Mis-recognized Utterance Detection Using Multiple Language Models Generated by Clustered Sentences

Katsuhisa FUJINAGA[‡], Hiroaki KOKUBO[†], Hirofumi YAMAMOTO[†],
Genichiro KIKUI[†], Hiroshi SHIMODAIRA[‡]

[†] ATR Spoken Language Translation Research Laboratories
[‡] School of Information Science, Japan Advanced Institute of Science and Technology
kfujina@jaist.ac.jp, {hiroaki.kokubo, hirofumi.yamamoto}@atr.co.jp
genichiro.kikui@atr.co.jp, sim@jaist.ac.jp

Abstract

This paper proposes a new method of detecting mis-recognized utterances based on a ROVER-like voting scheme. Although the ROVER approach is effective in improving recognition accuracy, it has two serious problems from a practical point of view: 1) it is difficult to construct multiple automatic speech recognition (ASR) systems, 2) the computational cost increase according to the number of ASR systems. To overcome these problems, a new method is proposed where only a single acoustic engine is employed but multiple language models (LMs) consisting of a baseline (main) LM and sub LMs are used. The sub LMs are generated by clustered sentences and used to re-score the word lattice given by the main LM. As a result, the computational cost is greatly reduced. Through experiments, the proposed method resulted in 18-point higher precision with 10% loss of recall when compared with the baseline, and 22-point higher precision with 20% loss of recall.

1. Introduction

In the recent progress of ASR, word accuracy is widely used to measure system performance. Word accuracy is a good measure for speech to text systems, such as dictation. However, it is not a good measure for speech translation systems, since only one mis-recognized word in an utterance sometimes gives incorrect (or negative) information and word accuracy does not measure whether all of the recognized words are correct. In speech translation systems, mis-recognized utterances must be discarded to avoid incorrect translation, and hence the performance of ASR is measured on the basis of utterance accuracy instead of word accuracy.

To detect mis-recognized utterances, we focus on ROVER [1]. Recently, it has been shown that integrating the outputs given by multiple ASR systems can improve the reliability of the confidence measures [2]. ROVER is one of the most successful integration approaches based on a voting scheme for word error reduction. The voting scheme employed in ROVER is based on the simple idea that "The hypothesized words that are commonly found in many ASR system's output would be reliable".

Despite the high performance of ROVER, it has two serious problems. The first problem is the difficulty in constructing multiple ASR systems. ROVER needs two or more ASR systems, and each ASR system's characteristics should be greatly different from the others in order to improve the recognition performance. Therefore, ROVER needs a lot of resources, multiple corpora, multiple models, multiple decoders, and so on. The second problem is the computational cost for decoding. Since the decoding process of each ASR system runs in parallel with those of other ASR systems, the overall computational cost increases in proportion to the number of ASR systems.

In this paper, we propose a new method that detects mis-recognized utterances, based on a ROVER-like voting scheme. In contrast to the conventional ROVER, the proposed method has two advantages: 1) the construction of ASR systems is easy, 2) the computational cost is low.

2. Mis-recognized Utterance Detection Using Multiple Language Models

2.1. Outline of the proposed method

This section describes the outline of the proposed method. In the proposed method, a baseline LM and sub LMs are used where outputs of the baseline LM are verified on the basis of the voting among the sub LMs' outputs.

To solve the problems of ROVER mentioned in the Introduction, the proposed method differs from the conventional voting scheme in the following two points:

- Using multiple LMs generated by clustered sentences, instead of multiple ASR systems.

- Using re-scoring, instead of parallel decoding.

2.2. Multiple LM generation using clustered sentences

The baseline LM is trained by using the whole set of available corpora, while the sub LMs are trained by using clustered sentences which are sub sets of the corpora. In the present study, the corpora were split automatically into sets of clustered sentences based on a minimum entropy optimization criterion whose entropy is calculated using unigram probabilities. Because of the automatic clustering, a large number of sub LMs can be generated easily even if only one corpus is available.

Since the baseline LM uses a much larger amount of training data than the sub LMs, the baseline LM is expected to be more accurate than the sub LMs. Therefore, the baseline LM is used to decode the input speech, while the sub LMs are used to verify the decoder's output and detect mis-recognized utterances.

A problem arises when increasing the number of clusters to obtain multiple sub LMs. The more we increase the number of clusters, the less amount of training data remains for each sub LM, and thus the less accurate sub LMs we obtain. To solve the insufficient training data problem, each sub LM is linearly interpolated as is given by
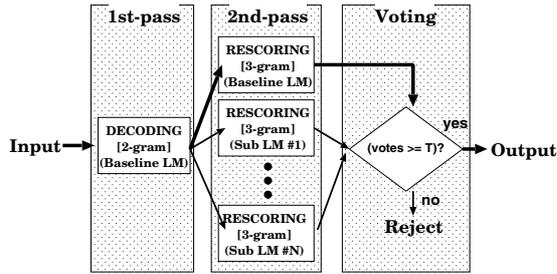
Figure 1: Outline of the proposed ASR where the function of rejecting mis-recognized utterances is implemented.

$$\hat{P}_{sub_n}(W_i \mid \cdot) = (1-\lambda)P_{base}(W_i \mid \cdot) + \lambda P_{sub_n}(W_i \mid \cdot), \quad (1)$$

where $P_{base}(W_i \mid \cdot)$ and $P_{sub_n}(W_i \mid \cdot)$ denote the N-gram probability of the baseline LM and the n-th sub LMs, and λ is the linear interpolation coefficient.

2.3. Rejection of mis-recognized utterances using multiple LMs

Figure 1 shows an outline of the proposed ASR system where the function of rejecting mis-recognized utterances is implemented. The proposed framework consists of a 2-pass decoder and a voting machine:

- In the 1st-pass process, the ASR system recognizes the input speech using the baseline 2-gram LM, and outputs a word lattice.

- In the 2nd-pass process, the ASR system re-scores the word lattice using each of the 3-gram baseline LM and 3-gram sub LMs, and outputs the 1-best hypothesized utterances respectively. The 1-best utterance obtained by the baseline 3-gram LM is treated as the recognition result of the ASR.

- In the voting process, to estimate the confidence for the recognition result, the 1-best utterances generated by each sub LM are utilized. In the present study, the voting machine counts the number of the same 1-best utterances generated by sub LMs as the recognition result of the ASR. If the voting count is greater than or equal to a threshold, the recognition result is considered to be reliable, and is accepted. If the count is less than the threshold, the recognition result is rejected.

It should be noted that the computational cost of the 1st-pass process is much larger than that of the 2nd-pass process where only re-scoring is performed. Since only the baseline LM is used in the 1st-pass, total computational cost is greatly reduced in comparison with the conventional ROVER architecture.

3. Experimental setup

In the present study, the performance of the proposed method was evaluated with respect to whether it can detect mis-recognized utterances effectively.

3.1. Experimental conditions

A travel-task corpus named the Basic Travel Expression Corpus (BTEC)[3] was used for evaluation since we have been developing a speech translation system for the travel conversation task.

Table 1: Experimental conditions

Sampling frequency	16KHz
Frame length	25ms
Frame shift	10ms
Features	12 MFCC, 12 ΔMFCC, Δpower
Acoustic model	gender-dependent, 1,400 states, 5 mixtures HMM
Language models	1st-pass: word 2-gram 2nd-pass: word 3-gram
Decoder	ATRSPREC[4]
Dictionary size	36,810 words

The BTEC consists of a lot of basic expressions that are often used in travel conversations.

In the evaluation experiments, 161,744 utterances were used to train the models, and 510 utterances were exclusively used for evaluation. Table 1 shows the other experimental conditions.

Under the conditions described above, the baseline performance of the ASR without rejection function is 88.89% in word accuracy and 68.63% in utterance accuracy.

3.2. Measures for performance evaluation

When ASR is used as the input of a speech translation system, the recognition performance of the ASR is measured on the basis of the accuracy of the accepted utterances. To that end, in the present study, we evaluated the proposed method by two measures: recall and precision defined by

$$\text{Recall} = \frac{\text{\# of accepted correct utterances}}{\text{\# of correct utterances}},$$

$$\text{Precision} = \frac{\text{\# of accepted correct utterances}}{\text{\# of accepted utterances}}.$$

If all the utterances are accepted without rejection, the precision is identical to the utterance accuracy of the baseline ASR (68.63%), and recall becomes 100%.

4. Experimental Results

4.1. Evaluation of the proposed method

Figure 2 shows experimental results obtained when the parameters were chosen so that the highest precision was achieved when recall was between 80% and 90%. In the figure, N denotes the number of sub LMs, λ is the coefficient of linear interpolation of the sub LMs, and "Base Line" expresses the baseline precision rate of the ASR without rejection function. Each line in this figure shows the relationship between precision and recall by varying the threshold (T).

It can be seen from the figure that as we increase the threshold the precision increases while the recall decreases. The proposed method achieved 18-point higher precision with 10% loss of recall, and 22-point higher precision with 20% loss of recall, when compared with the baseline performance. Another observation from the figure is that the parameters, such as λ, have large influence on both the maximum precision and recall when the precision is fixed. This suggests that the parameters should be optimized depending on the precision and recall required.

4.2. Comparison with conventional methods

The performance of the proposed rejection method was compared with the conventional methods listed:

a) Rejection using normalized likelihood

Hypothesized utterances are rejected if their normalized

Table 2: Systems used in the ROVER approach

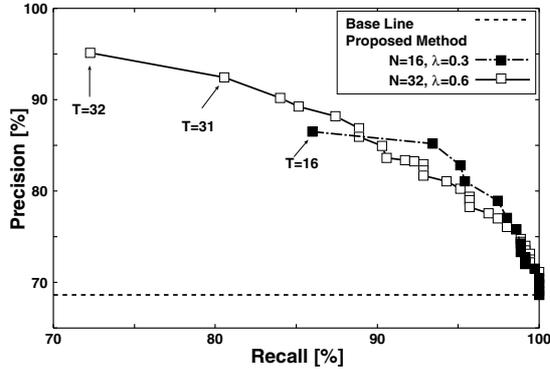|	Acoustic model	Language models		Accuracy [%]	
		1st-pass	2nd-pass	Utterance	Word
SYSTEM 1	gender-dependent, 1,400 states, 5 mixtures HMM	word 2-gram	word 3-gram	68.63	88.89
SYSTEM 2	gender-dependent, 1,400 states, 5 mixtures HMM	multi-class composite 2-gram[5]	–	65.69	88.55
SYSTEM 3	gender-dependent, 2,000 states, 16 mixtures HMM	word 2-gram	word 3-gram	67.25	88.89

Figure 2: Experimental results in the conditions that achieve the highest precision with about 80% or 90% in recall.
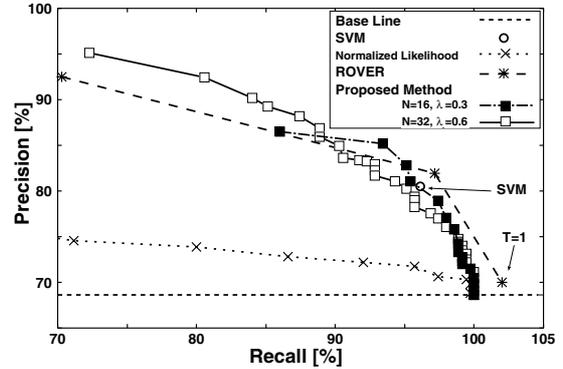

Figure 3: Comparison of the proposed method and previous works

likelihood per frame is lower than a threshold.

b) Rejection using support vector machine (SVM)

Hypothesized utterances are verified by using the support vector machine (SVM). In the present study, three features were used as the feature vectors for the SVM: normalized likelihood per frame, a posteriori probability, and ratio of likelihood of first candidate and second candidate were employed. The SVM was trained with 1,010 utterances which were not included either in the evaluation data set or the training data set used by ASR.

c) ROVER

ROVER was originally developed to reduce word recognition errors and hence it has not been used for the purpose of utterance rejection. Here, the ROVER approach was used in such a way that a majority decision per utterance was performed among the outputs of three ASR systems. If the number of majority votes for the selected recognition result is greater than or equal to a threshold, the selected recognition result is considered to be correct. Table 2 shows conditions and recognition rates of the systems used by ROVER. In this table, SYSTEM 1 is the same as the ASR system used for evaluation of the proposed method. The acoustic model of SYSTEM 3 was provided by the IPA Japanese dictation free software project[6]. Other conditions are the same as in Section 3.1. If each ASR system produces different output from the others and the threshold is $T = 1$, the recognition result of SYSTEM 1 is selected and is considered to be the correct recognition.

Both methods a) and b) use confidence measures based on features that are extracted from a single ASR, while method c) and the proposed method use confidence measures based on features that are extracted from the outputs of multiple ASRs/LMs.

Figure 3 shows the experimental results of the proposed method and the conventional methods, where the experimental conditions are the same as those described in Section 4.1.

When $T = 1$, ROVER achieved over 100% in recall, since many correct utterances, mis-recognized by the baseline ASR used for the proposed method, are chosen as a result of majority decision.

First, the proposed method and ROVER are compared with other methods. The experimental results indicate that the confidence measures extracted by multiple ASRs/LMs can achieve precision that is higher than or equal to the confidence measures extracted by a single ASR. Next, the proposed method was compared with ROVER. The proposed method achieved higher precision than ROVER when recall was lower than 95%. Additionally, in this experiment, the computational cost of the proposed method was a third of ROVER's cost.

4.3. Performance according to each parameter

In this section, we investigate how each parameter influences the performance of the proposed method.

4.3.1. Number of sub LMs

Here, we investigate how the number of sub LMs influences the performance of the proposed method.

Figure 4 shows the experimental results when $\lambda = 0.6$. It can be seen from this figure that the maximum value of precision increased with the number of sub LMs. However, the difference of precision yielded with the same recall was very little.

4.3.2. Linear interpolation coefficient

Here, we investigate how the linear interpolation coefficient, used for sub LMs, influences the performance of the proposed method.

Figure 5 shows the experimental results for $N = 10$. In this figure, as the linear interpolation coefficient increases, the maximum precision rises and the recall drops. Therefore, it can be seen that it is necessary to determine the linear interpolation coefficient depending on the recall required.
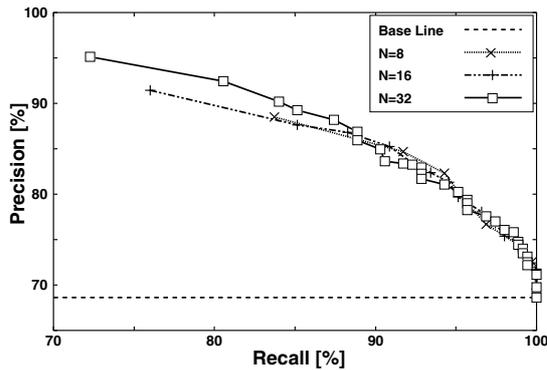
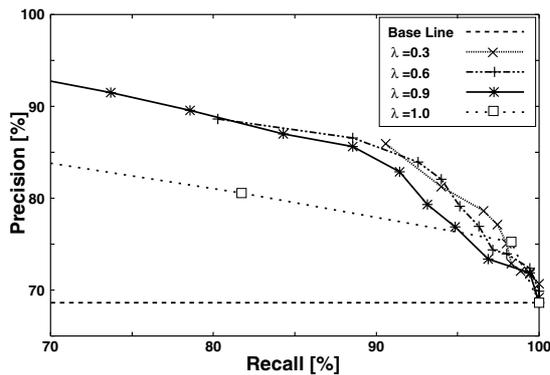Figure 4: Experimental results according to the number of sub LMs

Figure 6: Experimental results according to the clustering criterion (1)

Figure 5: Experimental results according to a linear interpolation coefficient

Figure 7: Experimental results according to the clustering criterion (2)

4.3.3. Clustering criterion

In the above experiments, a minimum entropy optimization criterion (ENTROPY) was employed for clustering the corpus, and its entropy was calculated by 1-gram. Here, we investigate how the clustering criterion influences the performance of the proposed method. In the evaluation experiment, we compare five criteria: 1-3) ENTROPY calculated by 1-gram, 2-gram, 3-gram, 4) random, and 5) topic. In case of the topic, each sentence of the corpus was classified using topic tags that indicate the scenes, e.g. "Shopping", in which it was used.

Figures 6 and 7 show the experimental results for $N = 10$, $\lambda = 0.9$. In these figures, the clustering criterion hardly influenced the performance. However, the performance dropped a little when random clustering was used. Since entropy-based clustering does not need topic tags, it is a more practical and inexpensive method compared with topic-based clustering.

5. Conclusion

In this paper, we proposed a new method that detects mis-recognized utterances, based on a ROVER-like voting scheme. In contrast to the conventional ROVER, the proposed method has two advantages: 1) the construction of ASR systems is easy, 2) the computational cost is low. The experimental results indicated that the proposed method achieved 18-point higher utterance precision with 10% loss of recall, and 22-point higher utterance precision with 20% loss of recall, when compared with the baseline performance of the ASR without rejection function. Additionally, in a Japanese to English speech translation task, our experiments indicated that the proposed method is effective in restraining mis-translations.
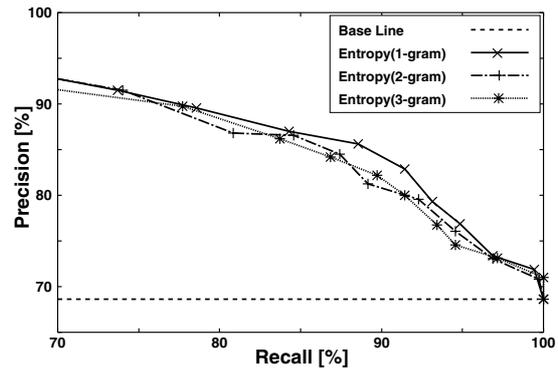
7. References

[1] J. G. Fiscus. "A Post-processing System to Yield Reduced Error Word Rates: Recognizer Output Voting Error Reduction (ROVER)". Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 347–354, 1997.

[2] Y. Kodama, T. Utsuro, H. Nishizaki, S. Nakagawa. "Experimental Evaluation on Confidence of Agreement among Multiple Japanese LVCSR Models". Proc. EUROSPEECH 2001, pp. 2549–2552, 2001.

[3] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," Proc. 3rd International Conference on Language Resources and Evaluation, Vol. I, pp. 147-152, 2002.

[4] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga, and Y. Sagisaka. "Spontaneous Dialogue Speech Recognition Using Cross-word context Constrained Word Graphs". Proc. ICASSP 1996, pp. 145–148, 1996.

[5] H. Yamamoro, S. Isogai, Y. Sagisaka. "Multi-Class Composite N-gram Language Model for Spoken Language Processing Using Multiple Word Clusters". Proc. ACL2001, 2001

[6] K. Itou, K. Shikano, T. Kawahara, K. Takeda, A. Yamada, A. Itou, T. Utsuro, T. Kobayashi, N. Minematsu, M. Yamamoto, S. Sagayama, and A. Lee. "IPA Japanese dictation free software project". Proc. ICSLP 2000, pp. 1343–1350, 2000.