Improving Non-native Speech Recognition Performance by Discriminative Training for Language Model in a CALL System

Hongcui Wang*, Tatsuya Kawahara† and Yuguang Wang*
* Tianjin University, Tianjin, 300072
E-mail: hcwang@tju.edu.cn
† Kyoto University, Kyoto, 606-8501, Japan
E-mail: kawahara@ar.media.kyoto-u.ac.jp

***Abstract*— High non-native speech recognition performance is always a challenge for a CALL (Computer Assisted Language Learning) systems using ASR (Automatic Speech Recognition) for second language learning. Conventionally, possible error patterns, based on linguistic knowledge, are added to the ASR grammar network. However, the effectiveness of this approach depends much on the prior linguistic knowledge. In this paper, we design a new scheme for error prediction using two sequential machine learning methods. The first step of the prediction method is aiming at the generality, in which decision tree-based error prediction is adopted in our previous work. The second step of the training is aiming at removing most of the redundant candidates. For the second step, we propose a method based on discriminative training to judge the error candidates that degrade the ASR performance and remove them from the ASR grammar network. An experimental evaluation shows that the proposed method can effectively improve non-native speech recognition performance by drastically reducing the False Alarm rate. Moreover, the smaller WER (Word Error Rate) also confirms the effectiveness of our method.**

I. INTRODUCTION

Computer-Assisted Language Learning (CALL) systems have become popular with the aid of the ASR technology, especially in the field of second language learning [1, 2]. So far CALL systems using the ASR technology mainly concentrate on practicing and correcting pronunciation of individual vowels, consonants and words, such as the system in [3]. Although some systems allow training of an entire conversation, such as the Subarashii system [4], little has been done to improve learners' communication ability including vocabulary skill as well as grammar skill. We developed a system (CALLJ) to aid students learning Japanese as a second language [5]. The system offers students the chance to practice elementary Japanese by creating their own sentences based on visual prompts, before receiving feedback on their mistakes. Considering a broad range of variations in learners' accent and the fact that the system has an idea of the desired target sentences in a CALL system, a dedicated grammar network for each question, including error candidates, is dynamically generated.

Conventionally, the prior linguistic knowledge in a CALL system is often used to improve the detection accuracy of typical pronunciation errors [6, 7]. According to the second language learning theory, the typical pronunciation errors made by non-native speakers are mostly produced by the influence of their mother language and can be predicted empirically. However, this method is effective only if the learner of the system is limited to one country. A much more amount of error patterns exist if the system allows any non-native speakers, as in CALLJ system [5]. It is hard to learn an appropriate number of error patterns to predict and keep its generality when new lessons or vocabulary words are added to the system.

To keep the generality of the error prediction method, a decision tree-based error prediction for the ASR grammar network is proposed in [8]. However, redundant error candidates may be generated for some words using this method, e.g. candidate "よっつう" is predicted for "よっつ". Also different error candidates in the grammar network have different degradation impacts for ASR performance, e.g. for the target word "きっぷ", error candidate "きぷ" is presumed to be easier to make ASR errors than the candidate "チッキ".

In order to generate a set of less redundant error candidates, in this paper, we design a discriminative model to train the impact weights for ASR errors with different error patterns and then get a simpler language model. The discriminative model is used to judge error candidates, which easily lead to system errors which are probably not made by students. We will delete such a kind of error candidates from the ASR grammar network. For each error pair, a set of features is generated based on the linguistic knowledge and the error type classified by decision-tree learning. As a result, false alarm should be reduced, so that students would not be bothered by the errors while they speak a correct answer.

II. DISCRIMINATIVE TRAINING WITH PERCEPTRON ALGORITHM

In this section, we describe our methods that improve non-native speech recognition performance by reducing errors that tend to be made by the system rather than the students. Two steps are conducted in the error prediction process for

grammar network generation in our system. First, given a question (sentence), we dynamically generate a grammar network based on the effective prediction error method using decision-tree method, which is completed in [5]. We examine each specific error candidate if it easily leads to a recognition error by the system. Based on the judgment, we decide whether or not to add this candidate to the language model. That is, if P(System Error| Error Candidate) > P(Student Error| Error Candidate), then we will abandon this predicted error candidate, otherwise we will keep it in the ASR language network. The step two is the main task in this paper.

A. Errors in System

A system error and a student error are defined in CALL systems. A system error is the error made by the system. It includes two cases. If the student gives a correct answer, but the system recognizes it as an error candidate, this is one kind of the system error. If the student gives a wrong answer, but the system recognize it as a correct answer or another error candidate, then this is another kind of the system error. A student error is defined as the wrong answer made by the student.

Let TS=(tw_1,tw_2,…,tw_n) denote a target sentence (answer given by the CALL system), LS=(lw_1,lw_2,…,lw_m) denote the actual sentence (transcription of the student's utterances) and RS=(rw_1,rw_2,…,rw_h) denote the recognition sentence (hypnosis of the ASR). We align the three sentences and pick every error triple denoted by (tw_i,lw_j,rw_k) where $tw_i \neq lw_j$ or $lw_j \neq rw_k$. In the error triples, either the student or the ASR system made mistakes. These are listed in Table I.

B. Discriminative Training

Generative training and discriminative training are two different types of model training. In reality, the generative learning method often falls in sub-optimal because of the lack of training data. Also, generative learning only aims to maximize the likelihood of the correct model rather than to minimize the error ratio. Thus, the discriminative training has been investigated, as in [9, 10].

For an ASR system, suppose X is the acoustic data and H is a recognition sentence. Our goal is to find the best H^* which maximizes $P(H_i \mid X)$ as follows.

$$H^* = \arg\max_{Hi} P(H_i \mid X) = \arg\max_{Hi} P(H_i \mid X)P(H_i) \quad (1)$$

If we can discriminate between the correct hypothesis and the incorrect hypotheses, the performance of ASR system can be improved. So we need to strengthen the correct hypothesis and weaken the incorrect hypotheses. Here, we introduce a parameter γ to strengthen or weaken the hypothesis. The evaluation function of (1) is modified to (2).

$$P(H_i \mid X) \cdot \gamma \cdot P(H_i) \quad (2)$$

Where $0 \le \gamma < 1$ is the weaken parameter and $\gamma \ge 1$ is the strength parameter. Here in our case, the language model is the grammar network, so we can simply add or remove a candidate from the network utterly. This means if the error candidate is more likely to cause a system error, then we remove the candidate from the grammar network, which makes $P(H_i) = 0$. In order to judge whether each specific error candidate causes the system's error or not, we need to train a binary classifier which will be introduced in the following section.

TABLE I
STUDENT ERROR AND SYSTEM ERROR COMBINATIONS

Student \ System	Correct Recognition	System Error
Correct answer	Correct ($tw_i=lw_j$ & $lw_j=rw_k$)	False Alarm (FA) ($tw=lw$ & $lw \neq rw$)
Wrong answer (SErr)	Error Detected (ED) ($tw_i \neq lw_j$ & $lw_j=rw_k$)	Error Undetected (EU) ($tw_i \neq lw_j$ & $lw_j \neq rw_k$)

Features are generated based on the linguistic knowledge and the output of the decision tree-based error classification result.

The training data of word error pairs were collected through trials of the prototype of the CALL system with speech input. All trial data consist of 140 sentences. And 145 student's error pairs are contained.

C. Error Pair Features

We could simply count the number of errors for a target word incorrectly recognized or pronounced to estimate whether it is more likely to cause an ASR system error or a student error. However, to archive better prediction performance needs a large number of error samples, since each error item is regarded as a pattern. It is impossible to collect enough data and classify a new error pair with high confidence. Hence, we need to design features to represent the characteristics of error pairs. This provides generality of the discriminative model training. Except the error specific type classified based on the decision tree (for details, see [8]), other features are listed in Table II.

TABLE II
FEATURES USED IN DISCRIMINATIVE TRAINING

#	Feature description	Example
1	Number of phonemes	
2	POS of target word	verb
3	Error type using decision tree	DW(Different form)
4	Confusion of long pronunciation	<[to :],[to]>
5	Confusion of short phoneme	<[da],[d da]>
6	Replacement of [t] with [d]	<[ta],[da]>
7	Replacement of [b] with [p]	<[ba],[pa]>
8	Replacement of [k] with [g]	<[ka],[ga]>
9	Omission or addition of [u]	<[shyo u],[shyo]>
10	Number of different phoneme pairs	
11	Position of different phoneme pair	M (between middle phones of the word)

TABLE IV
THE FEATURE CONTRIBUTIONS FOR CLASSIFYING ERRORS

Feature Selection	Precision	Recall	F-measure
- Number of syllable	0.781	0.777	0.778
- POS feature	0.807	0.805	0.806
- Error type feature	0.781	0.777	0.778
- Omission/addition phoneme features	0.751	0.750	0.750
- Replacement phoneme features	0.734	0.719	0.722
- The position of error phone	0.700	0.692	0.694
All Features	0.807	0.805	0.806

Algorithm 1. Training of perceptron algorithm

Inputs: Error pairs and error labels$\{(w_1, w_2)_i, y_i\}$, i=1,2,…,m;

Outputs: Parameters $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n)$;

1. Transform $(w_1,w_2)_i$ into an input vector x_i;
   $x_i = (f_1,f_2,\ldots,f_n)$, i=1,2,…,m ;
   $f_j(w_1,w_2)$, j=1,…,n; (feature functions)
2. Initialize parameters:
   $\lambda_i = 0$,i=1, 2, …,n ;
3. For t = 1,2,..,T (T is the number of iterations)
   For i = 1,2,…,m
   a) Calculate $z_i = sigmoid(\lambda \cdot x_i + \delta)$;
   b) If $z_i \neq y_i$ Then: (η – learning rate)
   $\lambda^{(t+1)} = \lambda^{(t)} + \eta \cdot (z_i - y_i) x_i$;
4. Repeat step 3 until λ cease to change;
5. Output parameters λ ;

Fig. 1 Training process of perceptron algorithm

D. Training with Perceptron Algorithm

We applied a perceptron algorithm to train the discriminative model. Combining the situation in this work, the specific procedure is showed in Fig. 1.

This training is conducted for the system's errors and the student's errors using TS-LS and LS-RS pairs, respectively. Then, the probabilities for the system' error P(System Error | Error Candidate) and the student's error P(Student Error | Error Candidate) are given by the respective sigmoid function, z_i .

E. Training Result

In the training process, 75% data were randomly selected for the training, and the remaining 25% were used for testing the method. Here, precision, recall and F-measure are used to evaluate the performance of the final classifiers. Precision is the ratio of errors that are correctly classified. Recall is the coverage of correctly classified errors. And F-measure is a harmonic mean of the precision and recall.

TABLE III
OVERALL PRECISION, RECALL AND F-MEASURE OF CLASSIFIER

Error Type	Precision	Recall	F-measure
System Error	0.824	0.778	0.800
Student's Error	0.789	0.833	0.811
Weighted Avg.	0.807	0.806	0.805

In Table III, the overall performance result is listed. 82.4% of system errors are correctly classified with F-measure of 80%. The contribution of the features by eliminating one by one is showed in Table IV. We can see the position of the error phone is the most important feature for the classification because it affects the performance most when we eliminate it. The second most critical feature is the error type derived from the decision tree-based classification method. The POS does not affect at all as it is actually included in the error type feature.

III. EXPERIMENT

To evaluate the effectiveness of the proposed approach for the improvement of the non-native speech recognition performance, we conducted a closed experimental evaluation. Ten foreign students (eight males, two females) from eight different countries took part in the experiment. The entire collected utterance data set is used for testing

A. Evaluation Measure

In the experiment database, the three sequences of a target sentence, its correct transcript, and the recognizer's output are aligned word by word. To evaluate the performance of ASR, we use the standard measure of Word Error Rate (WER), together with Error Detection Rate (EDR) and False Alarm Rate (FAR), as well as the perplexity of the ASR language model. The EDR is defined as the number of detected errors (ED) divided by the total number of errors the students made (SErr). The FAR is the number of words erroneously flagged as student errors (FA), divided by the total number of words students spoke correctly.

B. Experimental Result

We compared the performance of the language models based on two different error prediction methods. One is the decision tree-based error prediction (baseline). The other is the proposed discriminative training based method. According

TABLE V
ASR PERFORMANCES

Method	WER	EDR	FAR	Perplexity
Decision tree-base Method	11.2%(78/694)	75.7%(87/115)	8.6%(50/579)	4.1
Proposed Method	8.1%(56/694)	67.8%(78/115)	3.3%(19/579)	3.4

to preliminary experiments, the ASR system achieves the best performance when we remove a candidate from the network as P (System Error | Error Candidate) - P (Student Error | Error Candidate)> 0.1.

Table V shows the result of the close evaluation. The proposed method realized smaller perplexity than the decision tree-based method. Especially, it drastically reduced the FAR (8.6%) to more than half (3.3%), which demonstrates the effectiveness of the learning in reducing the system errors. Moreover, the WER of the proposed method is improved compared with the baseline result. However, the EDR of the proposed method is decreased because the deletion of some error candidates induced the new detection failures of them. For example,"ぎゅうにゅう" can be pronounced wrongly as "ぎゅにゅう" by a student. But in most of the cases it is wrongly recognized as "ぎゅにゅう" even when students pronounce correctly. According to the discriminative model, this error candidate is classified into a system error. So it is removed from the grammar network. Then, even if a student make such a mistake in fact, it will be more likely to be undetected.

IV. CONCLUSIONS

We have proposed a discriminative training approach to improve non-native speech recognition performance by reducing False Alarm errors. Given an ASR grammar network, including error candidates, the trained discriminative model will judge whether each error candidate leads to the ASR system error, and abandon this redundant error candidate to ensure better performance of the system. In a closed evaluation of the experiment, the language model based on the proposed method significantly reduced the FAR to more than half and improved the WER.

REFERENCES

[1] Dean Luo, Yu Qiao, Nobuaki Minematsu, Yutaka Yamauchi, Keikichi Hirose, "Regularized-MLLR speaker adaptation for computer-assisted language learning system", In Interspeech, pp. 594-597, 2010.

[2] Daniel Felps, Heather Bortfeld, and Ricardo Gutierrez-Osuna," Foreign accent conversion in computer assisted pronunciation training", In Speech Communication, 51(10): 920–932, 2009.

[3] Goh Kawai and Keikichi Hirose, "A CALL system using speech recognition to train the pronunciation of Japanese long vowels, the mora nasal and mora obstruent," in *Eurospeech*, 1997.

[4] Jared Bernstein, Ami Najmi, and Farzad Ehsani, "Subrashii: Encounters in Japanese spoken language education," *CALICO*, vol. 16, pp. 361–384, 1999.

[5] Hongcui Wang, Christopher J. Waple and Tatsuya Kawahara, "Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition", Speech Communication, Elsevier, vol. 51, pp. 995-1005, 2009.

[6] Yasushi Tsubota, Tatsuya Kawahara, and Masatake Dantsuji, "Recognition and verification of English by Japanese students for computer-assisted language system," in *ICSLP*, 2002.

[7] Shuang Xu, Jie Jiang, Zhenbiao Chen, Bo Xu, "Automatic pronunciation error detection based on linguistic knowledge and pronunciation space", in Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP), 2009.

[8] Hongcui Wang and Tatsuya Kawahara, "Effective Error Prediction using Decision Tree for ASR Grammar Network in CALL System", in Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP), 2008.

[9] Zheng Chen, Kai-Fu Lee, and Ming Jing Li, Discriminative training on language model, In Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP), Beijing, China, 2000.

[10] Brian Roark, Murat Saraclar, Michael Collins, Mark Johnson, Discriminative Language Modeling with Conditional Random Fields and the Perceptron Algorithm, In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 2004.