TOWARD SUMMARIZATION OF BROADCAST NEWS SPEECH

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

The technology of speech summarization will be of use in many applications including automatic closed captioning of TV programs and making minutes of meeting and conferences. This paper proposes a new method of automatically summarizing Japanese broadcast news speech based on topic words extracted by using a large vocabulary continuous speech recognition system.

The proposed method summarizes each sentence independently, and is, therefore, different from major conventional methods that aim choosing one or several key sentences from a set of sentences. To summarize a sentence, we extract a limited number of relatively important words from the sentence according to the target compression ratio to the number of words. To determine a word set to be extracted, we define a summarization score consisting of the topic score (significance measure) of words and the linguistic score (likelihood) of the word concatenation. A set of words which maximizes the score is efficiently selected using the dynamic programming (DP) technique. In this paper, it is also confirmed that speech recognition accuracy can be improved by rescoring word sequence hypotheses using the summarization score.

1. INTRODUCTION

Recently, large-vocabulary continuous-speech recognition (LVCSR) technology has been advanced and dictation of speech reading newspapers can achieve word accuracy of 90% or above with a real time system. Now broadcast news dictation and natural conversational speech recognition are actively investigated. Major applications of the LVCSR systems in the near future will include automatic closed captioning and meeting/conference summarization. Since transcribed speech usually includes some redundant information, summarization technique is indispensable in indexing and making abstracts for automatic retrieval of speech data. In the closed captioning, the number of words presented on the TV screen for professional announcers' broadcast news speech sometimes exceeds the number of words that the people can read and understand. This problem can be solved by summarizing the speech.

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 consisting a paragraph or a text using several significance measures. The measures are defined based on keyword frequencies in a sentence, location of a sentence in a text, words used in titles and headlines in a sentence, text structure indicated by function words such as conjunctions, relationships between words and sentences, and similarity between sentences. Wakao et al.[1] proposed a technique of summarizing broadcast new text. They selected important sentences using significance measures associated with keywords based on their frequency in news manuscripts, and applied Japanese language-specific summarization rules to the selected sentence. The technique was used in an experiment using TV news text to produce closed captions. Their local summarization rules delete useless words, connect short phrases split by conjunction words, replace polite-form nouns and verbs at the end of the sentence with simple and short nouns having the same meaning.

In this paper, 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 set of transcribed words so that the sequence becomes a correct Japanese sentence including topic words. We employ the dynamic programming technique to determine the words to be extracted. We investigated a method of automatically detecting topic words from the nouns obtained by a speech recognition system on the basis of their significance scores. In a preliminary experiment, we investigated many measures which were originally proposed for information retrieval from text databases [2]. The best measure selected in the experiment is used in the summarization method proposed in this paper as a measure to choose topic words. Eight human subjects evaluated the goodness of the summarizing sentences derived using this method.

It is expected that a set of words which maximize the summarization score is likely to overlap with correct words. Therefore we try to rescore the N-best word sequence hypotheses obtained by the speech recognition system using the summarization score, and evaluate the effectiveness of this method for improving the recognition accuracy.

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 words keeps a specified ratio to the number of words in the original sentence. The words are extracted using a summarization score consisting of the topic score (significance measure) of extracted words and the linguistic score (likelihood) of the word concatenation. A set of words that maximizes the summarization score is selected using the dynamic programming (DP) technique. This method is expected to be effective to reduce information without loosing important one.

2.1 Summarization score

The summarization score consisting of the topic score (significance measure) and the linguistic score (likelihood) 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 the following equation.

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

where trigram probability $P(v_m|v_{m-1}v_{m-2})$ 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 to nouns. The score is calculated as follows using the significance measure chosen in our previous experiment[2].

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

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 the words other than nouns. λ is a weighting factor for balancing the topic score and the linguistic score. The larger λ gives more weight to important words and the smaller λ gives more weight to linguistic feasibility as Japanese. The 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 from `<s>`, and ending w_l, w_n $(0 \leq l < n \leq N)$

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

2. initialization

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

3. DP process

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

$$\begin{aligned} g(m,l,n) &= \max_{k<l} g(m-1,k,l) + \log P(w_n|w_k w_l) + \lambda I(w_n) \\ B(m,l,n) &= \operatorname*{argmax}_{k<l} g(m-1,k,l) + \log P(w_n|w_k w_l) + \lambda I(w_n) \end{aligned}$$

4. select the optimal path

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

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

5. traceback

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

The two-dimensional space for performing the dynamic programming process is shown in Fig. 1.

3. SUMMARIZATION SCORE-BASED RESCORING

A set of words that gives relatively large summarization score is expected to have a large potential as correct words. Therefore, N-best word sequence hypotheses obtained by the speech recognition system are rescored using the summarization score in order to evaluate its effectiveness on improvement of the recognition performance.

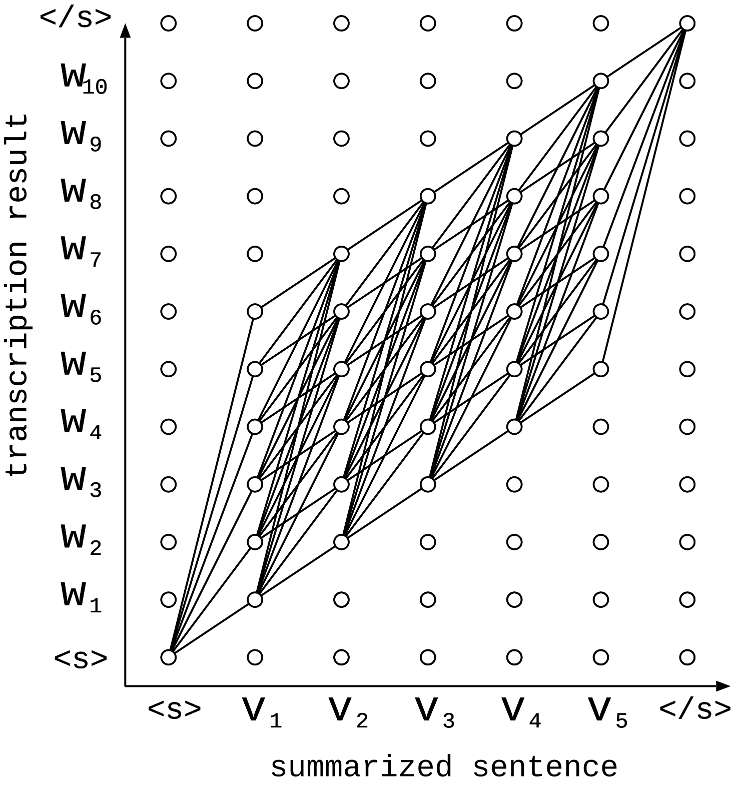

Figure 1: An example of DP alignment for speech summarization.

4. STRUCTURE OF THE BROADCAST NEWS TRANSCRIPTION SYSTEM

4.1 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 coeffrcients were normalized by the CMS (cepstral mean subtraction) method. The acoustic models we 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 (roughly 20 hours in total). They are completely different from the broadcast news task. All of the speakers were male, thus 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.

4.2 Language Model

Broadcast-news manuscripts recorded from August 1992 to May 1996, comprising roughly 500k sentences consisting of 22M words, were used for constructing language models. The vocabulary size is 20k. 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.

4.3 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 and the N-best sentences were derived.

5. Evaluation Experiments

5.1 Evaluation data

News speech data broadcast on TV in June 1996 were 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 1.7%.

5.2 Language models for summarized sentences

Another trigram model was built to measure the linguistic score for summarizing the sentences. This trigram model is different from the model used in the transcription, since the summarizing sentences are considered to be comprised of simpler expressions than original news manuscripts. Although we need a large text database for calculating the trigram, we do not have a large database of manually summarized sentences which should be fee of extraneous words in the original news. Therefore, we automatically made summarizing sentences using the following procedure, as a realistic solution. First, we asked eight human subjects to classify each word in 50 sentences, which were randomly extracted from the news manuscripts, into one of the following three levels of the significance.

(1) important : important word which cannot be removed from the sentence.
(2) unnecessary : unnecessary word;
meaning of the sentence does not change even if it is removed.
(3) others

Next, based on the word classification results, all the language-model training corpus consisting of 500k sentences were "summarized" automatically. That is, all the words belonging to the same grammatical word categy as one of the word classified into "unnecessary" were removed and ending of each sentence was substituted by its simple expression. Thus, a corpus of quasi-summarized sentences was generated. This process, however, could not always correctly summarize the sentences, and the generated corpus included inappropriate Japanese context. Since it was impossible to correct them manually, the summarized and the original corpora were merged with equal weighting to build the trigrams for summarization, so that the influences of the inappropriate contexts could be reduced.

5.3 Evaluation Results

Summarizing sentences were obtained using the transcription obtained by the recognition system described in Section 2. The summarization results were evaluated from two viewpoints; one is the performance of extracting important words from the transcription and the other is the difference between the meaning of summaries and originals. Utterances consisting of more than 25 words with word recognition accuracy above 90% were summarized under the condition that the number of words in the summary was reduced to 70% of the number of words in the original sentence.

1. Words to be extracted
   The same eight subjects as the above experiment classified each word in the test set transcription (recognition result) into one of the three classes of significance, "important", "unnecessary" and "others" in the same way as described above. 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.
   (1) same : the summarizing sentence has the same meaning as the original
   (2) inclusive : meaning of the summarizing sentence is included in that of the original
   (3) different : meaning of the summarizing sentence is different from that of the original

   Results show that 30% of the summarizing sentences belonged to "same", 47% to "inclusive" and 23% to "different". This means that 77% of the summarizing sentences could maintain the meanings of the original speech. In order to improve the performance of the remaining 23% of the speech, we probably need to use higher-level information such as semantics.

5.4 Effects of rescoring by the summarization score

The N-best hypotheses obtained from the second pass in the LVCSR system were rescored using the summarization score at various ratios of the number of words in the summarizing sentences. When the word error rate of the baseline recognition system without rescoring was 17.9%, those after rescoring according to the summarizing ratios of 50%, 70% and 90% were 17.9%, 17.5%, and 17.5%, respectively. These results indicate that rescoring by the summarization score slightly reduces the word error rate. It was also found that summarizing sentences that were most feasible as Japanese sentences did not always improve the recognition performance. Since the word error rate could be reduced in spite of the very crude language model for summarization which was made by using the quasi-summarizing sentences without including any summarization rules, more error reduction is expected by further improvement of the summarization language model. The language model may be improved by using condensed sentences such as those in newspaper articles.

6. CONCLUSIONS

This paper proposed a new method of automatically summarizing broadcast news speech based on topic words and linguistic likelihood, facilitated by the dynami programming technique. This method can efficiently keep 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 77% of the summarizing sentences could maintain the meanings of the original speech. This paper also tried to use the summarization score for rescoring the N-best hypotheses to improve the recognition performance and achieved a small improvement. Further research includes investigation of better language modeling for evaluating summarization sentences so that unnatural connection of words can be avoided. It is crucial to build a large-scale training corpus consisting of the pairs of sentences before and after summarization.

References

[1] T. Wakao et al., “Text Summarisation for Production of Closed-Caption TV Programs in Japanese”,computere processing of oriental language, Vol.12, No.1,1998.

[2] S. Furui et al. “Japanese Broadcast News Transcription and Topic Detection”, Proc. DARPA Broadcast News Trascription and Understanding Workshop,pp.144- 149,1998