Recognition of Utterances with Grammatical Mistakes based on Optimization of Language Model towards Interactive CALL Systems

Takuya Anzai* and Akinori Ito*
* Tohoku University, Sendai, Japan
E-mail: {takuya,aito}@spcom.ecei.tohoku.ac.jp Tel/Fax: +80-22-795-7084

***Abstract*—To realize a voice-interactive CALL system, it is necessary to recognize the learner's utterance correctly including the grammatical mistakes. In this paper, we proposed methods for improving recognition accuracy of speech with grammatical mistakes. The proposed method is based on the method that uses n-gram model trained from sentences that are generated using grammatical error rules. We introduced two improvements to the previous method: one is the utterance discrimination to avoid introducing errors into correct utterances, and the other one is optimization of language model where probability of grammatical mistakes in the generated training text is optimized using the score of utterance discrimination. As a result, we obtained 0.92 point improvement, which is 12% error reduction.**

Index Terms: speech recognition, interactive CALL system, grammatical mistakes, language model

I. Introduction

With the progress of globalization in recent years, population of English learners has been increased. Among various English learning methods, Computer-Assisted Language Learning (CALL) system is one of the most promising learning methods [1]. Most CALL systems focus on training for reading, writing and listening, and a few commercial CALL systems provide with training method for speaking. However, conventional CALL systems with speaking practice focus on training of pronunciation or intonation. To improve conversation skills, it is necessary for learners to practice conversation in a real dialogue. To realize the conversation practice using computer, several voice-interactive CALL systems have been developed [2], [3], [4], [5]. We are now developing an interactive CALL system of English for Japanese learners.

Speech recognizer is used to recognize the learner's utterance. Here, there are two problems for recognizing language learners' utterances. The first one is that pronunciations of learners are different from those by native speakers, and the difference greatly depends on the learner. The second problem is that the utterances made by the learner inevitably contain grammatical mistakes, which are not assumed in ordinary speech recognizers.

The first problem can be solved using acoustic models for non-native speaker [6], [7], [8]. To improve the recognition accuracy, speaker adaptation technique was also proposed for non-native speech recognition [9]. For the second problem, Kweon et al. proposed a rule-based method that expands a network grammar so that utterances with popular mistakes can be accepted [3]. Ito et al. used grammatical error rules similar to the privious method to generate sentences containing grammatical mistakes, from which an n-gram model is trained [4]. Anzai et al. improved the error rules for generating training sentences [5].

In this paper, we introduce two improvements into speech recognition method based on n-gram trained from generated sentences. The first one is introduction of utterance discrimination, which determines whether the input utterance contains grammatical mistakes or not before performing speech recognition. The second one is optimization of language modeling. If the input utterance has many grammatical mistakes, we need to use a language model trained from sentence with many mistakes; if the input utterance is grammatically correct, grammatical errors in the training data may introduce recognition errors. Therefore, we developed a method to estimate how erroneous the input utterance is, and choose the optimum language model to recognize the input utterance.

II. N-gram training from generated sentences

A. Interactive CALL system with pre-exercise

In this work, we assume a dialog with a CALL system with pre-exercise [3], where the learner first studies words and grammars used in the conversations in the lesson, then the learner actually converses with the CALL system. Pre-exercises make it easier for learners to produce speech when using the system. In addition, assuming a pre-exercise before the dialogue session with the system had the effect of suppressing off-task utterances by the learners [3].

As we assume a pre-exercise before the conversation with the CALL system, we can expect the learner to respond to the system using the same expressions as those appearing in the pre-exercise. Therefore, we assume there is a "correct" sentence to be uttered by the learner at a certain situation. We refer to such a sentence, a correct sentence expected to be uttered by a learner, as *the target sentence*. In a real session, however, not all user utterances match the target sentences. We refer to a sentence actually uttered by a learner as *the uttered sentence*. An uttered sentence often contains grammatical and lexical mistakes. The uttered sentences are recognized using the speech recognizer, and the recognition results often have recognition errors. We call the result of the automatic speech recognition as *the recognized sentence*.

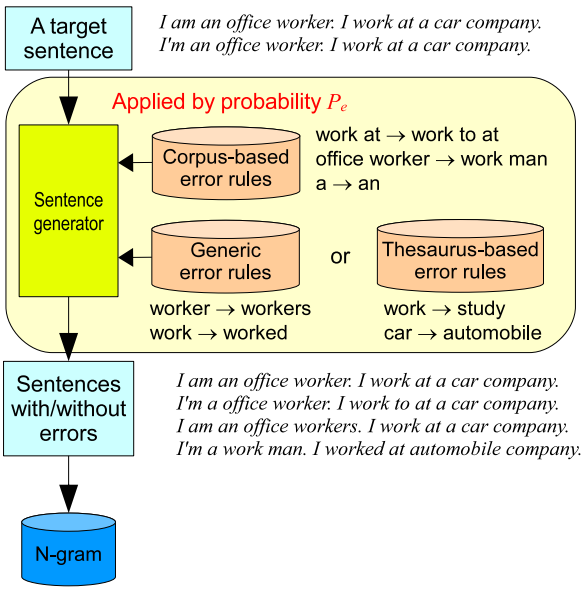


Fig. 1. Language model generation

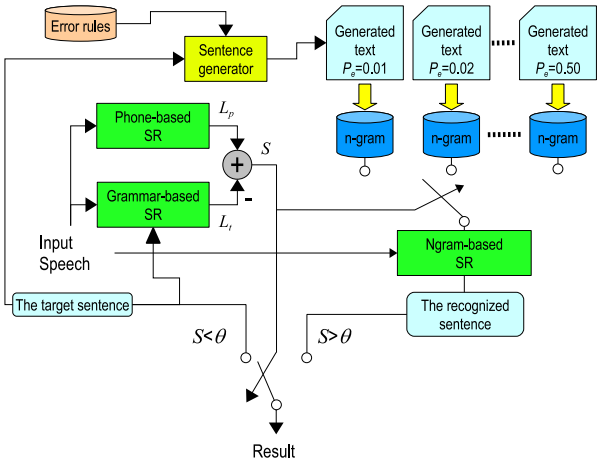


Fig. 2. Overview of the proposed method

B. Sentence generation and n-gram training

Next, we explain how to train the n-gram language model for recognizing the input utterance [4], [5]. Basically, we prepare an n-gram utterance by utterance, assuming that we know the target sentence of the utterance beforehand. Figure 1 shows the procedure of language model training.

First, we prepare grammatical error rules that are frequently made by Japanese learners. We prepare three kinds of rules: the corpus-based error rules extracted from the transcription of English utterances spoken by Japanese speakers [10], the generic error rules such as confusion of singular and plural, and the thesaurus-based error rules generated from WordNet. Then the rules are applied to the target sentences by probability P_e (thus the sentence is unchanged by probability $1-P_e$), and a sufficient amount of sentences are generated. Finally, a back-off N-gram is trained from the generated sentences.

III. Utterance Discrimination and Language Model Optimization

A. Overview

There are two issues in the previous framework. One is the tradeoff between coverage of error rules and recognition accuracy. We need to incorporate more and more error rules to obtain high coverage, but it raises perplexity of the n-gram and deteriorates the recognition performance. The other issue is how to determine P_e. P_e should be high when recognizing utterances with many mistakes, but if the utterance has no grammatical mistakes, high P_e just increases recognition error rate.

To solve the above two problems, we introduce two improvements. The first improvement is the utterance discrimination to determine whether the input utterance contains any words that are not in the target sentence. We use acoustic score as a feature of the discrimination. If the utterance is judged to be correct, we do not perform any further recognition. As this discrimination is independent from the sentence generation, this method prevents the sentences without grammatical mistakes from recognition errors caused by the grammatical error rules. The second one is the optimization of language model. In this method, we first prepare many n-gram models trained from generated sentence sets with different P_e. When recognizing the input utterance, the best language model is selected using the acoustic score difference. Figure 2 shows the overview of the proposed method.

B. Discrimination of utterance with mistakes

First, we explain the input utterance discrimination. The discrimination is based on acoustic score (log-likelihood). We calculate score of the input utterance twice, once using phone recognition without linguistic constraint, then using the grammar that accepts only the target sentence. Let the recognition scores calculated by these processes be L_p and L_t, respectively. Then we calculate the acoustic score difference

$$S = L_p - L_t. \tag{1}$$

Figure 3 shows histogram of utterances with and without grammatical mistakes. We can see that correct utterances have smaller score difference. Therefore we use S as a feature of the discrimination. An input sentence is classified as "correct" when the score difference S is smaller than the threshold θ. If sentence is classified as correct, the target sentence was used as the recognition result. Otherwise, the utterance was recognized using the speech recognizer with the n-gram.

We carried out an experiment to investigate the effectiveness of the utterance discrimination. The experimental conditions are shown in Table I. The test utterances were collected by the following procedure: first, learners were asked to memorize the English target sentences, then utter the sentences by only seeing the Japanese translation of those sentences. Word accuracy of the uttered sentences with respect to the target sentences was 88.3%. The probability P_e was determined *a*

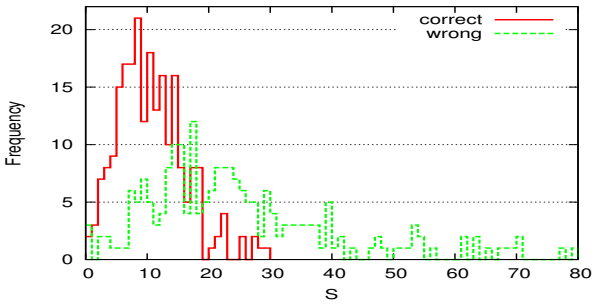


Fig. 3. Histogram of utterances with and without mistakes with respect to acoustic score difference S

TABLE I
EXPERIMENTAL CONDITIONS

Acoustic model	512-mixture 5-state HMM trained using the ERJ database
Acoustic feature	MFCC, ΔMFCC, ΔΔMFCC, Δpow, ΔΔpow
Decoder	Julius 4.1.2
P_e	0.08
Number of generated texts	100,000/target sentence
Test utterances	441 utterances spoken by 15 speakers (14 male and 1 female)

posteriori so that the best word accuracy was obtained on the test set.

Figure 4 shows the word accuracy given by the proposed method. The leftmost part of Fig. 4 shows the word accuracy when all utterances were recognized using the n-gram. As shown, we could obtain an improvement of 0.48 point when setting θ to the value around 12.

C. Optimization of language model

In the previous method, the n-gram language model for recognizing the utterances with mistakes were trained using the generated sentences with fixed error probability P_e. However, the optimum value of P_e differs from utterance to utterance. Figure 5 shows the word accuracy for four utterances with respect to the error probability P_e for generating training sentences. Figure 5 (a) shows the result for correctly uttered utterances (A and B) where the lower P_e gave better results. Figure 5 (b) is that for utterances with mistakes (C and D)

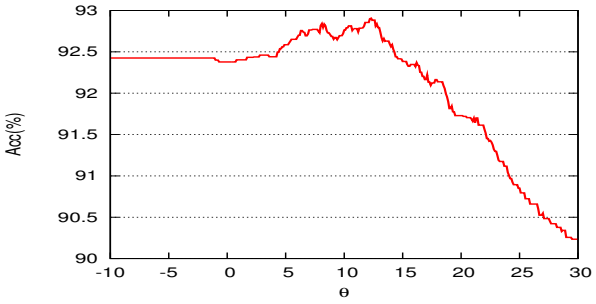
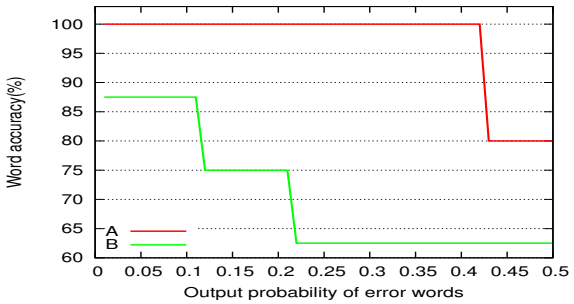
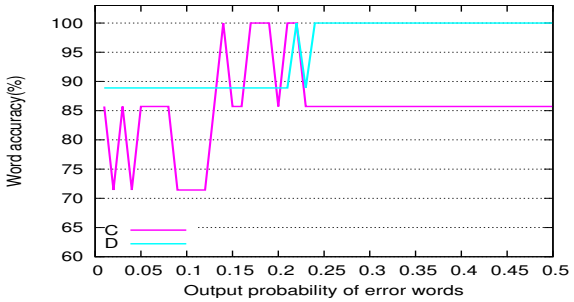


Fig. 4. Word accuracy with respect to the threshold θ



(a) Utterances without mistakes



(b) Utterances with mistakes

Fig. 5. Word accuracy with respect to P_e

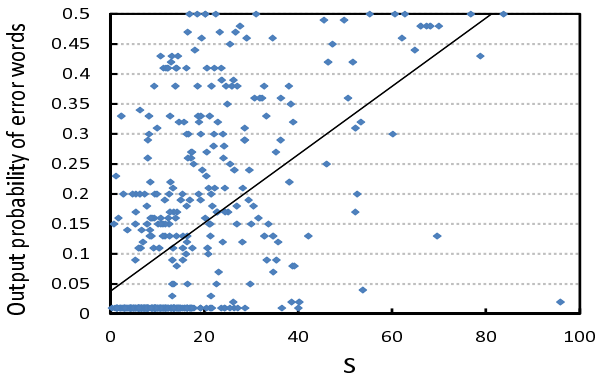


Fig. 6. Acoustic score difference S vs. the optimum error probability (0.01∼0.50)

where the maximum accuracy was obtained with higher P_e. These results suggest that the word accuracy can be improved if we can choose a language model trained with sentences with appropriate error probability.

We use the acoustic score difference S for prediction of the optimum error probability. As explained above, S becomes smaller for sentences without mistakes. Figure 6 is a scatter plot of S and the optimum error probability for each utterance, as well as the result of linear regression. We used the result of linear regression for estimating the optimum P_e.

This framework is a similar approach to the language model selection. In the researches of spoken dialogue, language models are selected depending on dialog state [11], [12]. Our method is different from these approaches, where the language model is optimized using the input utterance. This kind of

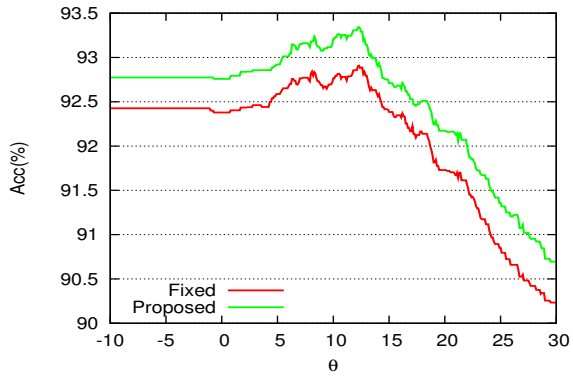


Fig. 7. Word accuracy using the language model optimization

language model optimization is new attempt compared with the conventional language model adaptation schemes.

We carried out an experiment for confirming the effectiveness of the error probability estimation method. The evaluation data was the same as that in the previous section, and 15-fold cross validation (opened for each speaker) was performed for estimating the linear regression coefficients. We prepared 10 language models using $P_e = 0.01, 0.02, \ldots, 0.10$. When the estimated P_e was larger than 0.10, the model trained with $P_e = 0.10$ was used, because the model with larger P_e gave less word accuracy even for utterances with many grammatical errors.

Figure 7 shows the results, where the red line shows the result when $P_e = 0.08$, and the green line is the result when P_e was predicted using linear regression. We can see that the proposed method improved the word accuracy constantly compared with the fixed error probability. The best result was 0.44 point better than the method with fixed error probability, and 0.92 point higher than the result when utterance discrimination and language model optimization were not used.

IV. Conclusion

In this paper, we proposed methods for improving recognition accuracy of speech with grammatical mistakes. The proposed method is based on the method that uses n-gram model trained from sentences that are generated using grammatical error rules. We introduced two improvements to the previous method: one is the utterance discrimination to avoid introducing errors into correct utterances, and the other one is optimization of language model where probability of grammatical mistakes in the generated training text is optimized using the score of utterance discrimination. As a result, we obtained 0.92 point improvement, which is 12% error reduction.

As we could improve the word accuracy, there is still more room for further improvement. If we could perfectly choose the error probability of the language model, the word accuracy raises to 95.1%, that is 1.8 point higher than the above result. We still need to improve language modeling, as well as the acoustic modeling.

Acknowledgment

Part of this work was supported by Grant-in-Aid for challenging Exploratory Research by Japan Society for the Promotion of Science (JSPS), No. 24652111.

References

[1] M. Levy, “Technologies in Use for Second Language Learning,” The Modern Language Journal, 93, pp. 769-782, 2009.
[2] F. Ehsani, J. Bernstein and A. Najmi,“An interactive dialog system for learning Japanese”, Speech Communication, 30, pp. 167-177, 2000.
[3] O. P. Kweon, A. Ito, M. Suzuki and S. Makino, “A grammatical error detection method for dialogue-based CALL system”, Journal of Natural Language Processing, 12(4), 137-156, 2005.
[4] A. Ito, R. Tsutsui, M. Ito and S. Makino, “Recognition of English utterances with grammatical and lexical mistakes for dialogue-based CALL system”, Proc. Interspeech, 2819-2822, 2008.
[5] T. Anzai, S. Hahm, A. Ito, M, Ito and S. Makino, “Grammatical error detection from English utterances spoken by Japanese,” Proc, APSIPA ASC, 482-485, 2010.
[6] S. Witt and S. J. Young, “Language Learning based on Non-Native Speech Recognition,” Proc. Eurospeech, 633–636, 1997.
[7] L. M. Tomokiyo, “Lexical and acoustic modeling of non-native speech in LVCSR,” Proc. ICSLP, 2000.
[8] J. J. Morgan, “Making a Speech Recognizer Tolerate Non-native Speech through Gaussian Mixture Merging,” Proc. InSTIL/ICALL2004, 2004.
[9] Y. Ohkawa, M. Suzuki, H. Ogasawara, A. Ito and S. Makino, “A speaker adaptation method for non-native speech using learners’ native utterances for computer-assisted language learning systems,” Speech Communication, 51(10), 875–882 2009.
[10] Y. Tono, T. Kaneko, H. Isahara, T.Saiga and E. Izumi, “A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography”, Proc. 2nd Asialex International Congress, 257-262, 2001.
[11] D. Duff, B. Gates and S.. LuperFoy, “An Architecture for Spoken Dialogue Management,” Proc. ICSLP, pp. 1024–1028, 1996.
[12] W. Xu and A. Rudnicky, “Language Modeling for Dialog System,” Proc. ICSLP, 2000.