SENTENCE EXTRACTION-BASED PRESENTATION SUMMARIZATION TECHNIQUES AND EVALUATION METRICS

Makoto Hirohata, Yousuke Shinnaka, Koji Iwano and Sadaoki Furui

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

ABSTRACT

This paper presents automatic speech summarization techniques and its evaluation metrics, focusing on sentence extraction-based summarization methods for making abstracts from spontaneous presentations. Since humans tend to summarize presentations by extracting important sentences from introduction and conclusion parts, this paper proposes a method using sentence location. Experimental results show that the proposed method significantly improves automatic speech summarization performance for the condition of 10% summarization ratio. Results of correlation analysis between subjective and objective evaluation scores confirm that objective evaluation metrics, including summarization accuracy, sentence F-measure and ROUGE-N, are effective for evaluating summarization techniques.

1. INTRODUCTION

One of the major applications of automatic speech recognition is for transcribing spontaneous speech documents, such as presentations, lectures and interviews. Compared to speech read from a text such as in broadcast news utterances, recognition accuracy for spontaneous speech is still limited (60-80%). Spontaneous speech is ill-formed and usually includes redundant information such as disfluencies, fillers, repetitions, repairs, and word fragments. Therefore, simple transcription is unusable because of lengthy expressions. In addition, recognition errors cause transcriptions obtained from spontaneous speech to include irrelevant or incorrect information. Therefore, the processes of removing incorrect information and extracting important information are necessary for transcribing spontaneous speech. Automatic speech summarization is one approach toward accomplishing this goal.

This paper investigates and evaluates sentence extraction-based speech summarization techniques under the condition of 10% summarization ratio, and assesses various objective evaluation metrics in comparison with summaries made by human subjects.

A sentence compaction-based summarization method has already been investigated for high summarization ratios, such as 30% to 70% [1]. We have proposed a two-stage summarization method consisting of important sentence extraction and word-based sentence compaction, as shown in Figure 1 [2]. It has been confirmed that sentence extraction plays an important role in improving summarization performance, especially when the summarization ratio is relatively low.

Although various metrics of objectively evaluating summarization techniques have been investigated (e.g. [3]), they have not yet been assessed in the framework of speech summarization.

The remainder of the paper is organized as follows. Section 2 introduces sentence extraction methods investigated in this paper, Section 3 presents objective evaluation metrics, and Section 4 describes experimental conditions and results. Finally, Section 5 concludes the paper.

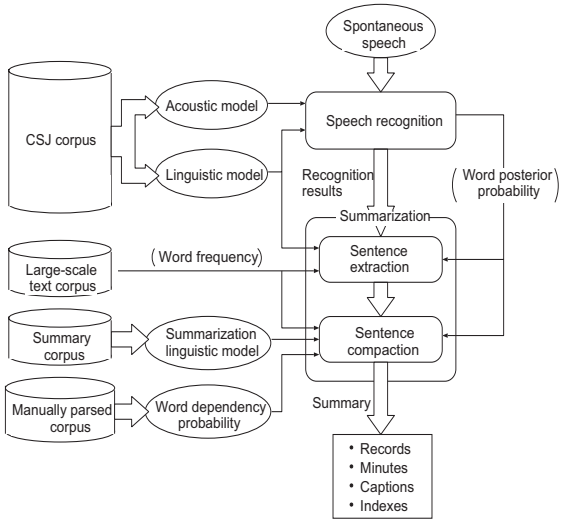

Fig. 1. The two-stage automatic speech summarization process consisting of sentence extraction and compaction.

2. SENTENCE EXTRACTION METHODS

2.1. Extraction using a significance score

A method for extracting important sentences according to linguistic, significance and confidence scores of each sentence has been investigated at 50% and 70% summarization ratios [2]. Since the significance score has proved extremely useful for summarization, this paper investigates a sentence extraction method simply using the significant score. Specifically, for each sentence obtained by speech recognition, $W = w_1, w_2, \ldots, w_N$, the following score is measured:

$$Score(W) = \frac{1}{N} \sum_{n=1}^{N} I(w_n) \quad (1)$$

where N is the number of words constructing the sentence W, and $I(w)$ is the significance score defined by Eq. 2 which is similar to a tf/idf measure.

$$I(w) = f_w \cdot icf = f_w \cdot \log \frac{F_A}{F_w} \quad (2)$$

Here, f_w is the number of occurrences of a word w in all the utterances in the presentation, F_w is the number of occurrences of w in a large corpus, and F_A is the number of all content words in the corpus. A fixed number of sentences having relatively high significance scores are selected.

In this paper, spontaneous presentations included in the Corpus of Spontaneous Japanese (CSJ) are used in the experiments. In order to measure the significance score, the number of occurrences of 50k kinds of words in the CSJ, consisting of transcribed presentations with 8M words, is computed.

2.2. Extraction using latent semantic analysis

Extraction using latent semantic analysis is one of the summarization techniques based on Singular Value Decomposition (SVD) [4]. The SVD semantically clusters content words and sentences, and derives a latent semantic structure for a presentation speech. The results of applying the SVD to each presentation is shown in Figure 2. The target matrix **A** represents the presentation speech, and the element a_{ji} of **A** is calculated as follows:

$$a_{ji} = f_{ji} \cdot icf \quad (3)$$

where f_{ji} is the number of occurrences of a content word j ($j \leq J$) in sentence i ($i \leq I < J$).

Each singular vector represents a salient topic. The singular vector with the largest corresponding singular value represents the topic that is the most salient in the presentation speech. Therefore, a fixed number of singular vectors having relatively large singular values are selected, and for each singular vector, a sentence having the largest index is extracted as an important sentence. The extracted sentences best describe the topics represented by the singular vectors.

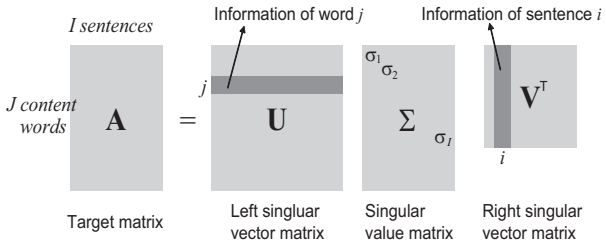

Fig. 2. Application of Singular Value Decomposition (SVD) to each presentation.

2.3. Extraction using dimension reduction based on SVD

Extraction using dimension reduction is another summarization technique based on the SVD. As shown in Figure 3, each sentence vector A_i is projected onto a weighted singular-value vector space. For evaluation of each sentence i, the score of each sentence is calculated by the norm in lower ($K \ll I$) dimensional space:

$$Score(i) = \|\Psi_i\| = \sqrt{\sum_{k=1}^{K} (\sigma_k v_{ik})^2} \quad (4)$$

A fixed number of sentences having relatively large scores in the K-dimensional space are selected. K is experimentally set at 5 in this paper.

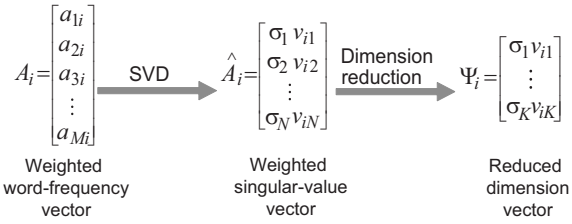

Fig. 3. Dimension reduction process using the results of SVD.

2.4. Extraction using sentence location

Figure 4 shows the ratio of extracted sentences as a function of sentence location in the presentation for manual summarizations with 10% and 50% summarization ratios carried out by three human subjects. 169 presentations were used in the analysis. Each presentation was split into 10 segments having approximately equal number of sentences. This result shows that human subjects tend to extract sentences from the first (introduction) and the last (conclusion) segments under the condition of 10% summarization ratio, whereas there is no such tendency at 50% summarization ratio.

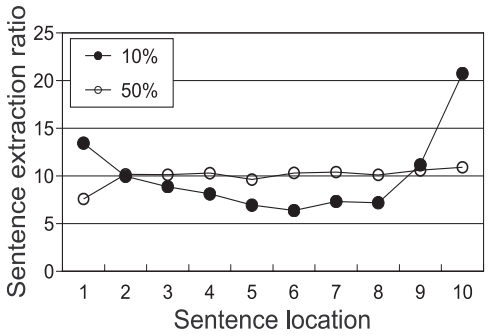

Fig. 4. Sentence extraction ratio in manual summarization as a function of sentence location in the presentation.

These results prompted investigating sentence extraction using sentence location, focusing on the introduction and conclusion segments, as shown in Figure 5. The introduction and conclusion segments are estimated based on the Hearst method [5] using sentence cohesiveness. The cohesiveness is measured by a cosine value between content word-frequency vectors consisting of more than a fixed number of content words. Each segmentation boundary is the first sentence from the beginning or end of the presentation speech, where cohesiveness becomes lower than a preset threshold. This extraction method is used in combination with the sentence extraction methods described above.

3. OBJECTIVE EVALUATION METRICS

Since it is impossible to always have human evaluation of automatic summarization results, it is indispensable to develop objective evaluation metrics.

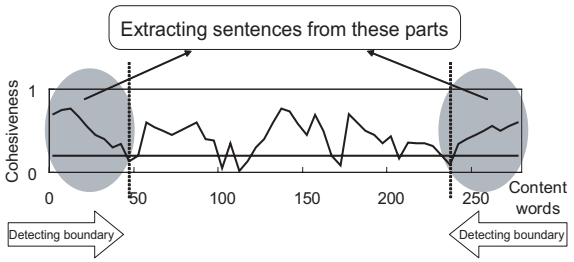

Fig. 5. Principle of sentence extraction from estimated introduction and conclusion segments in each presentation.

3.1. Summarization accuracy

Since manual summaries vary according to human subjects, all the human summaries are merged into a single word network, which is considered to approximately cover all possible correct summaries. Word accuracy of the automatic summary is then measured as a summarization accuracy SumACCY [6] in comparison with the closest word string extracted from the word network.

This metric works reasonably well at relatively high summarization ratios such as 50%, but has problems at low summarization ratios such as 10%, since the variation between manual summaries is so large that the network accepts inappropriate summaries. Therefore, we investigated word accuracy obtained by individually using the manual summaries (SumACCY-E). In this metric, the largest score of SumACCY-E among human summaries (SumACCY-E/max) or the average score of SumACCY-E (SumACCY-E/ave) is used. SumACCY-E/max is equivalent to the NrstACCY proposed in [6].

3.2. Sentence recall/precision

Sentence recall/precision is commonly used in evaluating sentence-extraction-based text summarization. Since sentence boundaries are not explicitly indicated in input speech, estimated boundaries based on recognition results do not always agree with those in manual summaries. In [7], extraction of a sentence in the recognition result is considered as extraction of one or multiple sentences in the manual summary with an overlap of 50% or more words. In this metric, sentence recall/precision is measured by the largest score (F-measure/max) or the average score (F-measure/ave) of the F-measures.

3.3. ROUGE-N

ROUGE-N is an N-gram recall between an automatic summary and a set of manual summaries [3]. ROUGE-N is computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{S \in S_H} \sum_{g_n \in S} C_m(g_n)}{\sum_{S \in S_H} \sum_{g_n \in S} C(g_n)} \qquad (5)$$

where S_H is a set of manual summaries, S is an individual manual summary, g_n is an N-gram, $C(g_n)$ is the number of g_n's in the manual summary, and $C_m(g_n)$ is the number of co-occurrences of g_n in the manual summary and the automatic summary. In our experiments, 1-grams, 2-grams, and 3-grams were examined.

4. EXPERIMENTS

4.1. Experimental conditions

30 presentations by 20 males and 10 females in the CSJ were automatically summarized at 10% summarization ratio. Mean word recognition accuracy was 69%. In order to automatically extract important sentences, sentence boundaries in the recognition results were automatically determined using language models, which achieved 72% recall and 75% precision. The technique of extracting sentences according to sentence location (IC) was combined with each of the three sentence extraction methods described in Subsections 2.1, 2.2 and 2.3. Thus, the following six summarization methods were evaluated: extraction using a significance score (SIG); latent semantic analysis (LSA); dimension reduction based on SVD (DIM); SIG combined with IC (SIG+IC); LSA combined with IC (LSA+IC); and DIM combined with IC (DIM+IC).

4.2. Subjective evaluation

In order to establish criteria for evaluating automatic summary, 180 automatic summaries (30 presentations x 6 summarization methods) were evaluated by 12 human subjects. The summaries were evaluated in terms of ease of understanding and appropriateness as summaries in five levels: 1-very bad; 2-bad; 3-normal; 4-good; 5-very good. The subjective evaluation results were converted into factor scores using factor analysis in order to normalize subjective differences.

4.3. Subjective evaluation results

Figure 6 shows the normalized subjective score for each summarization method averaged over 30 presentations. By combining the IC method using sentence location, every summarization method was significantly improved. SIG+IC achieved the best score, but the difference between SIG+IC and DIM+IC was not significant.

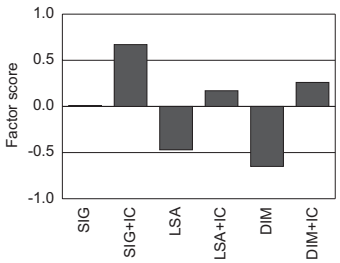

Fig. 6. Subjective evaluation results represented by the normalized score.

4.4. Correlation between subjective and objective evaluation results

In order to investigate the relationship between subjective and objective evaluation results, the automatic summaries were evaluated by eight objective evaluation metrics: SumACCY, SumACCY-E/max, SumACCY-E/ave, F-measure/max, F-measure/ave, ROUGE-1, ROUGE-2, and ROUGE-3.

Table 1 shows correlation coefficients between subjective and objective evaluation scores averaged over all the presentation

speeches, and Figure 7 shows the results of regression analysis in the cases of SumACCY-E/max and ROUGE-3, having relatively high correlation coefficients. All the objective metrics yielded correlation with human judgment. If the effect of word recognition accuracy for each sentence is removed, all the metrics, except ROUGE-1, yield high correlations. ROUGE-1 measures overlapping 1-grams, which probably causes the correlation between ROUGE-1 and the recognition accuracy.

Table 1. Correlation with subjective evaluation results for each objective evaluation metric. The evaluation scores were averaged over all the presentations for each sentence extraction method before calculating the correlation.

Objective metric	Correlation coefficient
SumACCY	0.82
SumACCY-E/max	0.96
SumACCY-E/ave	0.92
F-measure/max	0.95
F-measure/ave	0.96
ROUGE-1	0.85
ROUGE-2	0.94
ROUGE-3	0.96

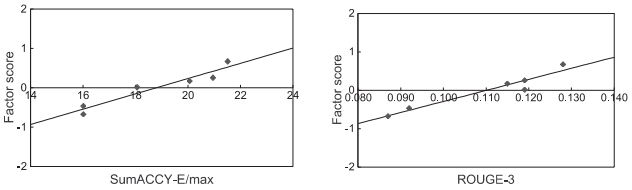

Fig. 7. Regression analysis for the relationships between subjective and objective evaluation scores in the cases of SumACCY-E/max and ROUGE-3. The evaluation scores were averaged over all the presentations for each summarization method.

Table 2 shows the correlation between subjective and objective evaluation scores for each presentation speech. The results of regression analysis are shown in Figure 8. In contrast with the results averaged over all the presentations, no metric has strong correlation. This is due to the large variation of scores over the whole set of presentations.

Table 2. Correlation with subjective evaluation results for each objective evaluation metric. Individual evaluation scores for the presentations were used in calculating the correlation.

Objective metric	Correlation coefficient
SumACCY	0.35
SumACCY-E/max	0.38
SumACCY-E/ave	0.33
F-measure/max	0.47
F-measure/ave	0.50
ROUGE-1	0.43
ROUGE-2	0.48
ROUGE-3	0.48

5. CONCLUSION

This paper has presented several sentence extraction methods for automatic presentation speech summarization and objective evaluation metrics. We have proposed sentence extraction methods using dimension reduction based on SVD and sentence location. Under the condition of 10% summarization ratio, it was confirmed that the method using sentence location improves summarization results. The method using a significance score or dimension reduction based on SVD, combined with the sentence location-based extraction method, achieved the best performance. Among the objective evaluation metrics, SumACCY, SumACCY-E, F-measure, ROUGE-2 and 3 were found to be effective. Although the correlation between the subjective and objective scores averaged over presentations is high, the correlation for each individual presentation is not so high due to the large variation of scores across presentations.

Future research includes investigation of other objective evaluation metrics, evaluation of summarization methods containing sentence compaction, and producing optimum summarization techniques by employing objective evaluation metrics.

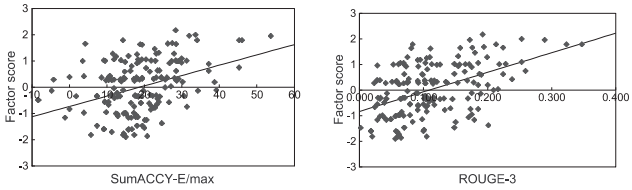

Fig. 8. Regression analysis for the relationships between subjective and objective evaluation scores in the cases of SumACCY-E/max and ROUGE-3. Individual evaluation scores for the presentations were used.

6. ACKNOWLEDGMENTS

This research has been supported by the 21st Century COE Program “Framework for Systematization and Application of Large-scale Knowledge Resources”.

7. REFERENCES

[1] C. Hori and S. Furui, “Advances in automatic speech summarization,” Proc. Eurospeech, pp. 1771-1774 (2001)

[2] S. Furui, T. Kikuchi, Y. Shinnaka and C. Hori, “Speech-to-text and speech-to-speech summarization of spontaneous speech,” IEEE Trans. Speech and Audio Process., 12, 4, pp. 401-408 (2004)

[3] C.-Y. Lin, “Looking for a few good metrics: ROUGE and its evaluation,” Working Notes of NTCIR-4 (Vol. Supl. 2), pp. 1-8 (2004)

[4] Y. Gong and X. Liu, “Generic text summarization using relevance measure and latent semantic analysis,” Proc. SIGIR, pp. 19-25 (2001)

[5] M. A. Hearst, “TextTiling: Segmenting text into multi-paragraph subtopic passages,” Computational Linguistics, 23, 1, pp. 33-64 (1997)

[6] C. Hori, T. Hirao and H. Isozaki, “Evaluation measures considering sentence concatenation for automatic summarization by sentence or word extraction,” Proc. ACL, pp. 82-88 (2004)

[7] T. Kitade, H. Nanjo and T. Kawahara, “Automatic extraction of key sentences from CSJ presentations using discourse markers and topic words,” Proc. the Third Spontaneous Speech Science and Technology WorkShop, pp. 111-118 (2004)