社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69 (26)
2007／12／20

日本語CALLシステムの音声認識における効果的な予測のための決定木に基づく誤りパターンの分類

王 洪翠† 河原 達也†

† 京都大学 情報学研究科
〒606-8501 京都市左京区吉田二本松町
E-mail: †wang@ar.media.kyoto-u.ac.jp

あらまし 音声認識を利用した外国語学習支援（CALL）システムが近年多く研究されているが、非母国語話者の誤りを含む音声の高精度な認識は依然大きな課題である。従来、言語学的知見に基づいて、誤りパターンを音声認識用文法ネットワークに追加する方法が一般的であったが このアプローチは誤りのカバレージと文法のパープレキシティのトレードオフの問題に直面する。本研究ではこの問題に対して、決定木を用いて非母国語話者が生じる発話誤りのパターンを効果的に予測する方法を提案する。自動的に文を生成する課題を用いて、本学の留学生により評価実験を行った結果、提案手法は、実際に生じた誤りに対する高いカバレージと、小さいパープレキシティの両方を実現する効果的な音声認識文法を作成し、実際に認識精度の改善を得ることができた。

キーワード 音声認識, CALL, ネットワーク文法, 決定木

Decision Tree based Error Analysis for Effective Prediction in ASR for Japanese CALL system

Hongcui WANG† and Tatsuya KAWAHARA†

† School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan
E-mail: †wang@ar.media.kyoto-u.ac.jp

Abstract CALL (Computer Assisted Language Learning) system using ASR for second language learning has received increasing interest recently. However, it still remains a challenge to achieve high speech recognition performance for users have various accents. Conventionally, possible error patterns, based on linguistic knowledge, are added to the ASR grammar network. However, this approach easily falls in the trade-off of coverage of errors and the perplexity of the grammar. To solve the problem, we propose a method based on a decision tree to learn effective prediction of errors made by non-native speakers. An experimental evaluation with a number of foreign students in our university shows that the proposed method can effectively generate an ASR grammar network, given a target sentence, to achieve both better coverage of errors and smaller perplexity, resulting in significant improvement in ASR accuracy.

Key words speech recognition, CALL, grammar network, decision tree

1. Introduction

Computer-Assisted Language Learning (CALL) system using ASR has received increasing attention in recent years[1-2]. Many research efforts have been done for improvement of such systems especially in the field of second language learning[2-5]. So far CALL systems using ASR technology mainly concentrate on practicing and correcting pronunciation of individual vowels, consonants and words, such as the system in [2]. Although some systems allow training of an entire conversation, such as the Subarashii system [3], little has been done to improve learners' communication ability including vocabulary skill as well as grammar skill. This work is part of an effort for this direction.

In this setting, the system must recognize learners' sentence utterances for a given scenario (sometimes the sentence itself is given). However, a broad range of variations in learners' accent makes it hard to get sufficiently high speech recognition performance in a second language learning system. On the other hand, since the system has an idea of the desired target sentences, it is natural to generate a dedicated grammar network for it. To be an effective CALL system, the grammar network should cover errors that non-native learners tend to make. Errors here mean answers that are different from the desired target one as well as mistakes including pronunciation errors.

To achieve better error prediction, the linguistic knowledge is widely used. In [4], 79 kinds of pronunciation error patterns according to linguistic literatures were modeled and incorporated to recognize Japanese students' English. However, the learner of the system is limited to Japanese students. Obviously, a much more amount of error patterns exist if the system allows any non-native speakers. Moreover, we need to handle more variations in the input, if we allow more freedom in sentence generation, as we proposed in CALLJ [6], in which a graphic image is given as a scenario and learners are prompted to generate a sentence to describe it. The system is covered in more detail in the next section. These factors would drastically increase the perplexity of the grammar network, causing adverse effects on ASR.

In this paper, we address effective error prediction for the ASR grammar network, which means predicting critical error patterns without a large increase in perplexity. Considering all possible errors easily leads to a large increase in perplexity. In order to find critical errors and avoid redundant ones, a decision tree is introduced for error classification. While a list of possible features (questions) are made based on linguistic knowledge, we introduce a coverage-perplexity criterion in order to derive a decision tree to find only effective features, which result in broader error coverage and a small increase in perplexity, thus are selected for prediction.

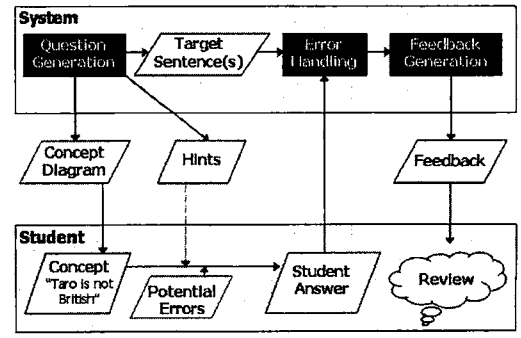

Fig. 1 System Overview

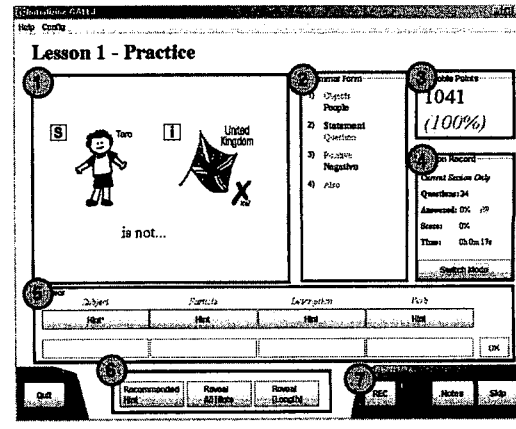

Fig. 2 Lesson Practice Screen; 1: Concept diagram, 2: Desired form guide, 3: Score, 4: Lesson statistics, 5: Answer area and hint display, 6: Further hint functionality, 7: Control button panel

The remainder of this paper is organized as follows. In Section 2. we give an overview of the CALLJ system.In Section 3. we introduce the method of error classification using a decision tree. In Section 4. we present experimental results. Section 5. concludes the paper with a summary.

2. CALLJ Overview

An overview of the CALLJ system is depicted in Figure 1. The system generates questions, on the fly, based on a key grammar point that the students are to practice. Each question involves the students being shown a "Concept Diagram", which is a picture representing a certain situation or scene. The students are then asked to describe this situation with an appropriate Japanese sentence. The interface through which the students carry out these exercises is shown in Figure 2.

In order to reduce the repetitiveness of the questions offered by the system, we dynamically generate each question

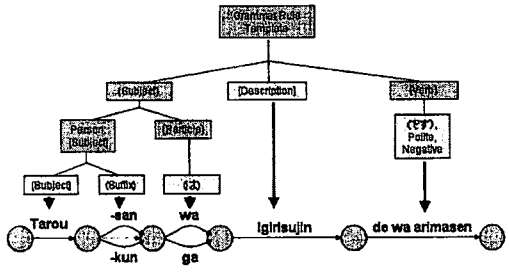

Fig. 3 Grammar-based Sentence

Table 1 Typical Question List

no answer
in dictionary
same POS
same base form
similar concept
same form
pronunciation confusion error
wrong inflection of target word

at run time from the set of vocabulary and grammar rules available. The first task in generating a question is to seltect a sentence concept from the template file. And then by taking the information in the concept instance and applying a set of grammar rules, a set of target sentences are created in a network form, as shown in the lower half of Figure 3.

To expresses the situation or concept the students have to describe in the interface, the system choose to display such information graphically. This helps avoid the problem of expressing the situation via a specific language, which could be problematic in cases where the native language of the students vary. Also, a hypothesis has been put forward that suggests that pictures are easier for the students to process and recall (a phenomena known as the *Picture Superiority Effect* [7]), that they enable the students to comprehend the semantical meaning behind the situation quicker than with text [8], and that they may lead to more satisfying and effective learning [9].

3. Error Classification Using Decision Tree

3.1 Decision Tree

A decision tree is introduced to identify critical errors, or classify error patterns to critical ones and others. The decision tree allows expert knowledge to be incorporated via the questions, and finds an optimal classifier given a training data set. In this work, features or questions are prepared based on the linguistic knowledge, and training data of erroneous patterns actually made by foreign students are also prepared. Then, the data are classified using questions, according to some criterion. In this work, the criterion should be effective in the error prediction. After the training, for all leaf nodes of the final classification tree, "to predict or not to predict" the error patterns are labeled. This decision tree is used to selectively predict error patterns for a given sentence.

The training data were collected through trials of the prototype of the CALLJ system with text input. All trial data consist of 880 sentences. Among them, 475 contain errors.

3.2 Error Categorization

For decision tree learning, an important setup is to identify the features of the data and choose questions for classification. In this work, we assume that all sentence inputs are aligned with the target sentence word by word. Thus, an error pattern could be a wrong word or no word (null string). For wrong words, several kinds of linguistic features can be attributed to the errors.

There are different features and error tendencies among different part-of-speech (POS: verb, noun, etc.), for example, verbs in Japanese take a role of representing sentence tense and voice. Therefore, we make a decision tree for each POS though some of the features are shared. This provides flexibility of using special questions, for example "same base form" is a unique question to verb. Typical features are listed in Table 1.

3.3 Coverage-Perplexity Criterion

In order to select effective features and find critical error patterns, we introduce two criteria of error coverage and perplexity in the grammar network. If we add all possible error patterns in the ASR grammar network, it can detect any errors in consideration in theory, however the ASR performance is actually degraded as a whole because of the increased perplexity in the language model. Thus, we need to find the optimal point in the tradeoff of the coverage and perplexity, which are described below:

- Error coverage

The error coverage is defined as the proportion of errors being predicted among all errors. It is measured by using the training data set, so that more frequent errors are given a higher priority. We can easily measure the increase in the coverage obtained by predicting a specific error pattern.

- Perplexity

The perplexity is defined as an exponential of the average logarithm of the number of possible competing candidates at every word in consideration. In this work, for efficiency and convenience, we approximate it by the average number of competing candidates of every word that appear in the

training data set. Then, we can compute the increase in perplexity when we predict some specific error pattern. For example, if we predict "th→d" confusion, the increase in perplexity is measured by the number of "th" sounds observed in the data (divided by the data size).

In the decision tree learning, we need a measure to expand a certain tree node and partition the data contained in the node. Thus, we define a coverage-perplexity measure (=*impact*) for a given error pattern as below:

$$impact = \frac{\text{increase in error coverage}}{\text{increase in perplexity}}$$

The larger value of this impact, the better recognition performance can be achieved with this error prediction. Thus, our goal is reduced to finding a set of error patterns that have large impacts. If a current node in the tree does not meet this criteria (threshold), we expand the node and partition the data iteratively until we find the effective subsets or the subset's coverage becomes too small (or all questions are applied).

3.4 Training Algorithm

Now we explain the concrete training algorithm: After initializing the classification tree with common baseline questions (no answer, same as the target word, in dictionary, and same POS), all samples fall within one of the classes (=leaf nodes). Then traverse the tree from top to down, from left to right. When finding a leaf node, split the node till the coverage-perplexity impact becomes larger than its threshold, or the coverage becomes smaller than its threshold. In the former case, when the coverage-perplexity criterion is satisfied, the error pattern is identified as effective "to predict". In the latter case, when the coverage criterion is not met, the errors in the node is decided as "not to predict". The recursive process can also be terminated when no more applicable questions are found. In each split, we test features (questions) that can be applied to the current node, and partition this node into two classes. There are constraints in application of the questions, since some of them are subsets of another, and can be applied only after that, for example, "same surface form" is applied after "same base form".

3.5 Example of Classification Result

The classification result for verb is shown in Figure 5. The coverage-perplexity impact threshold used is 0.01 and the error coverage threshold is 0.02. Attached to each type of the errors are the error occurrence frequency (in the training data) and the increase in perplexity. In Figure 5, "similar concept" means that target words are substituted to words having the same meaning or being related potentially. Among this category, we identified as effective subsets "DW_SForm" and "DW_DForm". For words that are not in dictionary, the same principle is applied to identify

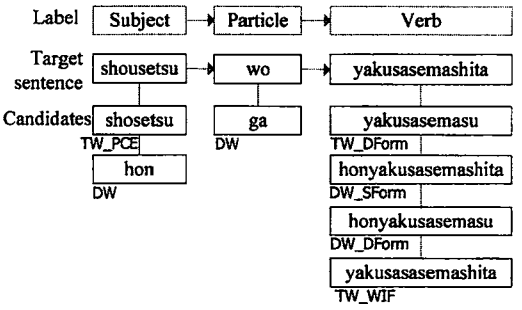

Fig. 5 Prediction Result for Given Sentence

"TW_WIF" (wrong inflection forms of the target word, such as "masu" stem + "te"). On the other hand, "TW_OForm" is predictable in nature, but the expected effect is so small (0.0012) and it may cause adverse effects on ASR, thus it is not included for prediction.

3.6 Error Prediction Integrated to Language Model

As we identified the errors to predict and errors not to predict, we can exploit this information to generate a finite state grammar network. Given a target sentence, for each word in the surface form, we extract its features needed such as POS and the base form, and compare the features with error patterns to predict using the decision tree. Then, we create potential errors of the corresponding error pattern with prediction rules and add them to the grammar node. Figure 6 shows an example of a recognition grammar based on the proposed method for a sentence "shousetsu wo yakusasemashitaka".

4. Experimental evaluation

To evaluate the prediction performance of the proposed error classification and generated grammar networks, we conducted an experimental evaluation.

4.1 Experiment Setup

The platform used for data collection and evaluation is CALLJ, designed for self-learning of the basic level of Japanese language. For this experiment, we have incorporated an ASR system based on Julius to accept speech input.

Ten foreign students of Kyoto University took part in the experiment. They are from seven different countries including China, France, Germany, and Korea. They had no experience with the CALL system before the trial, but were briefly introduced before undertaking the task. Seven lessons were chosen for this experiment. Each student tried two questions for each lesson. Total of 140 utterances were collected. Speech recognition results were presented to the students in the interface after they spoke their answers via a

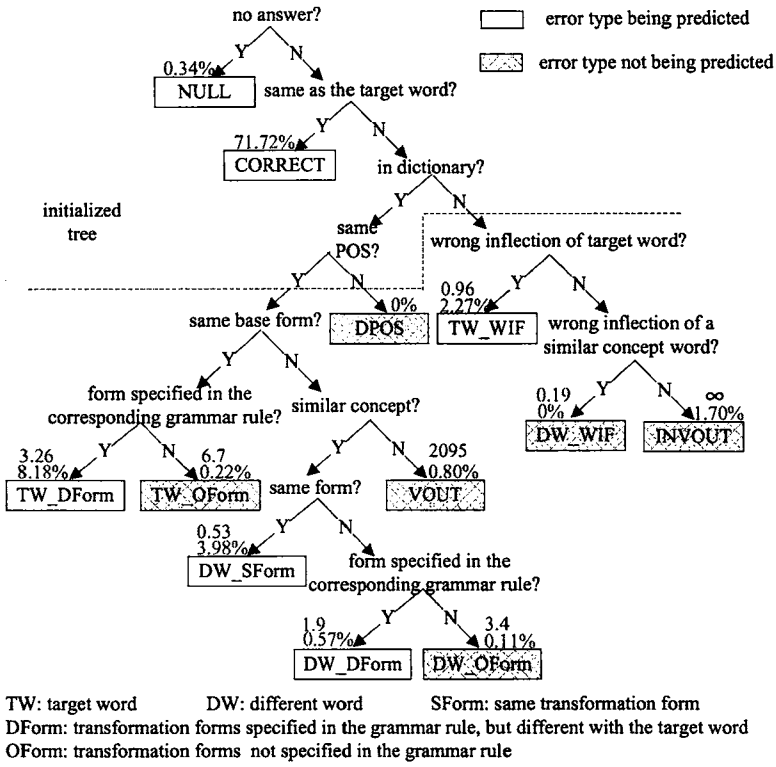

Fig. 4 Error Classification of Verb

microphone. The acoustic model is based on Japanese native speakers. And the language model was built with the proposed method. After the trials, all utterances were transcribed including errors by a Japanese teacher.

4. 2 Experiment Results

We compared three language models based on different error prediction methods:

- Baseline: This is a hand-crafted grammar for the text-input prototype system. It does not consider errors made by foreign students and simply includes all words in the same concept such as foods and drinks in the grammar network, and can be applied to any sentences in the same lesson.
- General method: In this method, we made an error analysis (as categorized in Section 2.2) and predict errors based on the heuristic knowledge to generate a grammar network. Various possible forms of the verbs are added, however, surface forms that are not found in the dictionary are not predicted.
- Proposed method

In Table 2, we present the results for the data set collected via the text-input prototype system, which was used for decision tree learning. This is a closed evaluation. The proposed method realizes significantly better coverage and smaller perplexity. The result validates the proposed learning algorithm. Then, we made an evaluation with the newly collected data via the ASR-based system. The results of the open evaluation are shown in Table 3. It is observed that the error coverage and perplexity are almost comparable to those of Table 2, demonstrating the generality of the learning. The effectiveness of the proposed method is also confirmed by the ASR performance (WER).

Table 2 Performance with Training Data (text input)

Method	Error Coverage	Perplexity
Baseline	37.96%	31.8
General Method	49.58%	22.3
Proposed Method	77.93%	5.05

Table 3 Performance with Test Data (speech input)

Method	Error Coverage	Perplexity	WER
Baseline	44.76%	33.78	28.53%
General Method	53.33%	21.48	24.06%
Proposed Method	85.71%	4.12	11.20%

5. Conclusion

In this study, we have proposed an approach to effective error prediction in ASR for second language learning systems.

A decision tree is successfully applied to identify critical error patterns which realize large coverage without increasing perplexity. In the experiment with the CALLJ system, the language model based on the proposed method significantly outperformed the conventional method and reduced the word error rate to less than a half.

References

[1] Imoto, K., Tsubota, Y., Raux, A., et al., "Modeling and Automatic Detection of English Sentence Stress for Computer-Assisted English Prosody Learning System", Proc. ICSLP, pp.749–752, 2002.

[2] Kawai,G., hirose,K., "A Call System using Speech Recognition to Train the Pronunciation of Japanese Long Vowels, the Mora Nasal and Mora Obstruent", Proc. Europspeech,657-660, 1997.

[3] Bernstein, J., Najimi, A., Ehsani, F. "Subarashii: Encounters in Japanese Spoken Language Education", CALICO Journal, 1999.

[4] Tsubota,Y., Kawahara,T., and Dantsuji,M., "Practical Use of English Pronunciation System for Japanese Students in the CALL Classroom", ICSLP, 2004.

[5] Abdou,S.M., Hamid, S.E., Rashwan, M., et al., "Computer Aided Pronunciation Learning System Using Speech Recognition Technology", Interspeech, 2006.

[6] Waple, C., Wang, H., Kawahara, T., et al., "Evaluating and Optimizing Japanese Tutor System Featuring Dynamic Question Generation and Interactive Guidance ", Interspeech, 2006.

[7] Nelson, D. L., Reed, U. S., Walling, J. R. "Picture Superiority Effect", Journal of Experimental Psychology: Human Learning & Memory, 1976

[8] Smith, M. C., Magee, L. E. "Tracing the time course of picture-word processing", Educational Communications and Technology Journal, 1982

[9] Levie, W. H., Lentz, R. "Effects of text illustrations: A review of the research", Journal of Experimental Psychology: General, 109, 373-392, 1980