependent phone HMMs with 3000 states and 16 Gaussian mixtures for each state are made using a part of the CSJ consisting of 338 presentations with the length of 59 hours spoken by male speakers different from the speaker of the presentation for testing. The transcribed presentations in the CSJ with 1.5M words are automatically split into words (morphemes) by the JTAG morphological analysis program, and the most frequent 20k words are selected to calculate word bigrams and trigrams.

Decoder

A word-graph-based 2-pass decoder is used for recognition. In the first pass, frame-synchronous beam search is performed using the above-mentioned HMM and the bigram language model. A word graph generated as a result of the first pass is rescored in the second pass using the trigram language model.

3.2. Summarization accuracy

To automatically evaluate summarized sentences, correctly transcribed presentation speech is manually summarized by nine human subjects and used as correct 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 summarization covering subjective variations. Word accuracy of automatic summarization is calculated as the summarization accuracy using the word network[3].

Evaluation of a sentence removed at the sentence extraction stage is performed as follows; if there exists a direct path from the sentence beginning <s> to the sentence ending </s> in the word network, the summarization accuracy for that sentence is 100% (no error), and if the direct path does not exist, it is considered that there exist deletion errors of all the words in the sentence.

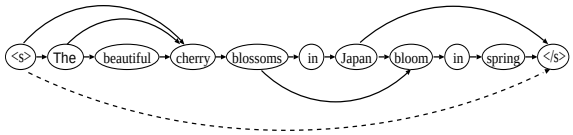

Fig. 2. Word network made by merging manual summarization results.

3.3. Evaluation conditions

Summarization has been performed under the following nine conditions; single-stage summarization without applying the important sentence extraction (NOS), two-stage summarization using seven kinds of the combination of scores for important sentence extraction (L_s, I_s, C_s, L_s-I_s, I_s-C_s, C_s-L_s, L_s-I_s-C_s), and summarization by random word selection. The weighting factors, λ_{L_s}, λ_{I_s} and λ_{C_s}, are set at optimum values for each experimental condition.

3.4. Evaluation results

Results of 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. In these experiments, the division of summarization ratio into the two stages has been experimentally optimized.

Figure 5 shows the summarization accuracy as a function of the ratio of compression by sentence extraction in the total summarization ratio at the 50% and 70% summarization conditions. This result indicates that the best summarization accuracy can be obtained when 2/3 and 1/2 of the

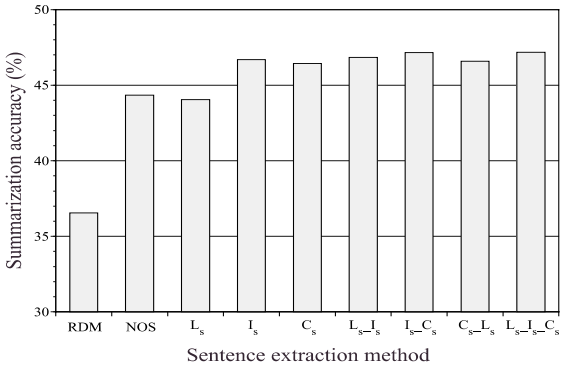

Fig. 3. Summarization at 50% summarization ratio.

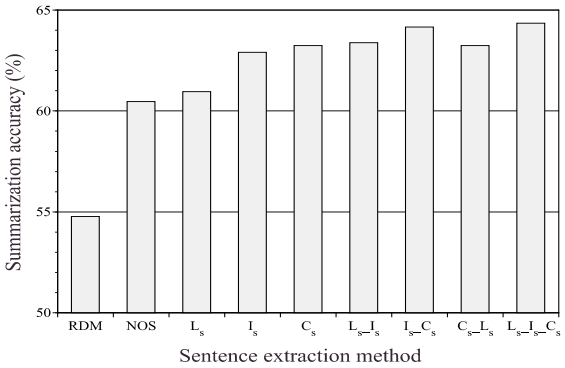

Fig. 4. Summarization at 70% summarization ratio.

compression is performed by the sentence extraction under the condition of 50% and 70% summarization ratio, respectively.

Comparing the three scores for the sentence extraction, the significance score (I_s) or the confidence score (C_s) achieves better results than the linguistic score (L_s), improving the summarization accuracy by 2% compared with the one-stage method. By combining the two scores (I_s_C_s) in the sentence extraction, improvement of the summarization accuracy compared with the one-stage method further reaches to 3%. Since the linguistic score is much less effective than other two scores, the combination of all three scores shows only a minor improvement compared with the combination of only the significance and the confidence scores.

4. 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 has been confirmed in spontaneous presentation speech summarization that combining

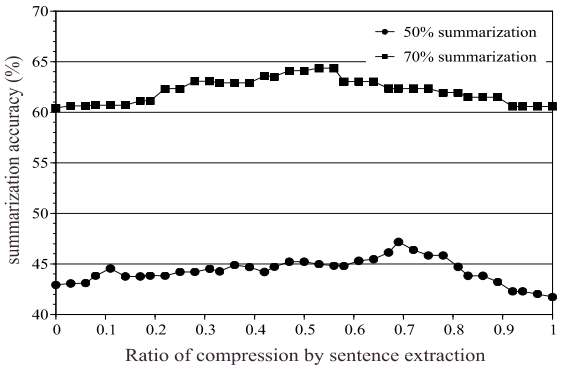

Fig. 5. Summarization accuracy as a function of the ratio of compression by sentence extraction in the total of summarization ratio.

sentence extraction with sentence compaction is effective, and the new method achieves better summarization performance than our previous one-stage method. It has also been found that the word significance score and the word confidence score are effective to extract important sentences. The two-stage method is effective to avoid producing short unreadable sentences, one of the problems that the one-stage method had.

Future research includes evaluation by a larger testing data with manual summary, investigation of other useful information/features for important sentence extraction, and automatically optimizing of the division of compression ratio into the two summarization stages.

5. ACKNOWLEDGEMENT

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

6. 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] C. Hori and S. Furui, “Summarized Speech Sentence Generation Based on Word Extraction and Its Evaluation,” Trans. IEICE, D-II, Vol. J85-D-II, No.2, pp.200-209 (in Japanese)

[3] 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)

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