Chinese and Japanese Word Learning System
by Estimation of Word Difficulty

Peiqing Zang†  Hiromitsu Shiina‡  Nobuyuki Kobayashi*  Fumio Kitagawa‡
†Graduate School of Informatics
Okayama University of Science, China
‡Faculty of Informatics, Okayama University of Science, Japan
* Faculty of Human Science, Sanyo Gakuen University, Japan
zangpq@gmail.com

Abstract
Many foreign students are also enrolling universities in Japan. When foreign students undertake lectures in Japanese, there is an assumption that they understand a good deal of Japanese. However, they do not understand all of the vocabulary needed for lectures. We believe that a Japanese word preparation learning system for providing the vocabulary required for a lecture is required. In addition, there is a change toward using mobile devices for learning, such as iPhone, iPad, Android Tablets, and so on. As such, we need to provide a simple method for learning. This research is developing a system for practicing Japanese and Chinese vocabulary using terms taken from a text for beginners, as well as technical terms taken from VOD lectures on a database server that have open access to documents such as lecture videos and slides. We are also utilizing the difficulty level of Chinese and Japanese terms to determine the order of appearance and to extract incorrect terms in multiple choice vocabulary problems. To determine the difficulty level of each word, we used support vector machine (SVM) to estimate the difficulty of words with no preexisting difficulty level based on words whose level is already established, and then used a bootstrap method which made use of the estimated difficulty levels of these words.

Keywords: Word difficulty, Machine Learning, Support Vector Machine, Learning Support System

1 Introduction
Many foreign students are now enrolling universities in Japan. It is assumed that a foreign student attending lectures in Japanese will be quite advanced in the study of the Japanese language, but in reality students often arrive after a short period of Japanese language study and have an incomplete grasp of the language. Because of this, they cannot understand all of the words used in their lectures. We believe these students need a self-learning system, before and after class, the Japanese terms used in their lectures.

Meanwhile, mobile devices such as the iPhone, iPad, and Android Tablet are changing the learning environment, and because of this we believe that simple learning methods are desirable. Many software programs have already been developed to help learners study foreign language vocabulary. Among these, nearly all of the multiple choice quiz applications are structured to present a target language prompt with answer choices in the user’s native language. Few applications present questions with a native-language prompt and target-language answer choices, and even fewer combine the two methods.

In this research, we address this by creating a system structure that combines Japanese language and Chinese language vocabulary training systems. The goal is to improve the results of vocabulary training, in particular by using Japanese-to-Chinese study sessions and multiple-choice questions in combination with Chinese-to-Japanese multiple choice questions and answers.

Our vocabulary training system was developed for the Android Tablet device, and in order to create this system it was necessary to estimate the relative difficulty levels of the words used as vocabulary items and incorrect answer choices.

To classify words by difficulty, we used the difficulty levels from the Chinese Proficiency Test (HSK[1]) for Chinese and the Japa-
nese-Language Proficiency Test (JLPT[2]) for Japanese. To determine the difficulty level of each word, we use SVM[3] to estimate the difficulty of words with no preexisting difficulty level based on words whose level is already established, and then used a bootstrap[4] method which made use of the estimated difficulty levels of these words.

Finally, for the subjects to be studied, we chose categories of words which foreign students would need to start life in Japan as well as those which would be useful to learn before attending lectures at a Japanese university.

2 Estimating vocabulary difficulty levels

2.1 Estimating vocabulary difficulty levels

When using data from the dictionary to estimate the level of difficulty for vocabulary, we theorized that the words used in the dictionary definition will be either easier than, or the same difficulty as, the entry word. Under this theory, the difficulty level distribution seen in the words that appear in the definition correlates to the level of difficulty of the entry word. We then carried out machine learning with an SVM, using the difficulty level distribution of the words in the definition as parameters.

This allowed us to estimate the level of difficulty of all the dictionary entry words. The difficulty estimation learning algorithm is a bootstrap method which applies words for which difficulty has been determined from teacher data to the definition, and then repeatedly assesses the difficulty of words for which difficulty has not been determined. The relationship between teacher data and the dictionary and the SVM learning process is shown in Figure 1. The following shows the steps involved:

(Step1) SVM learning with the initial learning data
If the level of difficulty of the words in the dictionary definition had not been estimated, the level of difficulty of those words was exempted from the difficulty level distribution, then learning parameters were made and learning was carried out by an SVM.

When the level of difficulty of the words was classified as L, and the frequency for each word difficulty level appearing in the semantic

Figure 1. A Summary of the Difficulty Learning Process

description for entry word \( w \) was \( D(l, w) \), \( l = 1, ..., L \), then the standardized word difficulty frequency distribution was defined as follows:

\[
DR(l, w) = \frac{D(l, w)}{\sum_{i=1}^{L} D(i, w)}.
\]

The SVM learning parameter (level of word difficulty of \( (w, DR(1, w), DR(2, w), ..., DR(L, w)) \) was composed of the level of difficulty the word \( w \) and \( L \) in the difficulty level distribution in the dictionary definition.

(Step2) Repeated SVM learning
(Step2-1) If the level of difficulty of the entry word was not in the initial data, we estimated it using the difficulty level distribution of the words in the dictionary definition.

(Step2-2) If the levels of difficulty for all of the words in the definition had not yet been determined, we made a difficulty level distribution from words whose difficulty level was known, then we estimated the difficulty level of the entry word.

(Step2-3) We updated the difficulty level of the words in the definition, then carried out classification learning with an SVM using parameters combining the level of difficulty of the entry words of the initial learning data.

(Step2-4) We then repeated (Step2-3). If the changes in results for the estimated level of difficulty converged, then we were finished.
2.2 Determining difficulty levels in Chinese

We used the test levels from the HSK as difficulty levels for Chinese. For the initial dataset[5],[6], we used a difficulty level list compiled from lists of words which have appeared on the test in the past. In addition, we used Hanyu Da Cidian as Chinese dictionary[7]. Difficulty levels of words not on this list were determined by extrapolation. The difficulty levels of Chinese terms were determined by the same method discussed for Japanese below.

The Chinese method differs from the Japanese in the way dictionary entries are partitioned into individual words when difficulty levels are assigned; for this purpose, we have used a form of morphological analysis[8] suited to Chinese.

The HSK is divided into six levels, with Level 1 being the easiest and Level 6 the most difficult. As an example, Figure 2 shows how the difficulty level is determined for 「加油」(jiayou). In Figure 2, the zeroth difficulty level estimation assigns difficulty levels only to the words whose difficulty is already known, and the SVM estimates a difficulty level of 4 based on the difficulty level distribution (4/11, 1/11, 1/11, 1/11, 2/11, 2/11). The first estimation uses the zeroth estimation to estimate the difficulty of the dictionary entry word and then uses that information to assign a difficulty level to each word; here, the SVM estimates a difficulty level of 5 based on the difficulty level distribution (4/17, 4/17, 2/17, 3/17, 2/17, 2/17). Note that symbols and words whose difficulty levels have not been determined are labeled “0” in Figure 2.

The word 「汽车」(feiji) appears in the dictionary definition used in the difficulty level estimations but is not included in the teaching data, and so it is assigned Level 2 during the first difficulty level estimation; it is then used in the estimation of 「加油」.

3 Selection of vocabulary items

3.1 Selection of vocabulary categories

Vocabulary categories were selected from the New Standard Japanese for Sino-Japan Communication Primary[9] and arranged in order of which were considered the most crucial to life in Japan. Lessons are divided into tenth study categories: (1) greetings; (2) transportation; (3) food; (4) people; (5) numbers; (6) facilities; (7) the week; (8) climate; (9) occupations; and (10) clothing.

3.2 Selection of vocabulary items and incorrect answer choices

Words which appear as questions were selected from the categories discussed above, and their difficulty levels in each language were used to arrange the words in order of increasing HSK difficulty in the case of Chinese and increasing JLPT difficulty in the case of Japanese. We think that when the difficulty level of both Japanese terms and their Chinese translations is low, they are conceptually easier to understand, and that these should be the easiest terms to study. Even if the Japanese is easy, it is still difficult to understand if the Chinese is difficult, and we suppose...
that studying is most difficult when both terms are difficult.

Incorrect answer choices were chosen based on the following criteria:
(1) Incorrect choices are chosen from the same category as the answer.
(2) Incorrect choices are of a similar level to the answer.

However, during the actual development of the system, incorrect answers were first selected automatically by the computer and then checked manually before being included. This is because the computer does not consider the meaning of the words (Figure 3).

4 The vocabulary training system

The vocabulary training system offers multiple-choice vocabulary questions with Chinese prompts and Japanese answer choices, as well as questions with Japanese prompts and Chinese answer choices. The following two training methods are provided:
(1) Normal Mode: The user studies multiple choice questions from one language at a time. If the user is studying Chinese, he or she will move on to the next Chinese prompt (Figure 4).
(2) Mixed mode: Correct answers appear again as questions. If the user is studying in Chinese, there will be a question on the Japanese word he or she selected, and the program will return to a Chinese question prompt when it is completed (Figure 5).

5 Future Works

In this research, we determined the difficulty of Chinese and Japanese vocabulary terms via an SVM and developed a training system which utilized these difficulty levels in creating incorrect answer choices. Study terms included words selected from the New Standard Japanese for Sino-Japan Communication Primary [15] and specialized terms used in university lectures. Furthermore, the vocabulary training method we developed not only includes a mode where the user studies a series of either Chinese or Japanese prompts, but also a mode in which some prompts appear in the opposite language. We consider the combined use of these two methods will improve results. In the future, we hope to construct a vocabulary training system which fits each student’s proficiency level. We also plan to work toward developing a training system which tells students how their vocabulary studies are progressing in real-time as they use it.

References