論文 / 著書情報
Article / Book Information

Title	Automatic Speech Summarization Applied to English Broadcast News Speech
Authors	Chiori Hori, Sadaoki Furui, Rob Malkin, Hua Yu, Alex Waibel
Citation	IEEE ICASSP 2002, Vol. 1, No. SP-L01.03, pp. 9-12
Pub. date	2002, 5
Copyright	(c) 2002 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
URL	http://www.ieee.org/index.html
DOI	http://dx.doi.org/10.1109/ICASSP.2002.5743641
Note	This file is author (final) version.

AUTOMATIC SPEECH SUMMARIZATION APPLIED TO ENGLISH BROADCAST NEWS SPEECH

Chiori Hori†, Sadaoki Furui†, Rob Malkin‡, Hua Yu‡ and Alex Waibel‡

†Department of Computer Science, Tokyo Institute of Technology,
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552 Japan
e-mail : {chiori,furui}@furui.cs.titech.ac.jp
‡Interactive Systems Labs, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
e-mail : {malkin,hua,ahw}@cs.cmu.edu

ABSTRACT

This paper reports an automatic speech summarization method and experimental results using English broadcast news speech. In our proposed method, a set of words maximizing a summarization score indicating an appropriateness of summarization is extracted from automatically transcribed speech. This extraction is performed using a Dynamic Programming (DP) technique according to a target compression ratio. We have previously tested the performance of our method using Japanese broadcast news speech. Since our method is based on a statistical approach, it could be applied to any language. In this paper, English broadcast news speech transcribed using a speech recognizer is automatically summarized. In order to apply our method to English, the model of estimating word concatenation probabilities based on a dependency structure in the original speech given by a Stochastic Dependency Context Free Grammar (SDCFG) is modified. A summarization method for multiple utterances using two-level DP technique is also proposed.

1. INTRODUCTION

Currently various applications of LVCSR systems, such as automatic closed captioning [1], meeting/conference summarization [2][3] and indexing for information retrieval [4], are actively being investigated. Transcribed speech usually includes not only redundant information such as disfluencies, filled pauses, repetitions, repairs and word fragments, but also irrelevant information caused by recognition errors. Therefore, especially for spontaneous speech, practical applications using speech recognizer require a process of speech summarization which removes redundant and irrelevant information and extracts relatively important information depending on users' requirements. Speech summarization producing understandable sentences from original utterances can be considered as a kind of speech understanding.

We have proposed an automatic speech summarization technique which produces an understandable summarized sentence by extracting relatively important words with relatively high linguistic likelihood and excluding redundant and irrelevant information [5] [6]. In addition, word concatenations in a summarized sentence are restricted by the dependency structure of the original sentence[7]. This summarization process aims to maintain the original meaning as much as possible within a limited number of words. In order to make abstracts, we have proposed a summarization technique which is applicable to multiple utterances[7]. This paper proposes a summarization method for multiple utterances using a two-level Dynamic Programming (DP) technique in order to reduce the amount of calculation.

We have previously investigated the performance of our method using Japanese broadcast news speech. In order to evaluate automatic summarization results, a summarization accuracy score using a word network generated by manual summarizations has been proposed[7]. As a result using this evaluation method, it was ascertained that the summarization method effectively extracts relatively important information and excludes redundant and irrelevant information. Since our method is based on a statistical approach, it can be applied not only to Japanese but also other languages. In this paper, English broadcast news speech transcribed using a speech recognizer[8] is automatically summarized and evaluated.

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 summarization ratio. The summarization ratio is the ratio of the number of characters in the summarized sentence to that in the original sentence. The summarization score indicating the appropriateness of a summarized sentence is defined as the sum of a word significance score I, a confidence score C of each word in the original sentence, a linguistic score L of the word string in the summarized sentence[5][6] and a word concatenation score T_r[7]. The word concatenation score given by SDCFG indicates a word concatenation probability determined by a dependency structure in an original sentence. This method is effective in reducing the number of words by removing redundant and irrelevant information without losing relatively important information. A set of words maximizing the total score is extracted using a DP technique[5].

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).

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

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 $I(v_m)$ indicates the relative significance of each word in the original sentence [5]. The amount of information based on the frequency of each word is used as the word significance score for topic words. We choose nouns and verbs as topic words for English. A flat score is given to words other than topic words. To reduce the repetition of words in the summarized sentence, a flat score is also given to each reappearing noun and verb.

2.2. Linguistic score

The linguistic score $L(v_m|\dots v_{m-1})$ indicates the appropriateness of the word strings in a summarized sentence and is measured by a trigram probability $P(v_m|\, v_{m-2}v_{m-1})$ [5]. In contrast with the word significance score which focuses on topic words, the linguistic score is helpful to extract other words necessary to construct a readable sentence.

2.3. Confidence score

The confidence score C (v_m) is incorporated to weight acoustically as well as linguistically reliable hypotheses [6]. 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 the confidence measure[8].

2.4. Word concatenation score

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

The word concatenation score between v_{m-1} and v_m is determined by a word dependency between them. That is, word concatenation in a summarized sentence is restricted by the dependency structure in an original sentence as exemplified in Fig.1.

Since the dependency structure between words is usually ambiguous, the word dependency is given by a probability that one word is modified by others based on the SDCFG as follows. The SDCFG for English consists of the following rules including both "forward dependency structure" in which a preceding word modifies a succeeding one and "backward dependency structure" in which a succeeding word modifies a preceding one.

$\alpha \rightarrow \beta\alpha$ `(forward dependecy structure)`
$\alpha \rightarrow \alpha\beta$ `(backward dependency structure)`
$\alpha \rightarrow w$

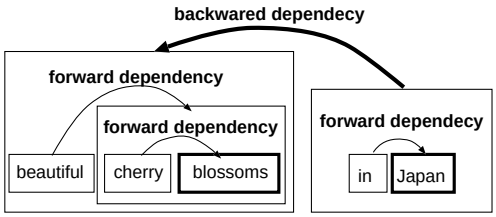

Figure 1: *An example of dependency structure.*

where α,β are nonterminal symbols and w is a terminal symbol (word).

The word dependency probability is a posterior probability estimated using the Inside-Outside probabilities. Suppose a sentence consists of L words, $w_1, \dots, w_L$. The probability that w_m and w_l has a dependency structure is calculated as a sum of the probabilities of the following sequence when a sentence is derived from the initial symbol S; 1) the rule of $\alpha \rightarrow \beta\alpha$ is applied, 2) $w_i \dots w_k$ is derived from β, 3) w_m is derived from β, 4) $w_{k+1} \dots w_j$ is derived from α and 5) w_l is derived from α. The probability of applying the rule of $\alpha \rightarrow \alpha\beta$ is also added. Using the dependency probabilities $d(w_m, w_l, i, k, j)$, the word concatenation score between w_m and w_n is calculated by

$$Tr(w_m, w_n) = \log \sum_{i=1}^{m} \sum_{k=m}^{n-1} \sum_{j=n}^{L} \sum_{l=n}^{j} d(w_m, w_l, i, k, j). \quad (2)$$

This score is defined as a logarithmic value of the sum of the dependency probabilities between w_m and each of $w_n \dots w_l$. Figure 2 illustrates the principle of the word concatenation score.

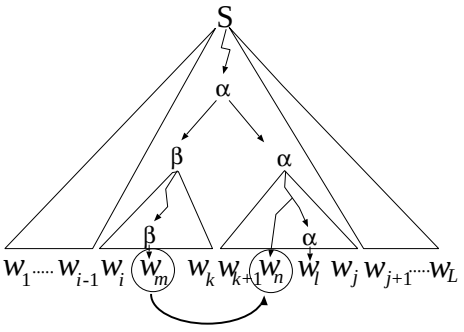

Figure 2: *The principle of word concatenation based on a dependency structure.*

We use the SDCFG to estimate the dependency structure of the original sentence. In our SDCFG, only the number of non-terminal symbols is determined and all combinations of rules are applied recursively. The non-terminal symbol has no specific function such as a noun phrase. All the probabilities of rules are stochastically estimated

based on data. Probabilities for frequently used rules become bigger, and those for rarely used rules become smaller. Even though transcription results by a speech recognizer are ill-formed, the dependency structure can be robustly estimated by our SDCFG.

3. SUMMARIZATION OF MULTIPLE UTTERANCES

Our proposed automatic speech summarization technique for each sentence has recently been extended to summarize a set of multiple utterances (sentences) [7]. A set of words maximizing the summarization score is extracted from multiple utterances under some restrictions applied at the sentence boundaries. These restrictions realizes the summarization of multiple utterances by handling them as a single long utterance. This results in preserving more words inside information rich utterances and shortening or even completely deleting less informative ones. However, the amount of calculation for selecting the best combination among all possible combinations of words in the multiple utterances increases as the number of words in the original utterances increases. In order to alleviate this problem, we have proposed a new method in which each utterance is summarized according to all possible summarization ratio and then the best combination of summarized sentences for each utterance is determined according to a target compression ratio using a two-level DP technique. Figure 3 illustrates the two-level DP technique for summarizing multiple utterances.

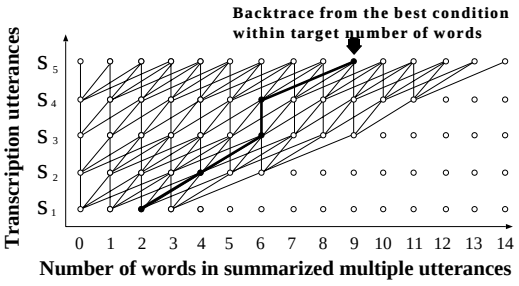

Figure 3: *An example of DP process for summarization of multiple utterances.*

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 which approximately expresses all possible correct summarization including subjective variations. A summarization accuracy of automatic summarization is calculated using the word network[7]. The word string that is the most similar to the automatic summarization result extracted from the word network is considered as a correct answer for the automatic summarization. The accuracy, comparing the summarized sentence with the set of words extracted from the network, is used as a measure of linguistic correctness and maintenance of original meanings of the utterance (summarization accuracy).

4.2. Evaluation data

English TV broadcast news utterances (CNN news) recorded in 1996 given by NIST as a test set of Topic Detection and Tracking (TDT) were tagged by Brilltagger[10] and used to evaluate our proposed method. Five news articles consisting of 25 utterances in average were transcribed by JANUS[8] speech recognition system. The multiple utterance summarization was performed for each of the five news articles at 40% and 70% summarization ratio. 50 utterances arbitrarily chosen from the five news articles were used for the sentence by sentence summarization with the summarization ratios of 40% and 70%. Mean word recognition accuracies of the utterances used for the multiple utterance summarization and those for sentence by sentence summarization were 81% and 80%, respectively.

4.3. Training data for summarization models

A word significance model, a trigram language model and SDCFG were constructed using roughly 35M words (10681 sentences) of the Wall Street Journal corpus and the Brown corpus in Penn Treebank[9].

4.4. Evaluation results

Manual transcription (TRS) and automatic transcription (REC) were both summarized. Table 1 shows an example of evaluation results based on a manual summarization word network. Figure 4 shows summarization accuracies of utterance summarization at 40% and 70% summarization ratio and Fig. 5 shows those of summarizing articles having multiple sentences at 40% and 70% summarization ratio. In these figures, I, L, C and T indicate that the word significance score, the linguistic score, the confidence score and the word concatenation score are used, respectively.

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

The averaged summarization accuracies of each manual summarizations (SUB) was considered to be the upper limit of the automatic summarization accuracy. To ensure that our method is sound, we made randomly generated summarized sentences (RDM) according to the summarization ratio and compared them with those obtained by our proposed methods.

These results show that our proposed automatic speech summarization technique is significantly more effective than RDM. Using the word concatenation score (I_L_T, $I\_L\_C\_T$) also reduce the meaning alteration compared to not using it (I_L, I_L_C). The result obtained when using the word confidence score ($I\_L\_C\_T$) compared with those not using it (I_L_T) shows that the summarization accuracy is improved by the confidence score.

5. CONCLUSIONS

Each utterance and a whole news consisting of multiple utterances of English broadcast news speech were summarized by our automatic speech summarization method based on the follwing scores: word significance score, linguistic likelihood, word confidence measure and word concatenation

Table 1: *An example of evaluation results based on a manual summarization word network.* upper: a set of words extracted from the correct summarization network which is the most similar to automatic summarization, lower: automatic summarization of recognition result.

Recognition result	VICE PRESIDENT AL GORE SAYS THE GOVERNMENT HAS A PLAN TO AVOID THE INEVITABLE PROSPECT OF INCREASED AIRPLANE CRASHES AND FATALITY IS
70% summarization	VICE PRESIDENT AL GORE SAYS THE GOVERNMENT HAS A PLAN TO AVOID THE INCREASED AIRPLANE CRASHES
	VICE PRESIDENT AL GORE SAYS THE GOVERNMENT HAS A PLAN TO AVOID <DEL> INCREASED AIRPLANE CRASHES
40% summarization	<INS> THE GOVERNMENT HAS A PLAN TO AVOID THE INCREASED AIRPLANE CRASHES
	GORE THE GOVERNMENT HAS A PLAN TO AVOID THE INCREASED AIRPLANE CRASHES

_: recognition error, <>: substitution, <INS>: insertion, <DEL>: deletion

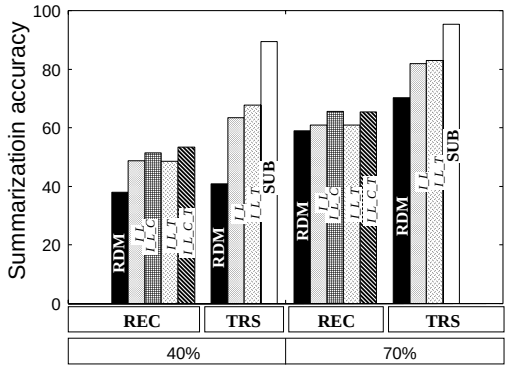

Figure 4: *Each utterance summarizations at 40% and 70% summarization ratio.*

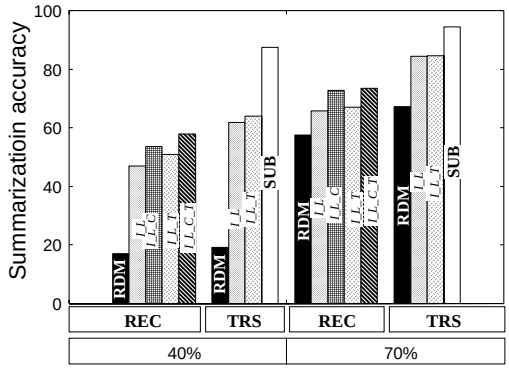

Figure 5: *Article summarizations at 30% and 70%summarization ratio.*

probability. Experimental results show that our proposed method can effectively extract relatively important information and remove redundant and irrelevant information from Englsih news speech as well as Japanese one.

In contrast with the confidence score which has been incorporated into the summarization score to exclude word errors by a recognizer, the liguistic score is effective to reduce out of context words extraction both from recognition errors and human disfluencies. In summarizing Japanese news speech, the confidence measure could improve the summarizing performance by excluding incontext word errors. In English, the confidence measure can not only exclude word errors but also help extracting clearly pronouced important words. This results in the use of the confidence measure causing a lager increace in the summarization accuracy for English rather than Japanese.

6. ACKNOWLEDGMENT

The authors would like to thank Dr. Yoshi Gotoh (Sheffield University) for an arrangement of generating the correct answers for automatic summarization.

REFERENCES

[1] T.Imai et al., "Progressive 2-pass Decoder for Real-Time Broadcast News Captioning," Proc. ICSLP2000, vol.I, pp.246-249, Beijing(2000).

[2] Z.Klaus, "Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains," Proc. SIGIR2001, New Orleans(2001).

[3] S.Furui et al., "Toward the Realization of Spontaneous Speech Recognition -Introduction of a Japanese Priority Program and Preliminary Results-," Proc. ICSLP2000, vol.III, pp.518-521, Beijing(2000).

[4] R.Valenza et al., "Summarization of Spoken Audio through Information Extraction," Proc. ESCA Workshop on Accessing Information in Spoken Audio, pp.111-116, Cambridge(1999).

[5] C.Hori et al., "Automatic Speech Summarization Based on Word Significance and Linguistic Likelihood," Proc. ICASSP2000, vol.III, pp.1579-1582, Istanbul(2000).

[6] C.Hori et al., "Improvements in Automatic Speech Summarization and Evaluation Methods," Proc. ICSLP2000, vol.IV, pp.326-329, Beijing(2000).

[7] C.Hori et al., "Advances in Automatic Speech Summarization," Proc. EUROSPEECH2001, vol.III, pp.1771-1774, Aalborg(2001).

[8] A.Waibel et al., "Advances in Meeting Recognition," Proc. HLT2001, pp.11-13, San Diego(2001)

[9] http://www.cis.upenn.edu/~treebank/

[10] http://www.cs.jhu.edu/~brill/