Evaluating and Optimizing Japanese Tutor System Featuring Dynamic Question Generation and Interactive Guidance

Christopher Waple[*], *Hongcui Wang*[*], *Tatsuya Kawahara*[*†], *Yasushi Tsubota*[†], *Masatake Dantsuji*[†]

[*]School of Informatics, Kyoto University,
[†]Academic Center for Computing and Multimedia Studies, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract

We are developing a new CALL system to aid students learning Japanese as a second language. This system is designed to allow students to create their own sentences based on visual prompts, receiving feedback based on their mistakes. The questions are dynamically generated, resulting in a large variety of challenges. The students may choose to receive guidance in order to complete each task, selecting the level of help that best suits their needs. A scoring system is also incorporated, which awards a grade to students based on the errors made and hints used. The trial of the system has been conducted with twenty one students, providing the statistics of actual errors and hint usages. With these data, we have trained the weights of the scoring system by taking into account the impact of each issue on the proficiency of the students. The validity of the estimated score is generally confirmed by predicting the proficiency of the students.

Index Terms: CALL, second language learning, Japanese

1. Introduction

Given the widespread and ever increasing exchange of knowledge, culture and manpower between different countries, the advantages and motivation for learning a foreign language are clear. Combined with the pervasiveness of the personal computers, it comes as no surprise that there is significant interest in the development of Computer Assisted Language Learning (CALL) systems.

There are many CALL systems that have already been developed both academically and commercially, covering different aspects of language study. However, these systems tend to be limited either by the repetitiveness of the learning material, or by the lack of freedom offered to the students. A study comparing the relative advantages and disadvantages of a system that allows a free form of input compared to those which restrict the students' answers has been carried out previously [1], showing strong benefits for the open-input approach. Also, it has been shown [2] that students who practice via sentence-production exercises will on average perform better when it comes to creating their own sentences.

Based on these observations, we have designed and developed a new CALL system, *CallJ*, to aid students learn elementary Japanese grammar and vocabulary via a set of dynamically generated sentence production exercises. This allows the students the freedom to create their own sentences, and receive feedback based on any errors they make. An interactive hint system is included, which allows the students to choose when to receive help, and how much help to receive in order to solve a task. A scoring system is also incorporated, to give the students an indication of improvement, thus increasing motivation. We previously presented the basic concept, along with some early prototyping work, in [3].

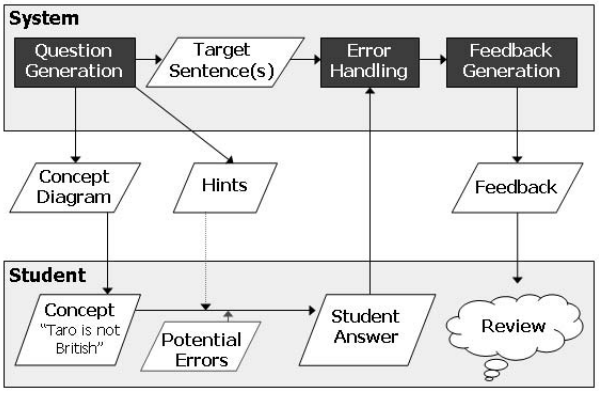

Figure 1: System overview

In this paper, we give an overview of the fully implemented system, focusing on the optimization of the weights for individual errors incurred and hints used, based on the experimental trials by foreign students. Section 2 gives the system description, and Section 3 presents the findings of the trials.

2. CallJ - System Design

An overview of the 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 the followings, we describe further detail regarding the main features of the system, namely question generation, error handling and feedback, and the scoring system.

2.1. Question Generation

In order to reduce the repetitiveness of the questions offered by the system, we dynamically generate each question at run time from the set of vocabulary and grammar rules available. This involves the creation of four main components: a concept or situation that the students must describe, a diagram that expresses this situation, target sentences that the students are expected to

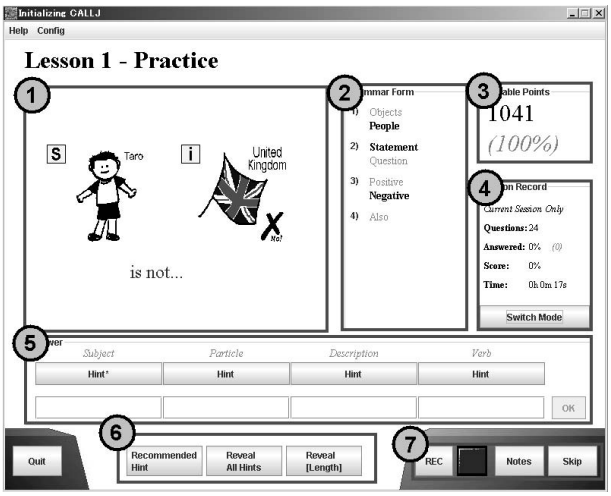

Figure 2: Lesson practice screen; 1: Concept diagram, 2: Desired form guide, 3: Score, 4: Lesson statistics, 5: Answer area and hint display, 6: Further hint functionality, 7: Control button panel

produce, and a set of hints that may be used by the students to reach their answer.

2.1.1. Concept Definition

The first task in generating a question is to generate the situation to be described. A template is prepared to cover a range of related situations. It defines the semantic components or slots that are required, optional or to be omitted when defining a specific situation. The system then selects which information slots are to be activated (the optional slots are chosen randomly). For these active slots, the system selects an appropriate concept instance, depending on the attributes of the slot specification.

2.1.2. Diagram Generation

The system generates a diagram that expresses the situation or concept the students have to describe. Displaying such information graphically 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 varies. 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* [4]), that they enable the students to comprehend the semantical meaning behind the situation quicker than with text [5], and that they may lead to more satisfying and effective learning [6].

Having the system generate the diagram offers a number of advantages. Firstly, it significantly reduces the cost time-wise in creating the images. Secondly, it leads to a greater consistency in style across the images. The diagram is created by combining a number of smaller sub-images, each representing a component of the concept instance.

2.1.3. Sentence Generation

A set of target sentences are created in a network form, as shown in the lower half of Figure 3. The network is created by taking the information in the concept instance and applying a set of grammar rules.

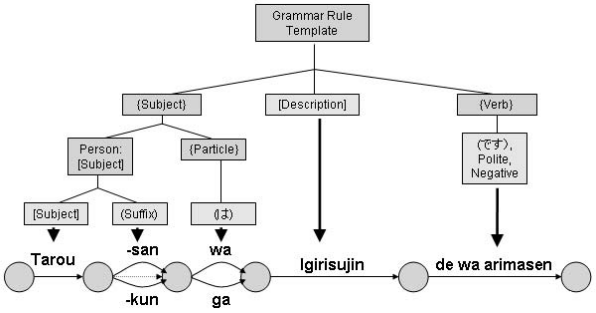

Figure 3: Grammar-based sentence generation

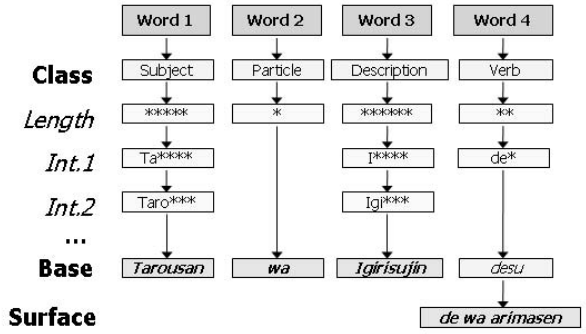

Figure 4: Component-based "hint chains"

Consider the example given in Figure 3. The top-level template specifies that the sentence should consist of three components: Subject, Description and Verb. These three components are each parsed in turn. Subject component, for example, is comprised of two sub-components: a sub-rule that expands into the subject itself (appending a suffix to the name if appropriate), and the associated particle. The complete network representing the valid sentences for the given question is generated in this way, the words being defined by the leaf nodes of the network.

2.1.4. Hint Generation

The hint system allows the students to reveal each word in the target sentence in stages, thus allowing them to receive just the amount of help they need to complete the task. The hints are generated by breaking down the target sentence (a typical sentence selected from the sentence network) into its constituent components, and then for each component creating an ordered set of hints.

Figure 4 shows an example of a sentence being broken down into a set of hints. All of the sentence components have a class hint and a base form hint. The number of the components and their class hints are presented in the first place, as shown in Figure 2. The base form hint is actually divided into a number of sub-hints, revealing the target word character by character. If the students actually know a word but have a trouble in remembering it, initially giving them small sections of the words may be enough to help them remember the word, and would thus be more useful than just giving them the whole word straight away. Usage of the hint system comes at a cost that is applied to the students' score. The scoring system is covered in more detail in Section 2.3.

2.2. Error Handling and Feedback

For the students to learn from their mistakes, it is vital that they be told where these mistakes are, the nature of each mistake, and how the mistake can be corrected. Thus, once the students enters their answer, the system must first detect if there are any errors in that answer. For each word in the students' answer, the sentence network was searched to find the closest matched word in the sentence position. If there is a mismatch, the input word is labeled as an error.

The error classification results from comparing the features of the input word to that of the closest matched word in the target answer. Features determined and analyzed for error classification include whether both words are of the same grammatical type, whether they share semantic tags, the string distance, any inflections etc. A decision tree is then used to take these features and determine the most appropriate error classification. The error classes were determined by initially considering the language error classes *Grammatical* and *Lexical* (as used in [7]), to which we added two further classes *Input* (to deal with mistakes in the input format, such as hiragana being used instead of katakana) and *Concept* (to deal with mistakes not in the language itself, but in the interpretation of the situation that the students need to describe). These classes were divided into further sub-classes based on the specific error features.

Each error class has a template feedback text string, describing the error, possible causes, and a suggested solution. For each error, this feedback is given to the students together with the correct word (the closest matched one).

2.3. Scoring System

To help motivate students and give them an idea of how they are progressing, we also implement a scoring system. This system penalizes students for making mistakes as well as for using hints in order to answer a question. Determining the values of these penalties is an important point, as we should penalize the errors and hint usages which are seen as having large effect on students' proficiency with a greater score reduction than for insignificant errors. To implement this functionality, it is clearly necessary to identify which error classes or error features have a large impact on proficiency. A set of weights are used for this purpose.

There are three different weight groups. The *component type weights* represent the cost incurred if an issue (either an error or a hint being used) occurs on a specific component type. The *error type weights* represent the cost associated with each possible error class. The *hint level weights* represent the cost for revealing hints at each hint level. Table 1 shows the estimated weights. The process for training these weights through the experimental trials is given in Section 3.2.

The penalty incurred for an error is calculated by summing together the weight for the associated error type with the weight for the component type. The costs for using all the hints on a particular word is also based on the maximum penalty associated with that word, and is distributed across the different hint levels, the percentage of the cost for each level being determined by the hint-level weights. The total score available for each question is based on the sum of the maximum error penalties for each word in the target sentences.

Table 1: The set of weights used by scoring system

Type	Feature	Value
Component Weights	Verb	10
	Noun	9
	Other	7
	Particle	6
	Location	4
	Definitive	3
	Counter	1
Error Type Weights	Grammatical	10
	Non-Spelling	10
	Lexical	6
	Concept	4
	Input	2
Hint Level Weights	Base Form	10
	Length	8
	Intermediate	8
	Surface Form	3

3. Experimental Results

A trial of the system was conducted by a number of students running through a set of lessons, and giving their feedback on the system. Twenty one students took part in the trial. All students were currently studying Japanese within the Kyoto University Japanese language course, and thus their approximate language proficiency was known based on the level of course to which they were assigned (Elementary, Intermediate 1 or Intermediate 2). The main goals of the experiments are summarized below:

- To investigate the tendencies and frequencies of errors made and hints used by the students
- To estimate the weights for the scoring system based on the above information
- To investigate whether the students' language proficiency can be estimated from the above

Each student ran through a set of eight lessons, answering a set of generated questions for each lesson. For most of the analysis, we combined the Intermediate 1 and Intermediate 2 students' results together, considering them as a single group, Intermediate.

3.1. Errors and Hint Usage by Students

Figure 5 shows the frequency of the types of errors detected by the system during the trial. The frequency is calculated by dividing the total occurrence of each error type by the number of components observed on which that error type may occur. It is observed that the most common form of problems are lexical errors. We also investigated the hint usage rate in a similar way, and found that the most commonly used hint level was the base-form. Both these results suggest that lexical errors were more common than the grammatical ones, and that the students had more issues with vocabulary than they did with the grammar structures.

3.2. Training and Evaluation of Scoring System

In order to train the various weights used in the scoring system, we determined how significant each of the features upon which these weights are based are to the students' overall language proficiency. To this end, we trained a Support Vector Machine (SVM) to take the hint usage and error data of the students, and estimate which proficiency group they belong to. For training

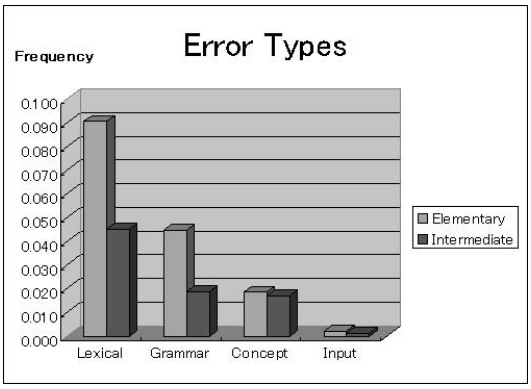

Figure 5: Frequencies of observed error types

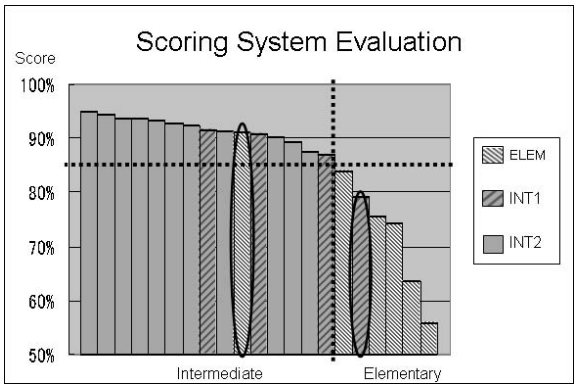

Figure 6: Evaluation of scoring system

purposes, each student was labeled as being either Elementary or Intermediate.

The SVM was initially trained using all of the available features, but the feature set was reduced using a greedy linear regression function (bottom-up construction and top-down reduction). We assume that the order the features were pruned/added during this process provides an approximation of each feature's significance to the overall proficiency estimation. However, there was a great deal of inconsistency in the derived rankings. This was caused by the fact that a number of the features are overlapped in terms of the data they represent, and thus are redundant. To avoid these inconsistencies, we adopted another approach. Instead of training an SVM including all the features, we trained three separate SVMs based on the three different feature categories: component types, error types, and hint-level usages. Within each of the groups, the overlap in information between the features is minimal, and as such we were able to obtain more consistent rankings for each of the features. The weights of the scoring system were determined based on the associated feature's relevant ranking within the feature group.

Once the weights were defined, we evaluated the scoring system's performance, by calculating each student's score from the answers he gave (and the hints he used) during the trial. This estimation was conducted in a "leaving-one-out" manner in which each test student's data is excluded in the weight estimation. Figure 6 shows the score obtained by each student using the trained weights, ordered from the highest on the left-hand side and the lowest on the right. The majority of the elementary students are clustered to the right-hand side, with the lowest scores. In this graph, we split the intermediate group into two groups: intermediate 1 and intermediate 2. From these results, we may consider setting a threshold of 85% as the boundary between intermediate and elementary classes. Although this would lead to two misclassifications, the overall result is encouraging, with a hit rate of 90.5% (=19/21). This shows that the trained score system offers a meaningful measure of proficiency of the students, and validates the approach we have taken to the cost estimation.

4. Conclusion

We have designed and implemented a new interactive CALL system, *CallJ*, for students of the Japanese language, with features aimed at reducing repetitiveness and increasing the freedom. We have successfully carried out a set of trials of the implemented system, capturing a significant amount of data regarding the errors the students made, along with the hints that they used throughout the trial. This data was used in training and evaluating the system's scoring system.

Ongoing work on the system includes the integration of speech recognition technology to allow the students to practice spoken Japanese. Because of the fundamental language differences between spoken and written Japanese, such functionality would be included via a separate set of practice schemes, as opposed to being included as an alternative input method to text within the current lessons. We also look at expanding the system so that it uses the students error record to influence the selection of questions and vocabulary they must face.

5. References

[1] Yang, J.C., and Kanji, A., "An Evaluation of Japanese CALL Systems on the WWW Comparing a Freely Input Approach with Multiple Selection", CALL Journal, Vol.12 No.1, pp.59-79

[2] Nagata, N., "Japanese Courseware for Distance Learning", AILA, 2000

[3] Waple, C., Tsubota, Y., Dantsuji, M., Kawahara, T., "Prototyping a CALL System for Students of Japanese Using Dynamic Diagram Generation and Interactive Hints", Interspeech, 2006

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

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

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

[7] Lister, R., "Negotiation of Form, Recasts, and Explicit Correction in Relation to Error Types and Learner Repair in Immersion Classrooms", Language Learning 48:2, June 1998, pp.183-218

[8] Yang, J.C., and Kanji, A., "Development of computer assisted language learning system for Japanese writing using natural language processing techniques: A study on passive voice", Proceedings of the workshop "Intelligent Educational Systems on the World Wide Web", 1997