Evaluation Methods 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 human subjects 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 human judgment 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 for automatic summarization and machine translation, automatic evaluation metrics to evaluate automatically processed sentences are needed. Sentences automatically processed can be compared to sentences manually processed by humans. The similarity between automatically and manually processed sentences can be used as an evaluation metric. However, the manual results for summarization and translation vary among humans, and correct answers for automatic results cannot be 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. In addition, to give a reliability that reflects the majority of the humans' selections, the **SumACCY** is weighted by a posterior probability of the manual summarization network to create a new metric. 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 spoken utterances, and its effectiveness has been confirmed [6] [7].

3. Evaluation Metrics

The automatic summarization results need to be evaluated from the viewpoints of excluding recognition errors, extracting important information, and maintaining original meanings. The simplest and probably the most ideal way of evaluating automatic summarization could be letting human subjects evaluate the appropriateness of automatic summarization. This type of 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.

The objective evaluation can be realized by comparing sentences automatically processed and sentences manually sum-

marized by humans. To generate target summaries, speech is manually transcribed and then manually summarized through word extraction. The similarities between the targets and the results automatically processed provide metrics indicating how much the task is accomplished. The similarity that can much reflect subjective judgments is a better metric. Since the order of words and the length of summarization results are restricted by the original sentences in our summarization approach, *word accuracy* is the straight-forward approach to measure similarities between the target and automatic summarizations.

3.1. Word accuracy

In the field of speech recognition, automatic recognition results are compared with manual transcription results. The conventional metric for speech recognition is a recognition accuracy calculated based on word accuracy as follows:

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

where Sub, Ins, Del and Len are the numbers of substitutions, insertions, deletions, and words in the manual transcription, respectively. Although *word accuracy* cannot directly evaluate the meanings of sentences, higher accuracy indicates that more original information is preserved. When the target for the automatically processed sentences can be set as only one sentence, word accuracy is the simplest and the most efficient metric.

However, there usually exist multiple targets for each automatic summarization sentence caused by the variation of manual summarization among humans. Therefore, it is not easy to apply the *word accuracy* to evaluate automatic summarization results. The subjective variation brings the following two problems:

1. how to consider all possible correct answers in the manual summarization, and
2. how to measure the similarity between the evaluation sentence and multiple manual summaries.

If we could collect all possible manual summarization sentences, the one that is most similar to the automatic result could be chosen as the correct answer and used for the evaluation. However, in real situations, the number of manually summarized sentences that could be collected is limited. The coverage of real answers in the collected manual summaries is unknown. When the coverage is low, the summarization results are compared with inappropriate targets, and the *word accuracy* obtained by such comparison does not provide an efficient measure.

3.2. Word string precision

One of the solutions to cope with the coverage problem is to use local matching of words or word strings with all the manual summaries instead of using the sentence-level matching. The similarity can be measured by counting the *precision*, the number of word/word-string components overlapping between the sentences.

Even if there are multiple targets for an automatic summarization sentence, the *precision* of components in each sentence can be used to evaluate an automatic summarization result. *Precision* is very efficient to evaluate the similarity of word occurrence between sentences with different lengths. 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. Word precision is calculated using eq. (2) simply neglecting the insertion errors, Ins.

Since meanings are basically conveyed by word strings rather than single words, we proposed ***word string precision*** [1] to evaluate linguistic precision and the maintenance of the original meanings of an utterance. In this method, word strings of various lengths, that is n-grams, are used as components for measuring the 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

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

where

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

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 corresponds to 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.

3.3. BLEU

Recently, **BLEU** was proposed as an automatic evaluation metric for machine translation based on the precision of word strings (n-grams) [3]. In this method, n-gram precision is calculated independently of the location of words in the sentence. The number of each n-word sequence in an automatic summarization that occurs at least once in any manual translation result is counted. When an n-word sequence 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 maximum frequency in a sentence in the manual results. Additionally, this precision is modified using the length of the n-word sequence and the length of the sentence. Consequently, **BLEU** is given by eq. (5).

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

where

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

In this equation, p_n is the n-gram precision, and c and r are the lengths of the sentence automatically processed and the target/correct sentence, respectively. N is the length of the n-gram, and ν_n is given by $1/N$. It has been reported that this metric is closely related to subjective evaluation of machine translation.

In machine translation, correct answers generated by humans vary in the same way as human summarization. Since word selection, order of words, and lengths of sentences are not explicitly restricted in machine translation, variation of manual translations are usually larger than manual summarizations generated by our method. The precision of components such as n-grams that overlap with components in multiple answers is very useful for measuring similarity between sentences with

large variation. One of the problems of the similarity of the n-word sequences is that it can only measure a local matching, especially when the n of an n-gram is small.

3.4. Summarization accuracy: SumACCY

In order to measure a global similarity and cope with the coverage problem at the same time, *summarization accuracy*: **SumACCY** has been proposed. In this method, a manual summary which is most similar to an automatic summarization result is considered to be a target answer, and word accuracy of the automatic summarization in comparison with the target answer is calculated.

To cover all possible correct answers for summarization using a limited number of manual summaries, all the manual summaries are merged to create a word network. A word sequence in the network, which is closest to the evaluation word sequence, is extracted and used for measuring the similarity based on the word accuracy [6].

Since our summarization process is based on sentence compaction, words cannot be replaced by other words, and the order of words cannot be changed. Therefore, multiple summaries can be easily combined into a network that represents the variations. Each sentence that could be extracted from the network consists of words and word concatenations occurring at least once in the manual summarization results. The network made by the manual summaries can be considered to represent all possible variation of correct summaries.

Table 1: An example of manual summarization by sentence compaction.

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					spring
D			cherry	blossoms			bloom	in	spring
E		beautiful	cherry				bloom	in	spring

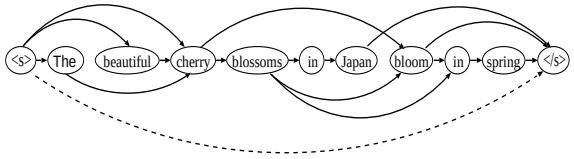

Figure 1: Word network made by merging manual summarization results.

"*The beautiful cherry blossoms in Japan bloom in spring.*" is supposed to be manually summarized as shown in Table 1. In this example, five words are extracted from nine words. Therefore, 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 beginning and ending symbols of a sentence. Although "*Cherry blossoms in Japan bloom.*" is not included in the manual answers in Table 1, this sentence that could be extracted from the network is considered to be one of the correct answers.

As a target answer for an automatic summarization result, a sentence that is most similar to the automatic summarization result is extracted from the network. *Summarization accuracy* of the automatic summarization result is calculated by comparing it with the extracted sentence.

3.5. Weighted SumACCY: WSumACCY

In the **SumACCY**, all possible sets of words extracted from the network of manually summarized sentences are equally used as target answers. However, the set of words containing word strings that are selected by many humans would presumably be better and more reliable answers. To obtain reliability that reflects the majority of the humans selections, the summarization accuracy is modified to be weighted by a posterior probability based on the manual summarization network. Reliability of the extracted sentence from the network is defined as a product of the ratios of the number of subjects who select each word to the total number of subjects. The weighted summarization accuracy is given by eq. (7).

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

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

3.6. Summarization experiments

Japanese TV news broadcasts aired in 1996 were automatically recognized and summarized sentence by sentence [4]. They consisted of 50 utterances by a female announcer. The out-of-vocabulary (OOV) rate for the 20k word vocabulary was 2.5%, and the test-set perplexity 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%.

3.7. Evaluation conditions

Summarization was performed using the possible combination of scores I, L, C and T. Nine automatic summarization results with various *summarization accuracies* from 40% to 70% and a manual summarization result (SUB) were selected as a test set. These 10 types of summarization results for each utterance were evaluated by 10 human subjects. The human subjects read these summarization results and rated each summarization from 1 (incorrect) to 5 (best). 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 a target set of automatic summarization results, and merged into a network. Note that a set of 24 manual summaries made by other subjects was used as the target of SUB.

3.8. 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 of the 10 types of summarization results by **SumACCY**, **BLEU** and

WSumACCY are shown in Figs. 2, 3 and 4, respectively. The correlation coefficients between human judgments and evaluation results by **SumACCY**, **BLEU** and **WSumACCY** are shown in Fig. 5.

Figs 2 and 3 show that the values of **SumACCY** and **BLEU** increase as the number of subjects making manual summaries increases, that is, as the variation in the manual summarization increases. On the other hand, the value of **WSumACCY** saturates when the number of subjects becomes larger than 10. This is because **WSumACCY** is weighted for the words selected by many subjects and de-weighted for the words selected by a few subjects. Therefore, this metric is robust against the variation of manual summarization especially when the selected words are concentrated.

When there exists a manual summary that is largely isolated from others, the correlation coefficients of all metrics decrease as shown in Fig. 5. However, the value for **WSumACCY** quickly recovers and it becomes more stable than other measures as the number of subjects making manual summarization increases. The correlation coefficients for **SumACCY** and **BLEU** are relatively unstable, since words selected by a few subjects are equally weighted as correct answers.

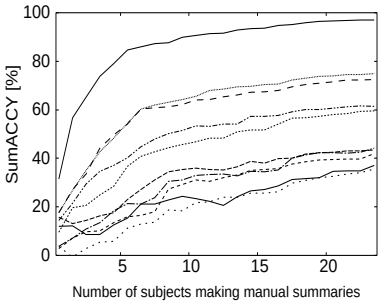

Figure 2: Variation of **SumACCY** depending on the number of subjects making manual summarizations.

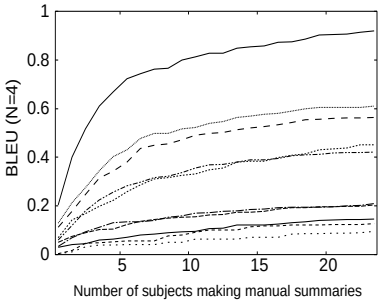

Figure 3: Variation of **BLEU** depending on the number of subjects making manual summarizations.

WSumACCY is a simple but robust and effective evaluation metric.

4. Conclusion

This paper has proposed a new metric, **WSumACCY**, 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 to create the new metric. Specifically, **WSumACCY** is weighted by the posterior probability of manual summaries in the network. 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 subjective judgments. In addition, this metric is relatively independent of the variations in manual summarization. Evaluation results show that **WSumACCY** is a simple but robust and effective evaluation metric for automatic summarization.

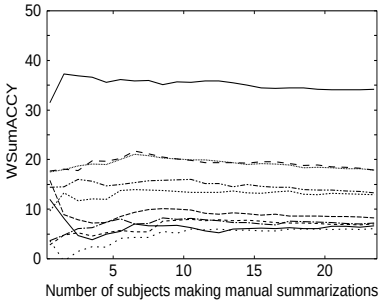

Figure 4: Variation of **WSumACCY** depending on the number of subjects making manual summarizations.

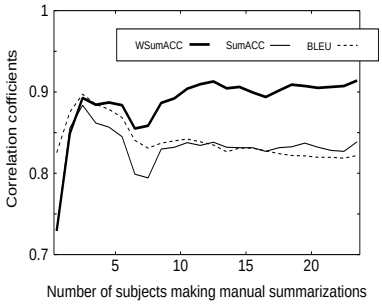

Figure 5: Correlation coefficients between subjective judgments of 10 humans and objective evaluation results depending on the number of subjects making manual summarizations.

5. Acknowledgment

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

6. 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.