Advances in Automatic Speech Summarization

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, Meguro-ku, Tokyo, 152-8552 Japan
e-mail: {chiori,furui}@furui.cs.titech.ac.jp

Abstract

This paper reports recent advances in automatic speech summarization method. In our proposed method, a set of words maximizing a summarization score is extracted from automatically transcribed speech. This extraction is performed according to a target compression ratio using a dynamic programming technique. The extracted set of words is then connected to build a summarized sentence. The summarization score consists of a word significance measure, a confidence measure, linguistic likelihood, and a word concatenation probability which is determined by a dependency structure in the original speech given by Stochastic Dependency Context Free Grammar (SDCFG). Japanese broadcast news speech transcribed using a large vocabulary continuous speech recognition (LVCSR) system is summarized using our proposed method and evaluated in comparison with manual summarization by human subjects. The manual summarization results are combined to build a word network, and word accuracy of each automatic summarization result is calculated comparing with the most similar word string in the network.

1. Introduction

Our goal is to build a system that extracts and presents information from spoken utterances according to the users' desired amount of information. Fig.1 shows our proposed system. In this system, spoken utterances are transcribed using a LVCSR system and summarized. The output of the system can be either a simple set of keywords, a summarized sentence for each utterance or summarization of an article consisting of multiple utterances. These outputs can be used for indexing and making closed captions, abstracts, etc. Since transcribed speech obtained from LVCSR systems not only includes redundant information such as disfluencies, filled pauses, repetitions, repairs, and word fragments, but also includes irrelevant information caused by recognition errors, a summarization technique is indispensable especially for spontaneous speech recognition and understanding.

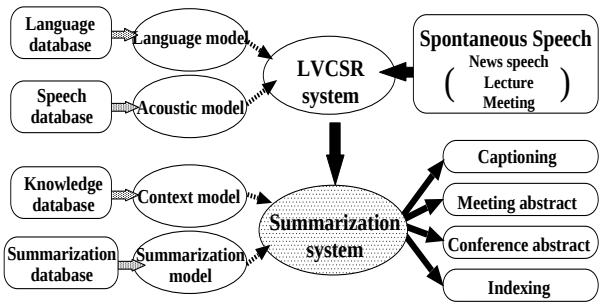

Figure 1: *Automatic speech summarization system.*

2. Summarization of each sentence utterance

Our proposed method to summarize speech, sentence by sentence, extracts a set of words maximizing a summarization score from an automatically transcribed sentence according to a target compression ratio. The summarization score indicating goodness of a summarized sentence consists of a word significance score I as well as a confidence score C of each word in the original sentence, a linguistic score L of the word string in the summarized sentence [1][2], and a word concatenation score T_r. The word concatenation score indicates a word concatenation probability determined by a dependency structure in the original sentence given by SDCFG [3]. The total score is maximized using a dynamic programming (DP) technique [1][2]. This method is effective in reducing the number of words by removing redundant and irrelevant information without losing relatively important information.

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).

$$S(V) = \sum_{m=1}^{M} \{L(v_m|\ldots v_{m-1}) + \lambda_I I(v_m) + \lambda_C C(v_m) + \lambda_T Tr(v_{m-1}, v_m)\} \quad (1)$$

where λ_I, λ_C and λ_T are weighting factors for balancing among L,I, C and T_r.

2.1. Word significance score

The word significance score I indicates relative significance of each word in the original sentence [1]. The amount of information based on the frequency of each word is used as the word significance score for each noun. A flat score is given to words other than nouns. To reduce the repetition of words in the summarized sentence, a flat score is also given to each reappearing noun.

2.2. Linguistic score

The linguistic score $L(v_m|\ \ldots v_{m-1})$ indicates goodness of word strings in a summarized sentence, and is measured by a trigram probability $P(v_m|\ v_{m-2}v_{m-1})$ [1].

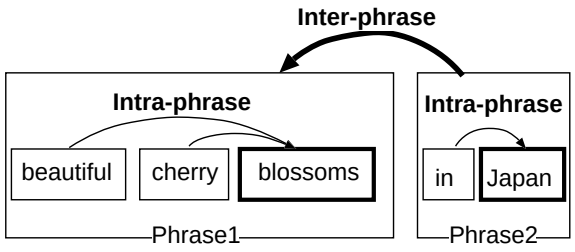

Figure 2: *Word concatenation rule.*

2.3. Confidence score

The confidence score $C(v_m)$ is incorporated to weight acoustically as well as linguistically reliable hypotheses [2]. Specifically, a posterior probability of 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 measure.

2.4. Word concatenation score

Suppose "the beautiful cherry blossoms in Japan" is summarized as "the beautiful Japan". The latter phrase is grammatically correct but semantically incorrect summarization. Since the above linguistic score is not powerful enough to alleviate such a problem, a word concatenation score $Tr(v_{m-1}, v_m)$ is incorporated to give a penalty for a concatenation between words with no dependency in the original sentence.

2.4.1. Word concatenation rule

Word concatenation in a summarized sentence is restricted by the dependency structure in the original sentence as exemplified in Fig.2. Whereas our experiments are conducted for Japanese, the example is shown in English for the sake of explanation. The word concatenation rule augments the words modified by many other words in the same phrase in the original sentence, such as the "blossoms" in Fig.2, so that they remain in the summarized sentence even when the number of words extracted for summarization decreases (intra-phrase rule). The word concatenation rule also gives a score to the concatenation of words in separate phrases in the original sentence based on the dependency structure of the phrases (inter-phrase rule).

Since the dependency structure within a phrase is deterministic, the word concatenation probability between words with dependency within a phrase of the original sentence is set to 1 and that between words without dependency is set to 0. On the other hand, since the dependency structure between phrases is ambiguous, the word concatenation probability between words in different phrases is determined by a probability that one phrase is modified by others based on the SDCFG [3] as follows.

2.4.2. Computation of word concatenation score

Suppose a sentence consists of H phrases, $P_1, \ldots, P_H$. When the kth word, w_k, belongs to a phrase $P_{h(w_k)}$ and the lth word, w_l, belongs to a phrase $P_{h(w_l)}$, the word concatenation score of w_k and w_l in the same phrase ($h(w_k) = h(w_l)$) is defined using the intra-phrase word concatenation rule($R(w_k, w_l) = 0, 1$). On the other hand, the word concatenation score w_k and w_l in the different phrases ($h(w_k) < h(w_l)$) is defined using the probability that $P_{h(w_k)}$ and $P_{h(w_l)}$ have the dependency structure. A word concatenation score $Tr(w_k, w_l)$ is defined as a logarithmic value of the word concatenation probability as shown in eq.(2).

$$Tr(w_k, w_l) = \begin{cases} \log \sum_{i=1}^{h(w_k)} \sum_{j=h(w_l)}^{H} \sum_{\alpha,\beta} g(\alpha \rightarrow \beta\alpha; i, h(w_k), j) \\ \qquad if\ h(w_k) < h(w_l) \\ \log R(w_k, w_l) \qquad if\ h(w_k) = h(w_l) \end{cases} \quad (2)$$

where α,β are nonterminal symbols of SDCFG.

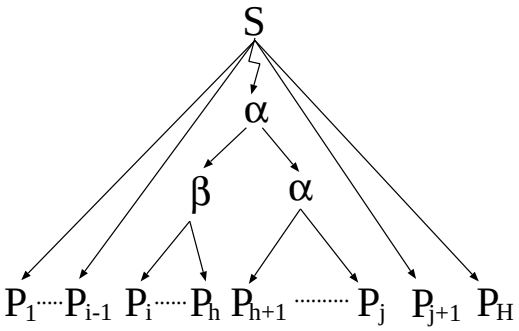

Figure 3: *Inside-Outside probability.*

$g(\alpha \rightarrow \beta\alpha; i, h, j)$ is a posterior probability that the rule of $\alpha \rightarrow \beta\alpha$ is applied and then $P_i \ldots P_h$ is derived from β and $P_{h+1} \ldots P_j$ is derived from α, when a sentence is derived from the initial symbol S as shown in Fig.3. The posterior probability is estimated using the Inside-Outside probability [3].

3. Summarization of multiple utterances with consistent meanings

Our proposed automatic speech summarization technique for each sentence can be extended to summarize a set of multiple utterances (sentences) having consistent meanings by combining a rule which is applied at sentence boundaries. As a result, the original sentences including many informative words are preserved and the sentences including few informative words are deleted or shortened.

Given a transcription result consisting of J utterances, $S_1, \ldots, S_J$ ($S_j = w_{j1}, w_{j2}, \ldots, w_{jN_j}$) the summarization is performed by extracting a set of $M (M < \sum_j N_j)$ 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

   $s_j(k, l, n)$: summarization score of each word

   $$s_j(k, l, n) = \log P(w_{jn}|w_{jk}w_{jl}) + \lambda_I I(w_{jn}) + \lambda_C C(w_{jn}) + \lambda_T Tr(w_{jl}, w_{jn})$$

$P(w_{jn}\|w_{jk}w_{jl})$	: linguistic score
$I(w_{jn})$	: word significance score
$C(w_{jn})$	: confidence score
$Tr(l, n)$	: word concatenation score
$\langle s \rangle$	: beginning symbol of a sentence
$\langle /s \rangle$	: ending symbol of a sentence

$g_j(m,l,n)$: local optimal score of $\langle s\rangle, w_{11}, \ldots,$ consisting of m words beginning with $\langle s\rangle$ of the sentence 1 and ending with w_{jl},w_{jn} in the sentence j ($0 \le l < n \le N_j$)

$G_j(m)$: local optimal score at the end of the sentence, consisting of m words beginning with $\langle s\rangle$ of the sentence 1 and ending with $\langle /s\rangle$ in the sentence j

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

$B_j(m)$: back pointer of the end of sentence

2. Initialization

$$G_0(m) = \begin{cases} \log P(w_{jn}|\langle s\rangle) + \lambda_I I(w_{jn}) + \lambda_C C(w_{jn}) \\ \quad if\ 1 \le n \le N_j\ \ \&\ \ m = 0 \\ -\infty \qquad otherwise \end{cases}$$

$$B_0(m) = \phi$$

3. DP process
for $j = 1$ to J
calculation for the beginning of sentence

$$g_j(m,0,n) = \begin{cases} G_{j-1}(m-1) + \log P(w_{jn}|\langle s\rangle) \\ \quad + \lambda_I I(w_{jn}) + \lambda_C C(w_{jn}) \\ \quad if\ 1 \le n \le N_j \\ -\infty \qquad otherwise \end{cases}$$

$$b_j(m,0,n) = \phi$$

calculation for the inside of sentence
for $m = j \times 2$ to N_j
for $n = 2$ to N_j
for $l = 1$ to $n-1$

$$g_j(m,l,n) = \max_{0\le k<l} \{g_j(m-1,k,l) + s_j(k,l,n)\}$$

$$b_j(m,l,n) = \mathrm{argmax}_{0\le k<l} \{g_j(m-1,k,l) + s_j(k,l,n)\}$$

calculation for the end of sentence

$$G_j(m) = \max_{\substack{0<n\le N_j \\ 0\le l\le N_j-1}} g_j(m,l,n) + \log P(\langle /s\rangle|w_{jl}w_{jn})$$

$$(\hat{n},\hat{l}) = \mathrm{argmax}_{\substack{0<n\le N_j \\ 0\le l\le N_j-1}} g_j(m,l,n) + \log P(\langle /s\rangle|w_{jl}w_{jn})$$

$$B_j(m) = (\hat{n},\hat{l})$$

4. Traceback
$j = J$
$m = M$
while $m > 0$
$v_m = w_{\hat{n}}$
$l' = b_j(m,\hat{l},\hat{n})$
$\hat{n} = \hat{l}$
if $l' \ne \phi$ then
$\hat{l} = l'$
$m = m-1$
else
$v_{m-1} = \langle s\rangle$
$v_{m-2} = \langle /s\rangle$
$(\hat{n},\hat{l}) = B_{j-1}(m-2)$
$m = m-3$
$j = j-1$

Fig.4 illustrates the DP process for summarizing multiple utterances. This summarization technique can be considered as a combination of the summarization method extracting important sentences investigated in the field of natural language processing and our sentence-by-sentence summarization method.

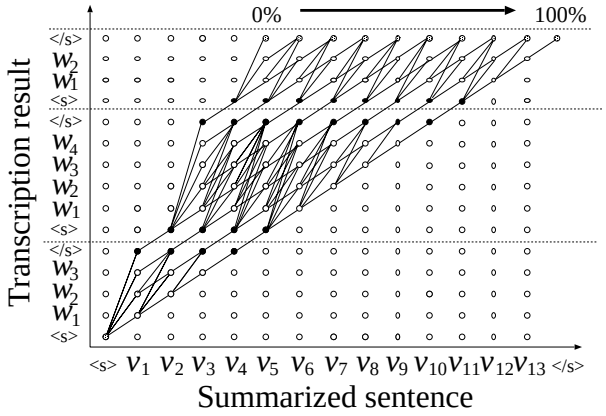

Figure 4: *An example of DP process for summarization of multiple utterances.*

4. Evaluation

4.1. Word network of manual summarization results for evaluation

To automatically evaluate summarized sentences, correctly transcribed speech are manually summarized by human subjects and used as correct targets. The manual summarization results are merged into a word network, and the word accuracy of automatic summarization is calculated using the word network. The network approximately expresses all possible correct summarization including subjective variations. The word accuracy based on the word string that is the most similar to the automatic summarization result extracted from the word network is used as a measure of linguistic correctness and maintenance of original meanings of the utterance (summarization accuracy).

4.2. Evaluation data

Japanese TV broadcast news utterances recorded in 1996 were used to evaluate our proposed method. 50 utterances with word recognition accuracy above 90%, which was the average rate over the 50 utterances, were selected and used for the evaluation. The summarization ratio was set to 70%. In addition, 5 news articles consisting of 5 sentences each were summarized using the summarization technique for multiple utterances at 30% summarization ratio.

4.3. Training data for summarization models

4.3.1. Word significance model

Broadcast-news manuscripts recorded from August 1992 to May 1996, comprising of approximately 500k sentences with 22M words, were used both for building a language model in speech recognition and calculating the word significance measure for summarization.

4.3.2. Language model

A trigram language model for summarization was built using text from Mainichi newspaper published from 1996 to 1998, comprising of 5.1M sentences with 87M words. We consider that the newspaper text is usually more compact and simpler than broadcast news text and therefore more appropriate for building language models for summarization than the latter. Our previous experiments confirmed that the automatically summarized sentences using word trigram based on newspaper text were much better than those by broadcast-news manuscripts [1].

4.3.3. SDCFG

SDCFG for word concatenation score was built using text from the manually parsed corpus of Mainichi newspaper published from 1996 to 1998, comprising of approximately 4M sentences with 68M words. The number of non-terminal symbols was 100.

4.4. Evaluation results

Manual transcription (TRS) and automatic transcription(REC) were summarized. In the summarization of REC, conditions with ($I\_L\_C\_T$) and without (I_L_T) the word confidence score were compared. In summarization for both TRS and REC, conditions with (I_L_T, $I\_L\_C\_T$) and without (I_L,I_L_C) the word concatenation score were compared.

To set the upper limit of the automatic summarization, manual summarization by human subjects for manual transcription (SUB_TRS) was performed. The results were evaluated using all other manual summarization results as correct summarization. In addition, as the upper bound of automatic speech summarization for transcription including speech recognition errors, manual summarization of automatically transcribed utterances was also evaluated (SUB_REC). To ensure that our method is sound, we made randomly generated summarization sentences (RDM) according to the summarization ratio and compared them with those obtained by our proposed methods.

Figure 5 shows results of utterance summarization at 70% summarization ratio and Fig. 6 shows those of summarizing articles having multiple sentences at 30% summarization ratio. These results show that our proposed automatic speech summarization technique is significantly more effective than RDM. The method using the word concatenation score (I_L_T, $I\_L\_C\_T$) can reduce meaning alteration compared with the method without using the word concatenation score (I_L, I_L_C). The better result using the word concatenation score ($I\_L\_C\_T$) compared with that without using the word concatenation score (I_L_T) shows that the summarization accuracy is improved by the confidence score.

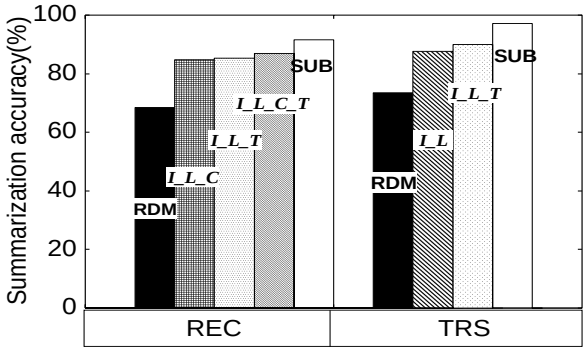

Figure 5: *Each utterance summarizations at 70% summarization ratio.*

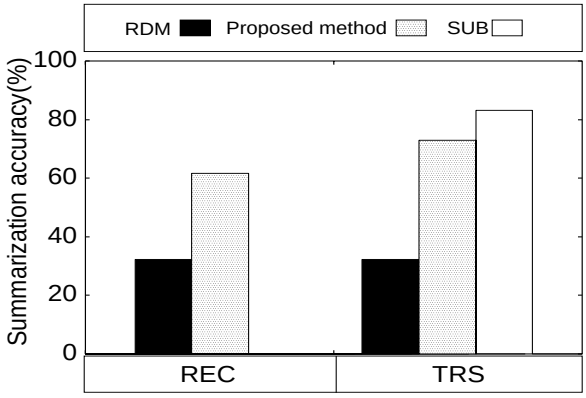

Figure 6: *Article summarizations at 30% summarization ratio.*

5. Conclusions

An automatic speech summarization method based on a word significance score, linguistic likelihood, a word confidence measure and a word concatenation probability has been proposed. A word set maximizing the total score was extracted using the dynamic programming technique and connected to build a summarized sentence. We proposed a new method for measuring the summarization accuracy based on a word network constructed using manual summarization results.

Each utterance and multiple utterances with consistent meanings of Japanese broadcast news speech was summarized by our proposed method. Experimental results show that the proposed method can effectively extract relatively important information and remove redundant and irrelevant information. A confidence score giving a penalty for acoustically as well as linguistically unreliable words could reduce the meaning alteration of summarization caused by recognition errors. A word concatenation score giving a penalty for a concatenation between words with no dependency in the original sentence could also reduce the meaning alteration of summarization.

Future research includes making abstracts of various monologues such as lectures and presentations.

6. Acknowledgment

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

7. References

[1] C. Hori and S. Furui, "Automatic Speech Summarization Based on Word Significance and Linguistic Likelihood", Proc. ICASSP2000, Istanbul, pp.1579-1582 (2000).

[2] C. Hori and S. Furui, "Improvements in Automatic Speech Summarization and Evaluation Methods", Proc. ICSLP2000, Beijing, pp.IV-326-329 (2000).

[3] A. Ito, C. Hori, M. Katoh and M. Kohda, "Language Modeling by Stochastic Dependency Grammar for Japanese Speech Recognition", Proc. ICSLP2000, Beijing, pp.I-246-249 (2000).