T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	Evaluation Method for Automatic Speech Summarization
Authors	Chiori Hori, Takaaki Hori, Sadaoki Furui
Citation	EUROSPEECH2003,, Vol. , No. , pp. 2825-2828
Pub. date	2003, 9

Powered by T2R2 (Science Tokyo Research Repository)

Evaluation Method for Automatic Speech Summarization

Chiori Hori and Takaaki Hori

NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
{chiori.hori}@cslab.kecl.ntt.co.jp

Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology
furui@cs.titech.ac.jp

Abstract

We have proposed an automatic speech summarization approach that extracts words from transcription results obtained by automatic speech recognition (ASR) systems. To numerically evaluate this approach, the automatic summarization results are compared with manual summarization generated by humans through word extraction. We have proposed three metrics, *weighted word precision*, *word strings precision* and *summarization accuracy* (SumACCY), based on a word network created by merging manual summarization results. In this paper, we propose a new metric for automatic summarization results, *weighted summarization accuracy* (**WSumACCY**). This accuracy is weighted by the posterior probability of the manual summaries in the network to give the reliability of each answer extracted from the network. We clarify the goal of each metric and use these metrics to provide automatic evaluation results of the summarized speech. To compare the performance of each evaluation metric, correlations between the evaluation results using these metrics and subjective evaluation by hand are measured. It is confirmed that **WSumACCY** is an effective and robust measure for automatic summarization.

1. Introduction

To validate the efficiency of new approaches, automatic evaluation metrics are needed to evaluate automatically processed sentences produced by automatic summarization and machine translation. Sentences automatically processed can be compared to sentences manually processed by humans. The similarity between automatically and manually processed sentences can be used for evaluation metrics. However, the manual results for summarization and translation vary among humans, and correct answers for automatic results are not unified. In consideration of this subjective variation, we have proposed three metrics for automatic summarization results, *weighted word precision*, *word string precision* [1] and *summarization accuracy* (**SumACCY**), based on a word network made by merging manual summarization results [2]. In the field of machine translation, an automatic evaluation metric based on n-gram precision, **BLEU**, was proposed [3].

This paper describes the goals of these automatic evaluation methods and the differences among the metrics that have already been proposed so far. In addition, to give a reliability that reflects the majority of the humans' selections, our **SumACCY** is weighted by a posterior probability of the manual summarization network. To compare these metrics, Japanese news broadcasts [1] is automatically recognized and summarized, and then the summarized results are evaluated by these metrics.

2. Automatic Summarization Method

We have proposed a sentence compaction-based statistical speech summarization technique. In this approach, a set of words maximizing a summarization score is extracted from automatically transcribed speech and then concatenated to create a summary [4] [5]. The word extraction is performed according to a target compression ratio. The summarization score indicates the appropriateness of summarization. This score consists 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, and a word concatenation score T. The word concatenation score indicates a word concatenation probability determined by a dependency structure in the original sentence given by stochastic dependency context-free grammar, **SDCFG**. The total score is maximized using a dynamic programming (DP) technique. 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 K words, $W = w_1, w_2, \ldots, w_K$, the summarization is performed by extracting a set of $M (M < K)$ words, $V = v_1, v_2, \ldots, v_M$, which maximizes the summarization score given by eq. (1).

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

where λ_I, λ_C and λ_T are weighting factors for balancing among L, I, C and T. Details of the scores are represented in our previous work [5][7]. The proposed technique can be applied to each sentence utterance as well as entire speech consisting of multiple utterances. This technique has been applied to Japanese as well as English, and its effectiveness has been confirmed [6] [7].

3. Evaluation Metrics

The process of summarizing speech involves excluding recognition errors and maintaining important information. In addition, the summarized sentences should be meaningful. The automatic summarization results are evaluated from the viewpoints of excluding recognition errors, extracting important information, and maintaining original meaning. In the first step of evaluating automatic summarization, humans subjectively evaluate the appropriateness of automatic summarization. This type of subjective evaluation is not only expensive but also insufficient for precisely comparing the efficiencies of different automatic summarization approaches. Therefore, it is necessary to adopt

automatic evaluation metrics to numerically validate the efficiency of automatic summarization.

3.1. Word accuracy

The most straight-forward approach to automatic evaluation is to directly compare results with goals based on similarity. In the field of speech recognition, automatic recognition results are compared with manual transcription results by humans. The conventional metric for speech recognition is a recognition accuracy calculated based on word accuracy as follows:

$$\mathrm{WACC} = \frac{Len - (Sub + Ins + Del)}{Len} \times 100[\%]. \qquad (2)$$

where Sub, Ins, Del and Len are the numbers of substitutions, insertions, deletions, and words in the manual transcription, respectively.

When the correct answer for the recognition result can be set as only one sentence, word accuracy is the simplest and most efficient metric. Although word accuracy cannot directly evaluate the meanings of sentences, higher accuracy indicates that more original information is preserved.

On the other hand, manual summarization sentences by humans can also be set as goals of automatic summarization. To generate manual summarization results, speech is manually transcribed by humans and then summarized through word extraction by humans. However, manual summarization results vary among humans. If we could collect all possible manual summarization sentences, the one that was the most similar to the automatic results could be set as the correct answer. As a result of unifying multiple answers to select the correct answer, the automatic result could be compared with the correct answer by using word accuracy.

However, in real situations, the number of manually summarized sentences that could be collected is limited. The word accuracy obtained by comparing dissimilar sentences does not provide an efficient metric. The problem of how to deal with multiple answers that vary among subjects has been addressed in the initial stages of our investigation of automatic evaluation metrics.

3.2. Precision

To test the similarity between two sentences, the overlap of components such as words and the word strings between them is used as a measure. Even if these sentences have different lengths, this measure is an efficient way to evaluate the similarity of word occurrence. The number of components overlapping between the sentences is calculated by using the ***precision*** of components. Word precision is calculated by eq. (2) without insertion errors, Ins. Therefore, word accuracy that includes insertion errors is more precise than word precision that uses a fixed pair of an answer and an automatic result.

The precision of components in each sentence can be applied to evaluate an automatic summarization result when using multiple answers. The number of components in the automatic summarization results that overlap with the components in the multiple answers is the ***precision*** of components. Note that a word occurring in a different location in the original sentence is considered to be a different word even though it is the same word as one in the result.

3.3. Word string precision

Precision based on each word can evaluate similarity between sentences at the level of isolated words. However, meanings are generated by combinations of words. Even if more important words are extracted in automatic summarization, word strings comprised of these words cannot always maintain their original meanings. To evaluate linguistic precision and the maintenance of the original meanings of an utterance, we proposed ***word string precision*** [1]. In this case, component units are word strings of various lengths. This ***word string precision*** is n-gram precision. The extraction ratio p_n of each word string consisting of n words in a summarized sentence $V = v_1, v_2, \ldots, v_M$ is given by

$$p_n = \frac{\sum_{m=n}^{M} \delta(v_{m-n+1}, \ldots, v_{m-1}, v_m)}{M - n + 1}, \qquad (3)$$

where

$$\delta(u_n) = \begin{cases} 1 & \text{if } u_n \in U_n \\ 0 & \text{if } u_n \notin U_n \end{cases},$$

u_n : each word string consisting of n words
U_n : a set of word strings consisting of n words in all manual summarizations.

When n is 1, p_n indicates the precision of each word, and when n is the same length as a summarized sentence ($n = M$), p_n indicates the precision of the summarized sentence itself.

3.4. BLEU

Recently, **BLEU** was proposed as an automatic evaluation metric for machine translation based on the precision of word strings (n-gram) [3]. In this method, the n-gram precision is calculated independently of the location of words in a sentence. Each n-gram in an automatic summarization that overlaps at least once in any manual translation result is counted. When an n-gram in an automatically processed sentence is more frequent than that occurring in any manual result, the frequency of the word string is limited to the max frequency in a sentence from the manual results. Additionally, this precision is modified in terms of the length of n-gram and the length of the sentence. **BLEU** is given by eq. (5).

$$\mathrm{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} \nu_n \log p_n\right), \qquad (5)$$

where

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}. \qquad (6)$$

In this equation, p_n is n-gram precision, and c and r are the lengths of the sentences automatically processed and the effective answer, respectively. N is the length of n-gram, and ν_n is given as $1/N$. They reported that this metric closely reflected subjective evaluations for machine translation by humans.

In machine translation, correct answers generated by humans vary as well as human summarizations. The order of words in a summarization generated through word extraction is restricted by that of the original sentence. However, word selection, order of words, and lengths of sentences in machine translation are not restricted explicitly. The differences among manual translations are bigger than those among manual summarization. The precision of components such as n-grams that

overlap with components in multiple answers is very useful for measuring similarity between sentences that are not so homogeneous. However, the similarity of the divided components can only measure a local accomplishment when the n of an n-gram is smaller than the length of sentences automatically generated. To evaluate a global accomplishment as a whole sentence, exact answers should be obtained.

3.5. Summarization accuracy: SumACCY

A manual summarization that is most similar to an automatic summarization result is considered to be an exact answer for this automatic summarization result. To set an exact answers for automatic summarization and evaluate global accomplishment, the word accuracy of the automatic summarization is calculated by using this exact answer. This word accuracy of the automatic summarization result relative to the exact answer is defined as the *summarization accuracy*.

However, the number of manual summarizations that are actually collected is limited. Accuracy is not efficient for measuring similarity between sentences that are not so homogeneous. To cover all possible correct answers for summarization, we have proposed **SumACCY**, which is calculated by using the word network generated through merging manual summarizations [6].

In our summarization process, words and the order of words in a summarized sentence are restricted by those in the original sentence. Therefore, manual summarizations are homogeneous with small differences. The homogeneous multiple answers can be combined into a network that represents concatenation structures between the divided components in the multiple answers. Sentences extracted from this network consist of words and word concatenations that occur at least once in manual summarization results. Quasi-correct answers are represented in the manual summarization network.

Table 1: Example of manual summarization through word extraction.

SUB	The	beautiful	cherry	blossoms	in	Japan	bloom	in	spring
A	The		cherry	blossoms	in	Japan			
B		beautiful	cherry	blossoms	in	Japan			
C		beautiful	cherry	blossoms				in	spring
D			cherry	blossoms			bloom	in	spring
E		beautiful	cherry				bloom	in	spring

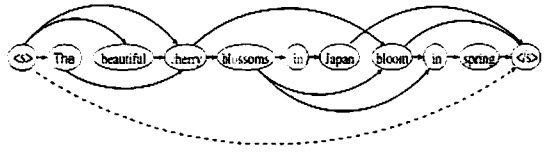


Figure 1: Word network made by merging manual summarization results.

"*The beautiful cherry blossoms in Japan bloom in spring.*" is assumed to be manually summarized as shown in Table 1. In this example, five words are extracted from nine words. The summarization ratio is 56%. Variations of manual summarization results in Table 1 are merged into a word network as shown in Fig. 1. <s> and </s> are the beginning and ending symbols of a sentence. Although "*Cherry blossoms in Japan bloom.*" extracted from the network is not included in the manual answers in Table 1, this sentence is considered to be one of the correct answers.

To set an exact answer for an automatic summarization result, the sentence that is most similar to the automatic summarization result is extracted from the network. *Summarization accuracy* of the automatic summarization result is calculated in comparison to the extracted sentence. If there exists a direct path from the sentence beginning with <s> to the sentence ending with </s> in the word network, the summarization accuracy for that sentence is 100% (no error).

3.6. Weighted SumACCY: WSumACCY

All possible sets of words extracted from the network of manually summarized sentences are equally set as exact answers. However, the set of words containing more word strings that are selected by more humans would presumably be better and more reliable answers. To obtain a reliability that reflects the majority of the humans' selections, the summarization accuracy is weighted by a posterior probability of the manual summarization network. The reliability of the extracted sentence from the network is defined as a production of the ratio of the number of subjects who select each word to the total number of subjects. The weighted summarization accuracy is given by eq. (7).

$$\text{WSumACCY} = \left(\prod_{m=2}^{\hat{M}} \frac{C(\hat{v}_{m-1}, \hat{v}_m)}{H} \right)^{\frac{1}{\hat{M}-1}} \times \text{SumACCY}, \qquad (7)$$

where $\hat{v}_m$ is the m-th word in the sentence extracted from the network as the exact answer. $\hat{M}$ represents the total number of words in the exact answer and the automatic summarization result. $C(v, w)$ indicates how many subjects selected the word connection of v and w. Here, word connection means an arc in the manual summarization network. H is the number of subjects.

4. Evaluation Experiments

4.1. Summarization experiments

Japanese TV news broadcasts aired in 1996 were automatically recognized and summarized sentence by sentence [4]. The set consisted of 50 utterances by a female anchor. The out-of-vocabulary (OOV) rate for the 20k word vocabulary was 2.5%, and the perplexity for the test set was 54.5. Fifty 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, the ratio of the number of words in the summarized sentences to that in the original sentences, was set to 40%.

4.2. Evaluation conditions

Summarization was performed using the possible combination of scores I, L, C and T. We selected nine automatic summarization results with various *summarization accuracies* from 40% to 70% and a manual summarization result (SUB) as test sets. These 10 types of summarization results for each utterance were evaluated by 10 humans. The human subjects read these summarization results and rated each summarization from 1 (incorrect) to 5 (best). In addition, these summarization results were also evaluated by using the objective metrics **SumACCY**, **WSumACCY** and **BLEU**. The scores were averaged over 50 utterances. To numerically evaluate the results

using the objective metrics, 25 humans generated manual summarization through word extraction. These manual summarization results were set as target sets of automatis summarization results. These manual summarization results were merged into a network. Note that a set of 24 manual summaries made by other subjects is used as the target of SUB.

4.3. Evaluation results

The set of 25 manual summaries was used for evaluating the automatic summaries by using the objective metrics while taking the subjective variations into account. Evaluation results by **SumACCY**, **BLEU** and **WSumACCY** are shown in Figs. 2, 3 and 4, respectively. Both **SumACCY** and **BLEU** increase as a function of the variation in manual summarization. On the other hand, **WSumACCY** is robust against the variation of manual summarization. The evaluation results by this metric are consistent independent of the variation in the taget sets, if the target set includes three manual summaries or more.

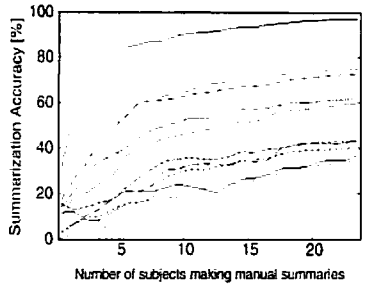


Figure 2: Variation of **SumACCY** depends on the number of subjects making manual summarizations.

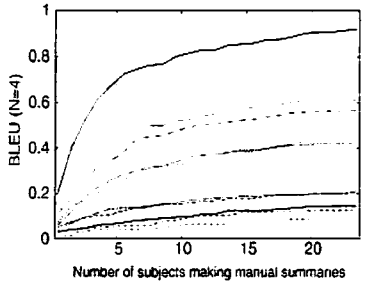


Figure 3: Variation of **BLEU** depends on the number of subjects making manual summarizations.

The correlation coefficients between these metrics and human judgments are shown in Fig. 5. In comparison with **SumACCY** and **BLEU**, **WSumACCY** can effectively reflect subjective evaluation results independent of the variation of answer sets. **WSumACCY** is a simple but robust and effective evaluation metric.

5. Conclusion

This paper has proposed **WSumACCY**, a new metric to evaluate the appropriateness of automatic summarization. Summarization accuracy based on a word network of manual summaries, **SumACCY**, was modified by incorporating a reliability of manual summaries. **WSumACCY** is a modification of **SumACCY** weighted by the posterior probability of manual

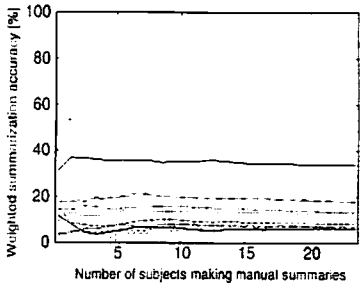


Figure 4: Variation of **WSumACCY** depends on the number of subjects making manual summarizations.

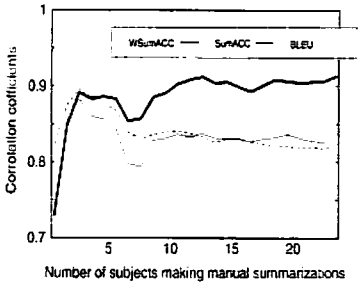


Figure 5: Coefficients of correlation between subjective judgments of 10 humans and objective evaluation results depend on the number of subjects making manual summarizations.

summaries in the network. The automatic summarization results for 50 utterances in Japanese TV news broadcasts have been evaluated by **SumACCY**, **WSumACCY** and **BLEU**. In comparison with **SumACCY** and **BLEU**, **WSumACCY** effectively reflects the subjective judgments. In addition, this metric is consistent independent of the variations in manual summarization. Evaluation results show that **WSumACCY** is a simple but robust and effective evaluation metric for automatic summarization.

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 et al., *Improvements in Automatic Speech Summarization and Evaluation Methods*, *Proc. ICSLP*, Beijing, China, Vol. 4, pp. 326-329, 2000.

[2] C. Hori et al., *Advances in Automatic Speech Summarization*, *Proc. Eurospeech*, Aalborg, Denmark, Vol. III, pp. 1771-1774, 2001.

[3] K. Papineni et al., *BLEU: a Method for Automatic Evaluation of Machine Translation*, *Proc. ACL*, Philadelphia, USA, 2002.

[4] C. Hori et al., *Automatic Speech Summarization based on Word Significance and Linguistic Likelihood*, *Proc. ICASSP*, Istanbul, Turkey, Vol. 3, pp. 1579-1582, 2000.

[5] C. Hori et al., *A New Approach to Automatic Speech Summarization*, To appear in *the IEEE Transactions on Multimedia*, 2003.

[6] C. Hori et al., *Automatic Summarization of English Broadcast News speech*, *Proc. HLT*, San Diego, U.S.A., 2002.

[7] C. Hori et al., *A Statistical Approach for Automatic Speech Summarization*, "Special Issue on Unstructured Information management" in *the EURASIP Journal on Applied Signal Processing*, Vol. 2, pp. 128-139, 2003.