AUTOMATIC SPEECH SUMMARIZATION BASED ON WORD SIGNIFICANCE AND LINGUISTIC LIKELIHOOD

Chiori Hori and Sadaoki Furui

Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
2-12-1, O-okayama, Meguroku, Tokyo, 152-8552 Japan
e-mail : {chiori,furui}@cs.titech.ac.jp

ABSTRACT

This paper proposes a new method of automatically summarizing speech by extracting a limited number of relatively important words from its automatic transcription according to a target compression ratio for the number of characters. To determine a word set to be extracted, we define a summarization score consisting of a topic score (significance measure) of words and a linguistic score (likelihood) of the word concatenation. A set of words maximizing the score is efficiently selected using a dynamic programming (DP) technique. Japanese broadcast news speech transcribed using a large vocabulary continuous speech recognition system was summarized. As a result 86% of important words in the original speech were correctly included in the summarizing sentences and 72% of the summarizing sentences could maintain the meanings of the original speech under the 60–70% summarization condition.

1. INTRODUCTION

Recently, large-vocabulary continuous-speech recognition (LVCSR) technology has been making significant advancements. Major applications of the LVCSR systems in the near future will include automatic closed captioning for broadcast news and meeting/conference summarization.

Since Japanese text is written with a mixture of three types of characters: Chinese characters (Kanji) and two types of Japanese characters (Hira-gana and Kata-kana), it is impossible even for professional typists to transcribe speech in real time. Therefore we are now developing an automatic closed captioning system using speech recognition technology with NHK broadcasting company. In closed captioning, the number of words presented on the TV screen for a professional announcers' broadcast news speech sometimes exceeds the number of words that people can read and understand. In addition transcribed speech usually includes some redundant information. Therefore a summarization technique is desired to be applied to the closed captioning system to compress the information of speech. Furthermore summarization for transcribed speech is also expected to be useful for indexing speech data for automatic retrieval and for making abstracts of presentations and minutes of meetings.

Techniques of automatically summarizing written text have been actively investigated in the field of natural language processing. The major goal of these investigations has been to select one or several important sentences from a set of sentences comprising a paragraph using several word/sentence significance measures. Wakao et al.[1] have recently proposed a technique of summarizing broadcast news text. They selected important sentences using significance measures associated with keywords based on their frequency in news manuscripts. Then they applied Japanese language-specific summarization rules to the selected sentence. The technique was used in an experiment to produce closed captions using TV news text.

In this study, we propose a new method of automatically summarizing broadcast news speech focusing on topic words and linguistic likelihood. In this method, speech summarization is considered as a process to extract a sequence of words from a transcribed sentence so that the sequence becomes a feasible Japanese sentence including topic words. We employ a dynamic programming technique to determine the words to be extracted. Validity of the summarizing sentences derived using this method is evaluated by eight human subjects.

2. APPROACH TO SUMMARIZATION OF SPEECH

To summarize a sentence, we extract a limited number of relatively important words from each sentence so that the number of characters maintains around a specified ratio to the number of characters in the original sentence. The words are extracted using a summarization score consisting of a topic score (significance measure) of extracted words and a linguistic score (likelihood) of the word concatenation. A set of words that maximizes the summarization score is selected using a dynamic programming (DP) technique. This method is effective in reducing the number of words without loosing important information.

2.1. Summarization score

The summarization score, consisting of a topic score and a linguistic score, is calculated as follows. Given a transcription result consisting of N words, $W = w_1, w_2, \ldots, w_N$, the summarization score of the extracted M ($M < N$) words, $V = v_1, v_2, \ldots, v_M$, is obtained by,

$$S(V) = \sum_{m=1}^{M} \{\log P(v_m|v_{m-2}v_{m-1}) + \lambda I(v_m)\} \qquad (1)$$

where trigram probability $P(v_m|v_{m-2}v_{m-1})$ is used as a linguistic score of the summarizing sentence and $I(v_m)$ is the topic score.

Since we found in our previous experiments using human subjects that most of the topic words are nouns, the topic score is only calculated for nouns. The score is calculated as follows using the significance measure chosen in our previous experiment[2].

$$I(w_i) = g_i \log \frac{G_A}{G_i} \qquad (2)$$

w_i :	a noun in the transcribed speech
g_i :	number of occurrences of w_i in the transcribed article
G_i :	number of occurrences of w_i in all the training news articles
G_A :	summantion of all $G_i$in all the training news articles($= \sum_i G_i$)

A flat score is given to words other than nouns. λ is a weighting factor for balancing the topic score and the linguistic score. A large λ gives more weight to important words and a small λ gives more weight to linguistic feasibility as Japanese. A dynamic programming method can be used to determine a word set which maximizes the summarization score as follows.

2.2. Dynamic programming for automatic summarization

Given a transcription result consisting of N words, $W = w_1, w_2, \ldots, w_N$, the summarization is performed by extracting a set of $M(M < N)$ words, $V = v_1, v_2, \ldots, v_M$, which maximizes the summarization score given by eq.(1). The algorithm is as follows.

1. definition of symbols and variables

<s>	: beginning symbol of a sentence
</s>	: ending symbol of a sentence
$g(m,l,n)$	: local optimal score

(summarization score of the sub-sentence, <s>$, \ldots, w_l, w_n$, consisting of m words, beginning with <s>, and ending with w_l,w_n $(0 \leq l < n \leq N)$)

$B(m,l,n)$	: back pointer

2. initialization

$$g(1,0,n) = \begin{cases} \log P(w_n|\texttt{<s>}) + \lambda I(w_n) & \\ \qquad\qquad if\ 1 \leq n \leq (N-M+1) & \\ -\infty \qquad\qquad\qquad otherwise & \end{cases}$$

3. DP process

```
for $m = 2$ to $M$
  for $n = m$ to $N - m + 1$
    for $l = m - 1$ to $n - 1$

      $g(m,l,n) = \max_{k<l}\{g(m-1,k,l)$
                  $+ \log P(w_n|w_k w_l) + \lambda I(w_n)\}$
      $B(m,l,n) = \mathrm{argmax}_{k<l}\{g(m-1,k,l)$
                  $+ \log P(w_n|w_k w_l) + \lambda I(w_n)\}$
```

4. select the optimal path

$$S(\hat{V}) = \max_{\substack{N-M<n\leq N \\ N-M-1<l\leq N-1}} g(M,l,n) + \log P(\texttt{</s>}|w_l w_n)$$

$$(\hat{n},\hat{l}) = \mathop{\mathrm{argmax}}_{\substack{N-M<n\leq N \\ N-M-1<l\leq N-1}} g(M,l,n) + \log P(\texttt{</s>}|w_l w_n)$$

5. traceback

```
for $m = M$ to 1
  $v_m = w_{\hat{n}}$
  $l' = B(m,\hat{l},\hat{n})$
  $\hat{n} = \hat{l}$
  $\hat{l} = l'$
```

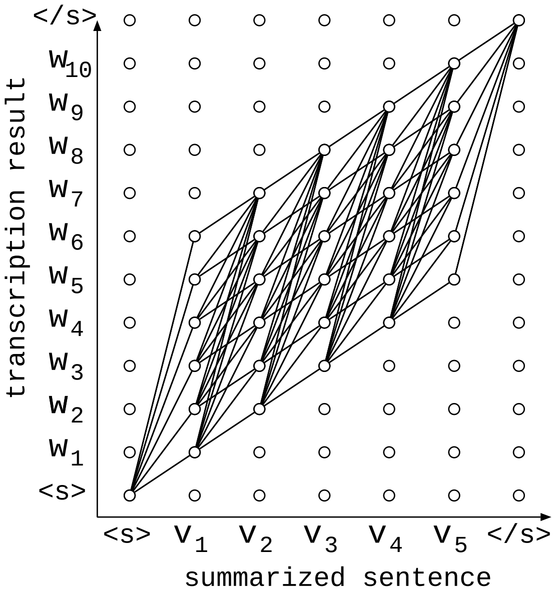

Figure 1: An example of DP alignment for speech summarization.

The two-dimensional space for performing the dynamic programming process is shown in Fig. 1. Since the summarization score obtained by the above equation increases as a function of the number of words in the summarized sentence, the summarization scores cannot be compared directly with summarized sentences having different number of words. In order to remove the effect of the sentence length, a normalization factor is applied to the summarization score and the sentence which maximizes the normalized score $S^*(M)$ is selected as follows.

$$S^*(M) = \bar{S}(M) - \bar{S}_{adj}(M) \qquad (3)$$

where

$$\bar{S}(M) = S(M)/M$$

$$\bar{S}_{adj}(M) = \frac{\bar{S}(M_{max}) - \bar{S}(M_{min})}{M_{max} - M_{min}}(M - M_{min})$$

$S(M)$: a summarization score for a summarized sentence of length M, which is derived by the DP process

M : a number of words($M_{min} \leq M \leq M_{max}$)

M_{min} : the minimum number of words

M_{max} : the muximum number of words

3. STRUCTURE OF THE BROADCAST NEWS TRANSCRIPTION SYSTEM

3.1. Acoustic Models

The feature vector extracted from speech consists of 16 cepstral coefficients, normalized logarithmic power, and their delta features (derivatives). The total number of parameters in each vector is 34. Cepstral coefficients were normalized using the CMS (cepstral mean subtraction) method. The acoustic models used were shared-state triphone HMMs designed using tree-based clustering. The total number of states was 2,106, and the number of Gaussian mixture components per state was 4. They were trained using phonetically-balanced sentences and dialogues read by 53 speakers (approximately 20 hours in total). They are completely different from the broadcast news task. All of the speakers were male, and so the HMMs were gender-dependent models. The total number of training utterances was 13,270 and the total length of the training data was approximately 20 hours.

3.2. Language Model

Broadcast-news manuscripts recorded from August 1992 to May 1996, comprising of approximately 500k sentences consisting of 22M words, were used for constructing language models. The vocabulary size is 20k words. To calculate word n-gram language models, we segmented the broadcast-news manuscripts into words by using a morphological analyzer since Japanese sentences are written without spaces between words. We constructed the language model so that it depends on the readings of words in order to prevent recognition errors caused by context-dependent readings of Kanji characters. Since our previous experimental results of summarization showed that inappropriate part-of-speech concatenation caused change of meanings after summarization [3], we decided to split a word into separate units for language modeling according to the part-of-speech, even if they share the same characters and the same reading.

3.3. Decoder

We used a word-graph-based 2-pass decoder for transcription. In the first pass, frame-synchronous beam search was performed using the above-mentioned HMMs and a bigram language model. A word graph was generated as a result of the first pass. In the second pass, the word graph was rescored using a trigram language model to derive the final transcription which was then used for summarization.

4. EVALUATION EXPERIMENTS

4.1. Evaluation data

News speech data broadcast on TV in June 1996 was used as a test set to evaluate our proposed method. The set consisted of 48 utterances by five anchor speakers, and was manually segmented into sentences. The out-of-vocabulary (OOV) rate for the 20k word vocabulary is 1.8%. 21 utterances with word recognition accuracy above 90%, which was the average rate over the 48 utterances, were selected and used for the evaluation. The summarization ratio, the ratio of the number of characters in the summarizing sentences to that in the original sentences, was varied between 60% and 70%. 70% was reported as an appropriate ratio for summarization in closed captioning so that the results could be easily read and still maintained the meaning of original sentences [1]. The optimum summarization sentence was selected for each input speech based on the normalized score described in Section 2.2.

4.2. Language models for summarizing sentences

A trigram language model for summarization was built using text from Mainichi newspapers published in 1996, comprising of approximately 1.7M sentences consisting of 29M words. We did this because we consider newspaper text to usually be more compact and simpler than broadcast news text and therefore more appropriate for building language models for summarization than broadcast news text.

4.3. Evaluation Results

Summarization results were evaluated from two viewpoints. One is the performance of extracting important words from the transcription. The other is the difference between meanings of the summaries and the originals.

(1)Words to be extracted

Eight human subjects classified each word in the test set transcription (recognition result) into one of the three classes of significance, "important", "unnecessary" and "others". Using these categories, the performance of extracting the "important" words in the summarizing sentences was evaluated. The results showed that 86% of the "important" words were correctly extracted.

(2)Meaning of summarizing sentences

The summarizing sentences were classified into one of the following three classes based on the difference of meanings compared with the original sentences.

Same : the summarizing sentence has the same meaning as the original

Inclusive : meaning of the summarizing sentence is included in that of the original

Different : meaning of the summarizing sentence is different from that of the original

Results show that 23% of the summarizing sentences belonged to "same", 49% to "inclusive" and 28% to "different". This means that 72% of the summarizing sentences could maintain the meanings of the original speech.

5. CONCLUSIONS

A new method of automatically summarizing broadcast news speech based on topic words and linguistic likelihood, facilitated by a dynamic programming technique has been proposed. This method can efficiently maintain the meaning of the original speech irrespective of reducing the number of words. Experimental results showed that 86% of important words were correctly included in the summarizing sentences and 72% of the summarizing sentences could maintain the meanings of the original speech under the condition that the number of characters was reduced to 60-70% of the original sentence. Further research includes investigation of better language modeling for scoring summarization sentences so that unnatural connection of words can be avoided. In order to reduce the percentage of incorrectly summarized sentences, we probably need to use higher-level knowledge such as semantics. It is crucial that a large-scale training corpus is constructed which consists of pairs of sentences before and after summarization.

ACKNOWLEDGMENTS

The authors would like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast news database. The authors would also like to thank Dr. T. Fuchi at NTT Cyber Space Laboratories for letting us use their morphological analysis program. This work is supported in part by the International Communications Foundation.

REFERENCES

[1] T. Wakao et al., Computer Processing of Oriental Language, Vol.12, No.1,1998.

[2] S. Furui et al., Proc. DARPA Broadcast News Transcription and Understanding Workshop,pp.144-149, 1998.

[3] C. Hori and S. Furui, Proc. 1999 Japan-China Symposium on Advanced Information Technology, pp. 75-82, 1999.