ENGLISH VOCABULARY LEARNING SUPPORT SYSTEM BASED ON
STOCHASTIC WORD PROFICIENCY MODEL ESTIMATING ELUSIVE WORDS
FOR STUDENTS

Daisuke Kitakoshi1*, Daisuke Kitajima2**, and Masato Suzuki1***
1Tokyo National College of Technology, 2Niigata University
*1220-2 Kunugida-machi, Hachioji, Tokyo JAPAN
*kitakosi@tokyo-ct.ac.jp, **t12i918g@mail.cc.niigata-u.ac.jp, ***suz@tokyo-ct.ac.jp

ABSTRACT

This research focuses on a learning environment for teenage Japanese students to improve their English vocabulary. We propose a system that helps learners (students) improve their English vocabulary by intensively memorizing words by means of a Word Proficiency Model (WPM) represented by a Bayesian network, which is known as one of the stochastic models. A student’s WPM can reflect relationships between features of English words (e.g., part of speech, word length, and so on) and the learner’s ease in learning the words. The proposed system generates a word list arranged in order of difficulty for respective students to learn based on the result of probabilistic inference using the WPM. Intensive learning of the words by means of the list generated by the proposed system improves the acquisition of English vocabulary by helping learners unconsciously memorize words they find difficult.

KEYWORDS

English vocabulary learning support system, Bayesian network, Probabilistic reasoning, Elusive word list, Word proficiency model.

1 INTRODUCTION

In non-English-speaking countries including Japan, EFL (English as a Foreign Language) programs have been given great importance. There has been a great deal of research on efficient methods of learning English, and instructional support for English teachers has been reported in a variety of fields such as pedagogical and educational engineering [1]. Japanese students have many opportunities to study English even after completing their compulsory education. Despite this, many of them have poor English communication skills. One reason for this is the students’ lack of vocabulary. English education in Japan stresses the learning of grammar. After completing compulsory education, students typically have the basic grammatical knowledge they need to communicate with native English speakers. However, since they do not have enough vocabulary to build sentences, they are not good at communicating in English.

In order to make English vocabulary learning more efficient, Hsieh et al. [2] proposed a personalized English article recommending system using accumulated learner (student) profiles to choose appropriate English articles for a user. Chujo et al. [3] attempted to classify vocabulary for learners at various levels by means of statistical measures. The goal of these studies are to improve the English learning efficiency of learners through identifying words and learning materials which are appropriate to the level of the learners.

In this article, we propose a system for improving the English vocabulary of teenage Japanese students. The system is based on the dependencies between word features and the proficiency of the students as extracted by a stochastic model. The proposed system is referred to as EWPM (an English vocabulary learning support system based on Word Proficiency Model). A stochastic Word Proficiency Model (WPM) for each learner (student) expresses the relationships between
the features of words and the “difficulty (or ease) of memorization” of the corresponding words, and its structure is decided by data consisting of the test results of each student and the features of words on the test. The EWPM generates a word list based on how hard they are to remember as determined by the WPM. The proposed system allows learners (students) to effectively learn words by intensively studying specific words that are judged by the WPM as subconsciously hard for these learners to memorize.

2 English Learning Environment in Japanese Compulsory Education

Japanese language has a distinctive system of grammar that is significantly different from the grammar of English and other Romance languages spoken in Western European and North American countries. The acquisition of English has become very important for many Japanese, as it has for non-native speakers around the world. However, most Japanese cannot acquire communication skills in English from their daily life, since they have few opportunities to speak English. While education in English is compulsory in Japan, students don’t have enough time to acquire vocabulary because “instruction in grammar” is generally the prime objective. For these reasons, although most junior high school students acquire some knowledge of English grammar, they lack the vocabulary to construct commonly-used sentence. Many students actually study English for ten years in school, until graduating from college. In addition, in recent years English has been made compulsory in elementary school, beginning in the 5th and 6th grades. However, no significant improvement has been confirmed. This seems to be because of the emphasis in compulsory English education on the learning of grammar rather than vocabulary acquisition. We think that improving vocabulary contributes to the improvement of English proficiency for Japanese. The proposed system supports the learning of “hard-to-memorize” words for students who are poor at learning English, and is therefore expected to improve English vocabulary. Hereafter, the term “elusive words” is used for “English words which a student has difficulty learning (memorizing)”.

3 Stochastic Model Representing Relationships Between English Word Features and Proficiency

This section describes a Bayesian network (BN) employed as a Word Proficiency Model (WPM). A BN is a stochastic model expressing dependencies between random variables (or nodes) [4] [5]. The structure of a BN is expressed in the form of a directed acyclic graph (DAG), and a directed link is drawn between two nodes if strong stochastic dependencies exist between two variables. Probabilities (or parameters) assigned to respective nodes are represented as a Conditional Probability Table (CPT) listing marginal or conditional probabilities corresponding to nodes in the BN. The joint probability distribution in the BN is expressed as the product of the conditional probabilities as follows:

\[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i)) \]

where \( pa(X_i) \) denotes a joint random variable consisting of random variables connected to \( X_i \). In this article, the structure of BN is decided by an algorithm called "Greedy Search"
implemented in BayoNet [6], a support software for structure learning in the BN. This article employs the maximum likelihood estimator (i.e., log likelihood in the field of BN structure learning) as an evaluation criterion for the structure learning. BNs can be applied to a variety of fields such as data mining and knowledge discovery [7] since their structural characteristics are easy to catch visually. In addition, many probabilistic reasoning algorithms using network connections of the BN have been proposed.

4. ENGLISH VOCABULARY LEARNING SUPPORT SYSTEM BASED ON WORD PROFICIENCY MODEL (EWPM)

This section describes the English vocabulary database and word proficiency model as the main components of the proposed system, and then presents a brief outline of the system EWPM.

4.1 English Vocabulary Database

The English vocabulary database in this article is composed of pairs of English words and features corresponding to them, and is used to construct the WPM. The word features (i.e., random variables in the WPM) and their possible states (values) are shown in Table 1. $F$ in this table stands for joint random variable representing a group of feature quantities with respect to words ($f$: value of $F$). The feature “Frequency” means the frequency of appearance of a word in books registered in the Project Gutenberg [8] which is a website to digitize and archive publications such as classic books. “Grade” denotes the grade at which a word appears for the first time in one of the English textbooks for junior high (titled “TOTAL ENGLISH”) authorized by the Ministry of Education, Culture, Sports, Science and Technology (if a word does not appear in the textbook, the value of the random variable “Grade” in the corresponding word equals “none”), and “Consonant clusters” is the maximum number of consecutive consonants (e.g., the value of consonant clusters equals 3 for the word strong). A value of $F$ is expressed as 5-tuple consisting of 5 feature quantities $<\text{Frequency, Part of speech, Word Length, Grade, Consonant clusters}>$ such as $<\text{Middle, Verb, 7, 2nd, 4}>$. The number of target words for learning in the proposed system EWPM is set to approximately 1800. These words are used as the scope of a touch-typing contest [9] for elementary and junior high students. Table 2 represents an example of the English vocabulary database consisting of words and their features.

4.2 Word Proficiency Model (WPM)

A WPM is represented by a BN, described in the preceding section, and has a total of 6 nodes. Five nodes, termed feature nodes, correspond to each of the word features shown in Table 1, and another node is termed answer node.

A Japanese-English word dictation test is conducted in order to collect the data for constructing the WPM. A data in the data set is composed of a pair of word features and whether a student can write the correct or incorrect answer (i.e., word) in the test. The number of data is therefore equal to that of the

<table>
<thead>
<tr>
<th>$F$</th>
<th>$f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>High, Middle, Low, Very low, None</td>
</tr>
<tr>
<td>Part of speech</td>
<td>Verb, Noun, Adverb, Adjective, Other 3 or less, 4, 5, 6, 7, 8, 9 and up</td>
</tr>
<tr>
<td>Word length</td>
<td>1st, 2nd, 3rd, none</td>
</tr>
<tr>
<td>Grade</td>
<td>1, 2, 3, 4</td>
</tr>
<tr>
<td>Consonant clusters</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English word</th>
<th>Meaning (in Japanese)</th>
<th>Part of speech</th>
<th>...</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>猫</td>
<td>Noun</td>
<td>...</td>
<td>Low</td>
</tr>
<tr>
<td>tell</td>
<td>～に話す</td>
<td>Verb</td>
<td>...</td>
<td>High</td>
</tr>
<tr>
<td>strong</td>
<td>強い</td>
<td>Adjective</td>
<td>...</td>
<td>Middle</td>
</tr>
</tbody>
</table>
words on the test. The answer node has two values, “correct” or “incorrect”, which denote whether the student writes a correct or incorrect answer for a word having a certain combination of features $F = f$.

The Bayesian network (WPM) constructed for each student expresses stochastic dependencies between words which the student finds easy (or hard) to memorize and features of the corresponding words. A WPM can therefore characterize some kind of proficiency in English vocabulary for a student since trends among the words that the student has difficulty memorizing can be extracted through probabilistic inference using this model. An example of a WPM corresponding to a student is shown in Figure 1. In this figure, there are 4 links directly connecting with the answer node. The existence of a link between the "Answer" and "Grade" nodes denotes that the accuracy rate for a word in the test depends on the grade in which the word appears for the first time in the textbook (e.g., a student can write a correct answer for a word with high probability if the word appears in the textbook in the 3rd grade). However, the detailed information on the student’s vocabulary proficiency is obtained not only by seeing where links exist, but by investigating the parameters (i.e., CPT) assigned to the corresponding nodes. Using this model, we can estimate elusive words for individual students, and the WPM can thus be employed to master such words.

### 4.3 Process to Support Vocabulary Learning based on EWPM

This section outlines the supporting process for improving the vocabulary of students in the EWPM. The proposed system employs the WPM, described in the preceding section, which can compute a variety of probabilities based on its connecting structure. In order to support word learning for each student, the EWPM computes the following conditional probability using the connection of the student’s WPM:

$$P(Answer = "incorrect" | F = f).$$ (2)

This formula represents the probability with which the student will write "incorrect answers" for words having the combination of feature quantities $F = f$ (described in section 4.1) in the dictation test. Consequently, if the formula (2) has large value, there is a high probability that a student will write a wrong answer when the words having features $F = f$ are on the test.

In this article, we assume that elusive words for a student have common features regardless of whether the student realizes that the words are “hard to memorize” or not, and considers a word in which the value of formula (2) $\geq 0.5$ as an “elusive word for the student”. The conditional probability described in (2) is computed based on the Loopy-BP algorithm [10] implemented in BayoNet. In the EWPM, a word list consisting of the words extracted in descending order of the above-mentioned conditional probability is defined as an “elusive word list” for a student. The word learning process in the proposed system is shown as follows:

i. An elusive word list is generated on the basis of the WPM constructed for a student and reasoning results computed using the WPM;

ii. A certain number of words is selected from the top of the elusive word list as
the “target” elusive word list for the student to learn;
iii. The student studies words in order from the top of the target elusive word list.

Generally, the number of words selected is determined depending on the amount of time spent studying for each student’s word learning. In other words, if the students want to spend a longer time on word learning, the EWPM can extract a greater number of words from the elusive word list as the target list.

5. EXPERIMENTS

Two kinds of experiments were conducted to evaluate the basic characteristics of the proposed vocabulary learning support system using WPM and its learning effects for students. We also verified the adequacy of the WPM. A total of 18 students (18 – 20 years old) at the Tokyo National College of Technology were used as subjects in the experiments.

5.1 Experiment (1)

The aim of this experiment was to evaluate the basic performance of the EWPM. First, all subjects took a Japanese-English word dictation test (number of questions: 100). After finishing this first test, the subjects were divided into four groups by two kinds of criterion:

i. whether the subjects would learn using the target elusive word list made by the proposed system (P) or using the list where the words were randomly arranged (N) towards the second and third tests;
ii. whether the subjects achieved a high score on the test (H) or not (L).

The four groups were termed the PH group (group of subjects satisfying the criteria P and H), PL (satisfying P and L), NH (satisfying N and H), and NL (satisfying N and L), respectively. The numbers of subjects were 3, 2, and 3 for each group. A WPM for each subject in PH and PL was constructed using sample data consisting of the test results (correct/incorrect) and the features of the words in the test.

In the experiments, each subject spent 10 minutes a day memorizing English words, and took a total of three dictation tests. The first test was conducted before starting the word memorization to collect sample data to construct the WPM, and the other tests were conducted every week for 2 weeks in order to evaluate the learning outcomes of each student. The range of questions in the test corresponded to a set $W$ consisting of words which subjects can acquire at a maximum of 1 week ($|W|$ was set to 600 in this experiment). Elements of $W$ were selected out of 1800 words in the touch typing contest [9], since students were supposed to have learned them in junior high. $W$ also played a role as the target set which the subjects were required to learn. The number of words in the list given to the subjects was equal to 600 for all four groups, and the subjects were instructed to study the words in order from the top to the bottom of the list.

The structure of the initial WPM was decided using the data regarding the results of the first dictation test (the number of data = the number of questions), and the first target elusive word list was generated on the basis of the reasoning results using the initial WPM. During the first week, students in PH and PL studied words using the first target elusive word list. During the second week, the students studied the second target elusive word list generated based on the second WPM. The structure of the second WPM was constructed using the data concerning the results of the second dictation test conducted after the end of learning in the first week.

We collected the number of words studied per day during the experimental period for each subject by self-report, and also confirmed the total number of acquired words in two weeks.
with each subject after finishing the third test. The total number of acquired words means the number of words that the subject realized to memorize.

5.2 Experiment (2)

Experiment (2) aimed to verify that the system can appropriately estimate the elusive words for students by means of BNs, employed as a word proficiency model (WPM). The number of subjects equals 18. In preparation for this experiment, a Japanese-English word dictation test (number of questions: 100) was conducted for the subjects independently of the test conducted in experiment (1), while the structures of WPM corresponding to respective subjects were decided based on the test results, as in experiment (1). The subjects were divided for comparison into two groups, one composed of the subjects who scored high on the test (number of correct answers $\geq 51$), and the other composed of the remaining subjects. Nine subjects were in the “high-score group”, and the others in the “low-score group”.

One-hundred questions were prepared for a dictation test to evaluate the characteristics of the WPM corresponding to each subject. The evaluation process was conducted in the following manner ($W$ stood for the set specifying the range of questions in the test, and is identical with the set used in experiment (1)).

1. Six-hundred words in $W$ were classified into 10 sets in order of the level of difficulty based on the value of conditional probability shown in formula (2), computed using the structure of WPM for each subject (60 words for 1 level);
2. One-hundred questions were prepared for the test by extracting every 10 words in a random manner from each word set;
3. Each subject took the test, and the test results were evaluated.

The words in difficulty level 1 were regarded as “the easiest words” for students to memorize, and those in level 10 as the most difficult words.

6 EMPIRICAL RESULTS AND DISCUSSION

This section shows the experimental results for the experiments (1) and (2), and discusses the basic characteristics and performance of the proposed system.

6.1 Results and Discussion for Experiment (1)

First, the effectiveness of the EWPM is discussed. Table 3 lists the average score for the first and third tests in students from the high score (H) and low score (L) groups.

<table>
<thead>
<tr>
<th>Subjects’ score</th>
<th>First</th>
<th>Third (students belonging to P)</th>
<th>Third (students belonging to N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>28.4</td>
<td>62.5</td>
<td>47.7</td>
</tr>
<tr>
<td>L</td>
<td>52.8</td>
<td>78.7</td>
<td>63.5</td>
</tr>
</tbody>
</table>

It is not surprising that the average scores for the first test in both groups were low, and the scores for the third test were improved in comparison due to the efforts of subjects. However, in the third test, the scores of the subjects belonging to group P (i.e., both PH and PL) were approximately 15 points greater than those for group N (i.e., both NH and NL). To discuss the properties of the learning method characterized by the proposed system,
transitions of the scores of three tests for all 4 groups are compared (Figure 2). As shown in this figure, the increases of scores between the first and second test for both PH (thick solid line) and PL (dotted line) were larger than that for NH (thin solid line with marker “*”), and also were similar to that for NL. Furthermore, the slopes of curves for PH and PL between the second and third test were sharper than those between the first and second test, and also were sharper for both the groups NH and NL, in which the students studied a “randomly arranged word list”. These results were caused by the characteristics of the target elusive word list made by EWPM. In the target elusive word list, since a certain number of words (600 in the experiment) was arranged in order of “difficulty to memorize” estimated by the conditional probability (2), the difficulty of memorizing a word is considered to decrease as the position of the word is located lower in the list (i.e., as the student proceeds with word learning). As a result, the learning speed improves with learning time and the improvement of learning effectiveness then contributes to the significant increase in the score on the third test. Table 4 also shows the total number of learning (studying) words and the total number of acquired words that the subjects declared in each of the groups during the experiment. This table shows that the number of acquired words for the groups using the target elusive word list was larger than that for the groups using the random word list, although the subjects in NH and NL actually declared they would study a greater number of words than the subjects in PH and PL. This result shows that the learning method using the target elusive word list generated by the proposed method is effective for the improvement of English vocabulary.

Second, the WPMs structured from the results of the three tests for all four groups were compared to evaluate the effectiveness of the learning method using the proposed system. For ease of comparison, we paid attention to only the feature nodes that had a direct connection (i.e., link) with the answer node. In each WPM for each test, the conditional probability \( P(\text{Answer} = \text{"correct"}|F = f) \), the accuracy rate of words with the feature \( F = f \), was computed (e.g., \( P(\text{Answer} = \text{"correct"}|\text{Frequency} = \text{"low"}) \)). Table 5 lists the occupation rates of word features (and values) of the words students found easy to memorize (i.e., \( P(\text{Answer} = \text{"correct"}|E = e) \geq 0.50 \)) for all combinations of features and their values. This table shows that the occupation rates of the features of “easy-to-memorize” words increased with the number of tests conducted for all subject groups; in other words, many students were able to write correct answers for any word regardless of its features. Moreover, for both the high score and low score groups, the occupation rates for group P (using the target elusive word list generated by the proposed method) were higher than those for group N (using the random word list), with one exception (i.e., the occupation rates for NL and PL in the first test). For these reasons, it was determined the learning method characterized by EWPM, in which students intensively studied the words that are hard to memorize, was effective for improving their English vocabulary. Because the students become proficient in English vocabulary regardless of the word features, the structures of WPM (BN) changed depending on the “proficiency of the students”. As a typical result, Figure 3
illustrates the WPMs which had structures determined using the sample data of (i) the results in the first test and (ii) those in the third test, for a subject in the NH group. There were 4 links between the answer node (circle colored gray) and the feature nodes in (i) while the number of links was equal to 2 in (ii). This result means that the stochastic dependencies between the “easy-to-memorize” words and the features of corresponding words weaken as the student proceeds with word learning.

Finally in this section, we focus on the features where the accuracy rates in the first and second tests were low (i.e., $P(Answer = "correct" | F = f) \leq 0.50$), and then compare the increment of accuracy rates for the word having these features from the first to the third test (Table 6). As shown in this table, the increments in the groups using the target elusive word list based on the EWPM (PL and PH) were larger than those in the groups using the random word list (NL and NH). In addition, it is noteworthy that the increment in the high score group (NH and PH), where the students already have a certain level of vocabulary, was larger than the increment in the low score group (NL and PL).

### 6.2 Results and Discussion for Experiment (2)

In this section, we discuss the expressive ability of the WPM in terms of stochastic dependencies between the individual students’ “difficulty (or ease) of memorizing” words and the features which the corresponding words have. Figure 4 depicts the transitions in the average number of correct answers in the dictation test per the difficulty levels estimated by WPM. This figure illustrates that the number of correct answers in the high score group (vertical-striped bar) is larger than that in the low score group (diagonal-striped bar). Moreover, the number of correct answers for both groups decreases with the difficulty levels. As for the regression lines for both groups, also plotted in Figure 4 (high score group : solid line, low score group : dashed line), we can see that there are negative correlations between the difficulty levels for each student specified by WPM and the number of correct answers for the student, since the slopes of both lines were negative. These results thus showed that the WPM can appropriately express the relations between the difficulty (or ease) of memorizing English words and the features of the corresponding words, and can also specify and then present elusive words for the students by using these relations. Both subject groups had distinctive subjects in terms of their test scores. The high score group had a subject who was able to write correct answers for almost all questions in the test regardless of difficulty level, and the low score group similarly had a
subject who could not write correct answers for almost any of the questions. The reason the slopes of these two lines are similar is probably because of subjects at these extreme end of the spectrum makes the slopes in both the high score and low score groups approach to 0. Consequently, to evaluate the performance and characteristics of EWPM more adequately, we need to prepare a set that contains a wider variety of words, including much harder words and much easier ones, so that the excellent students will get some incorrect answers and the underachievers will get some correct answers.

7 CONCLUDING REMARKS

This paper proposed an English vocabulary learning support system based on a word proficiency model (EWPM). The structure of the WPM (BN) is decided by means of sample data consisting of the results of a word dictation test and the features of the corresponding words. Empirical results showed that the English word learning method provided by the proposed system allows students to improve their vocabulary, and also confirmed that the WPM expressed the characteristics of respective students regarding stochastic dependencies between "difficulty (ease) to memorize" for words and features of the words. When a student keeps working at English word learning, structural and probabilistic properties in the WPM changes regardless of whether he (she) succeeds in improving vocabulary or not. We think that a more efficient way to apply the WPM to vocabulary improvement will be found by investigating the relationship between the features of WPM for each student and the vocabulary level of the student. Therefore, further investigations of these relationships are needed. In addition, since there are actually many methods to improve English vocabulary, it will be necessary to compare the performance of the EWPM with the other existing word learning approaches.

REFERENCES