Automatic Speech Recognition Framework for Multilingual Audio Contents

Hiroaki Nanjo, Yuichi Oku, Takehiko Yoshimi

Faculty of Science and Technology, Ryukoku University
Seta, Otsu 520-2194, Japan
nanjo@rins.ryukoku.ac.jp

Abstract

Automatic speech recognition (ASR) for multilingual audio contents, such as international conference recordings and broadcast news, is addressed. For handling such contents efficiently, a simultaneous ASR is promising. Conventionally, ASR has been performed independently, namely language by language, although multilingual speech, which consists of utterances in several languages representing the same meaning, is available. In this paper, we discuss a bilingual speech recognition framework based on statistical ASR and machine translation (MT) in which bilingual ASR is performed simultaneously and complementarily. Then, according to Japanese speech recognition with corresponding English text and MT, we show the framework works well.

Index Terms: multilingual speech recognition, statistical machine translation, multimedia contents

1. Introduction

According to the progress of information technologies and globalization, large-scale multimedia contents, such as broadcast news and recordings of international conference or meeting, can be distributed quickly all over the world through the wide-band networks. For foreigners, simultaneous interpretation is required for a quick understanding of such multimedia contents, and actually manual interpretation has been performed for many of such contents. To make such multimedia contents more universal, a captioning is significant. Especially, for large-scale multimedia contents, automatic captioning is strongly required. Automatic speech recognition (ASR) is promising for automatic captioning, and some captioning systems based on ASR have been realized.

Conventionally, ASR-based automatic captioning systems mainly target on monolingual speech [1] [2]. In this paper, we focus on multilingual speech, which consists of utterances in several languages representing the same meaning. For efficient use of such multilingual audio materials, ASR strategy which handles them appropriately is required. Specifically, ASR framework in which corresponding utterances of several languages are recognized simultaneously and complementarily is needed.

Based on the background, in this paper, we discuss bilingual ASR framework based on statistical ASR and machine translation (MT). Overview of the framework is shown in Figure 1. Conventional studies of combination of ASR and MT have been mainly focusing on a computer assisted translation of texts [3] [4]. In such works, users who want to translate some texts do not write down but just utter, and ASR with text-based MT is performed to dictate the utterances. On the contrary, our research target is ASR with speech-based MT.

2. Statistical Speech Recognition and Machine Translation

2.1. Framework of Statistical Speech Recognition

The orthodox statistical ASR is formulated as finding the most probable word sequence $\hat{W}$ for an input speech $X$, which is described as follows.

$$\hat{W} = \arg \max_W P(W|X)$$  \hspace{1cm} (1)

$P(W|X)$ can be rewritten $P(W)P(X|W)/P(X)$, and since $P(X)$ does not affect the maximization, Equation (1) is rewritten as below.

$$\hat{W} = \arg \max_W P(X|W)P(W)$$  \hspace{1cm} (2)

Speech recognition is a process to find the best word sequence $\hat{W}$ with two scores $P(X|W)$ and $P(W)$. Here, a model which gives $P(X|W)$ is called acoustic model, and a model which gives $P(W)$ is called language model.

Generally, a logarithm function is adopted, and then, scaling parameters $\alpha$ and $\beta$ are introduced as follows. Here, $N_w$ represents the number of the
words of word sequence $W$.

\[
\hat{W} = \arg\max_W \left( \log P(X|W) + \alpha \log P(W) + \beta \log N_w \right) 
\]  

(3)

2.2. Framework of Statistical Machine Translation

The orthodox statistical machine translation (MT) is formulated as finding the most probable word sequence of target language $\hat{J}$ for a word sequence of source language $E$. The process is described in Equation (4).

\[
\hat{J} = \arg\max_J P(J|E) 
\]  

(4)

Here, $P(J|E)$ is a translation score from English text $E$ to Japanese text $J$, namely, correspondence score of $E$ and $J$. A model which gives score $P(J|E)$ is a translation model.

2.3. Bilingual Speech Recognition Framework – Incorporation of Statistical Machine Translation to Speech Recognition

Simultaneous multilingual ASR is formulated as finding the most probable word sequence of target language for input speeches of target language and other languages. For example, in ASR of Japanese utterance $X$ using corresponding English utterance $Y$, word sequence $J$ which gives maximum $P(J|X, Y)$ is searched. The procedure is shown in Equation (5). Here, since $P(X, Y)$ does not affect maximization, it is eliminated.

\[
\hat{J} = \arg\max_J P(J|X, Y) 
\]  

(5)

Introducing $E_m$ which represents one of possible English word sequences for English speech $Y$, Equation (5) is described as follows. Here, since $X$ and $Y$ only depend on $J$ and $E$, respectively, $P(X|J, E_m, Y)$ and $P(Y|E_m, J)$ are rewritten as $P(X|J)$ and $P(Y|E_m)$, respectively.

\[
\hat{J} = \arg\max_J \sum_{E_m} P(J, E_m, X, Y) 
\]  

(6)

Adopting logarithm function to a right side of the Equation (6) and introducing scaling factors, Equation (6) is rewritten as follows.

\[
\hat{J} = \arg\max_J \left( \log P(X|J) + \alpha \log P(J) + \beta N_w 
\]  

(7)

Substituting $a - b$ with $\alpha$ and $b$ with $\gamma$, the ASR is described as Equation (8).

\[
\hat{J} = \arg\max_J \left( \log P(X|J) + \alpha \log P(J) + \beta N_w + \gamma \log \sum_{E_m} P(Y|E_m)P(E_m|J) \right) 
\]  

(8)

Right side of the Equation (8) consists of Japanese ASR score (log-domain) “$\log P(X|J) + \alpha \log P(J) + \beta N_w$”, English ASR score “$P(Y|E_m)P(E_m)$”, and translation model score “$P(J|E_m)$”. In the same
way, ASR of English is described below.

\[
\hat{E} = \arg\max_E \left( \log P(Y|E) + \alpha \log P(E) + \beta N_w + \gamma \log \sum_{J_n} P(X|J_n)P(J_n)P(E|J_n) \right)
\]  

These equations show that bilingual ASR can be performed with translation model and ASR systems of each language. The framework is shown in Figure 1.

In this paper, we evaluate proposed framework of bilingual ASR. To evaluate the framework properly, ASR for a specific language should be evaluated by getting rid of influences of ASR errors of another language. Here, Japanese ASR with corresponding English texts (correct transcript; not ASR result of English utterance) is evaluated. ASR with this assumption corresponds to the case that there is only one hypothesis \(E_m\) in Equation (8). Representing the English hypothesis as \(E\), Equation (8) is rewritten as follows.

\[
\hat{J} = \arg\max_J \left( \log P(X|J) + \alpha \log P(J) + \beta N_w + \gamma \log P(J|E) \right)
\]  

Equation (10) shows that in order to perform bilingual ASR, we need three components; 1) translation model, which gives translation score, 2) ASR system, which generates ASR results and their scores, and 3) a decoder searching for the best word sequence that maximizes a product (sum in log-scale) of translation score and ASR score. In this paper, as a decoding method, N-best rescoring is adopted.

3. Experimental Result

In this section, ASR system and translation model are described, and then experimental results are described.

3.1. ASR System

As for acoustic model, gender independent monophone model (129 stats, 64 mixture) and PTM (phonetic tied-mixture) triphone model [5] trained with 260 hours of read speech by 4130 speakers are used. They are based on continuous density Gaussian-mixture HMM. In PTM modeling, triphone states of the same phone share Gaussians but have different weights. Here, 129 codebooks of 64 mixture components were used. Speech analysis is performed every 10 msec. and a 25-dimensional parameter is computed (12 MFCC + 12 Δ MFCC + Δ Power). For language model, a word trigram model with the vocabulary of 60K words trained with 350M words of newspaper articles is used. We set up ASR system that consists of these models and a decoder Julius rev.3.4.2 [6].

3.2. Translation Model

As a translation model, we adopted IBM model-3. We trained a translation model with bilingual texts of Reuters newspaper articles, in which Japanese and English sentences are aligned [7]. In training, words that occur less than 2 times are regarded as unknown word. Statistics of unknown word are then used in a calculation of translation score. Specification of the training data is shown in Table 1.

3.3. Evaluation data

In this paper, bilingual ASR framework is evaluated on read speech recognition task. Evaluation data is designed as follows. Firstly, we selected 50 aligned Japanese-English sentences from an English textbook for Japanese learner, which consists of transcriptions of broadcast news (English) and their translation texts (Japanese). Then, we asked three Japanese speakers to read the Japanese parts (translation texts), and set them as a test data (150 utterances). The specification is shown in Table 2.

3.4. Result

ASR using statistical MT, namely bilingual ASR, was performed according to Equation (10). Here, we use monophone model for acoustic model. Scaling parameters \(\alpha\) and \(\beta\) are determined so that insertion error and deletion error are equivalent based on preliminary test. Figure 2 shows results of bilingual ASR for each speaker and scaling parameter \(\gamma\). It is confirmed that word accuracy has improved for all speaker. Although optimal value of \(\gamma\) differs among speakers, it is effective for all speakers when \(\gamma\) is set to 0.05 through 0.4. The results show that the proposed bilingual ASR framework works well. For
speaker C, ASR accuracy is relatively high, and the effect of bilingual ASR is small. This indicates that bilingual ASR is effective when ASR accuracy of 1-best result is lower because there are possibly more accurate candidates in the N-best list.

Then, we describe the result of bilingual ASR with PTM triphone model. Table 3 lists the result. For comparison, ASR result with monophone model is also shown. Here, scaling parameter $\gamma$ was optimized for each speaker. Using PTM triphone model in bilingual ASR, word accuracy was slightly improved (0.56% relative). The effect of bilingual ASR is smaller than when monophone model is used. In the read speech recognition task, ASR system with triphone model generally gives higher accuracy, and 1-best result is often the most accurate hypothesis among the N-best list. The result also suggests that bilingual ASR is effective when 1-best accuracy is lower. Lattice rescoring may be more effective since lattice accuracy generally outperforms N-best list accuracy.

For spontaneous speech such as broadcast news and international conference recordings, ASR 1-best accuracy is much lower, and the 1-best result is not so reliable. Therefore, the proposed bilingual ASR is promising for spontaneous speech.

4. Conclusion

Bilingual ASR framework based on statistical ASR and MT, in which corresponding utterances of two languages are recognized simultaneously and complementarily, was described. Based on the results of ASR of Japanese news readings with corresponding English texts, we showed that proposed ASR framework was effective, especially when ASR accuracy is not so high.

5. References


