Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing

Yo Ehara
Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology
2-4-7 Aomi, Koto-ku, Tokyo, Japan
i@yoehara.com, y-ehara@aist.go.jp

Abstract

We introduce a freely available dataset for analyzing the English vocabulary of English-as-a-second language (ESL) learners. While ESL vocabulary tests have been extensively studied, few of the results have been made public. This is probably because 1) most of the tests are used to grade test takers, i.e., placement tests; thus, they are treated as private information that should not be leaked, and 2) the primary focus of most language-educators is how to measure their students' ESL vocabulary, rather than the test results of the other test takers. However, to build and evaluate systems to support language learners, we need a dataset that records the learners' vocabulary. Our dataset meets this need. It contains the results of the vocabulary size test, a well-studied English vocabulary test, by one hundred test takers hired via crowdsourcing. Unlike high-stakes testing, the test takers of our dataset were not motivated to cheat on the tests to obtain high scores. This setting is similar to that of typical language-learning support systems. Brief test-theory analysis on the dataset showed an excellent test reliability of 0.91 (Chronbach's alpha). Analysis using item response theory also indicates that the test is reliable and successfully measures the vocabulary ability of language learners. We also measured how well the responses from the learners can be predicted with high accuracy using machine-learning methods.

Keywords: Vocabulary Test, Item Response Theory, Crowdsourcing

1. Introduction

Supporting someone involves enabling him/her is overcoming a difficulty. Thus, we first need to detect what he/she has difficulty with. This holds true for supporting second language learners: we first need to detect what they have difficulty with (Ehara et al., 2010; Ehara et al., 2012; Ehara et al., 2013; Ehara et al., 2014; Ehara et al., 2016). To detect such difficulties, we need to test them. The knowledge of a language is too large, and testing all types of knowledge imposes a heavy burden on them. This is where vocabulary tests come into play. Vocabulary tests measure language learners' knowledge of words, especially their meaning. Compared to other aspects of language, such as syntax, learners' vocabulary can be easily measured and can be used to understand what kind of difficulties that language learners face: the largeness of vocabulary, or a set of words, is easily measured due to its size. Its elements, or words, are also countable. The difficulty of words is roughly measured by their frequency in a corpus: the rarer a word, the more difficult it may be to learners. Most importantly, whether a learner knows a typical meaning of a word can be easily measured through multiple-choice questions, the results of which (i.e., learners' responses to the test) are machine-readable and can be easily automatically scored if the correct answers of the test are known. Therefore, the results of a vocabulary test given to a language learner provide essential information to support his/her language learning. This is why this type of test is frequently used as a placement test from which language learners are categorized into a "class" with similar difficulties in using the second language that they are learning.

Although the results of vocabulary tests are important for supporting language learners, few datasets are, however, available for building and evaluating language-support systems. This is probably because 1) most of the tests are used to grade test takers, i.e., placement tests; thus, they are treated as private information that should not be leaked, and 2) the primary focus of most language-educators is how to measure their students' ESL vocabulary, rather than the test results of other test takers.

In this paper, we introduce a freely available dataset for analyzing English vocabulary of English-as-a-second language (ESL) learners. In our dataset, 100 learners answered 100 well-tested vocabulary questions. Unlike a typical setting in which test takers are in classrooms of a language-learning course, we employed 100 learners via crowdsourcing, which means they were paid.

The test results in which test takers are paid are more suitable for a dataset to be used in building and evaluating language-support systems than those in which test takers are not paid. Vocabulary tests can be high-stakes tests when they are used for placement: test-takers have high stakes if they obtain good scores in the test, so they have strong motivation to obtain high scores. In a high-stakes test, language learners tend to pretend to have knowledge on the second language. Vocabulary tests as a high-stakes test are common as placement tests, but are not common in language-support systems because users know that they may be provided with incorrect support if they pretend. Thus, our dataset is more similar to a realistic environment in which language learners do not have motivation to exaggerate their knowledge of second language vocabulary.

We conducted a brief statistical analysis on our dataset, which showed that it is effective in terms of measuring ability. We also measured how accurately the responses from the learners can be predicted using machine-learning methods.

2. Related Work

Many previous papers have reported on the *analysis* vocabulary-test data using datasets not publicly available. Typically, one hundred to several hundred participants participated in the vocabulary tests in these studies. (Culligan, 2015) compared three typical test formats and showed that all three have high reliability and measure the vocabulary ability of language learners. Their results are based on vocabulary-test data of 54 words collected from 167 university students.

Recently, large experiments on an English vocabulary test were conducted with over 1,000 test takers and hundreds of words (Webb et al., 2017); however, their raw test-result data are not publicly available. This large number of test takers was provided by language teachers worldwide.

To the best of our knowledge, no vocabulary-test result datasets is publicly available. This is presumably because the data are usually collected from language classes on a volunteer basis. This method of collecting vocabulary-test results may be beneficial for classroom teaching because the environment under which a dataset is taken is a classroom and the applications to which the dataset are used are also for classrooms. However, this is not the case for developing educational software, in which participants are more diverse than in typical classrooms.

The vocabulary dataset by Ehara et al. (2010) is a publicly-available vocabulary knowledge data [1]. This dataset collects the results of $12,000$ word questions by 16 people. Notable differences between their dataset and our dataset are: 1) theirs were collected in a self-report manner, thus, strictly speaking, their dataset is not "test results". 2) the number of test-takers are 100 in our dataset while theirs are merely 16. Thus, our dataset is more accurate for small size of vocabulary while theirs focus on the size of vocabulary to be tested with sacrificing accurateness.

3. Dataset

The purpose of building our dataset was to use it for building and evaluating language-support systems. Such systems are used by a diverse population of users. This setting is very different from that of testing in classrooms of a language course in which the characteristics of test takers are usually limited to a category, for example, university students of the same year. To meet the purpose of this dataset, we employed a diverse population of test takers via crowdsourcing.

The data for the dataset were collected via a crowdsourcing service called Lancers, one of the major crowdsourcing companies in Japan. We used the vocabulary size test (VST) (Nation and Beglar, 2007) for this dataset. This test was designed to measure the vocabulary size of each learner. In this test, test takers are asked to answer 100 vocabulary questions. Each question has four options and only one of the options is correct. This means that each learner has a 25% chance to answer each question correctly regardless of their vocabulary knowledge. We employed 100 test takers. We paid each test taker 383 Yen (approximately 3.5 USD). An example question in the vocabulary size test (Nation and Beglar, 2007) is as follows:

microphone: Please use the <microphone>.

a machine for making food hot

b machine that makes sounds louder

c machine that makes things look bigger

d small telephone that can be carried around

To compare the test results to other test results, we limited the test takers to those who had previously taken the TOEIC test [2], which is a popular English proficiency tests in Japan. The choice of this test as a reference is simply because of the number of test takers and the internationality of the test. For example, although we know that the TOEIC test is popular in only certain Asian countries including Japan and other tests such as TOEFL [3] are more universally popular, we chose TOEIC as a reference because, we cannot collect enough test-takers on any crowdsourcing service popular in Japan. We did not include English proficiency tests popular almost solely in Japan, such as the "Eiken" test, an English proficiency test popular with Japanese high-school students While we required our test takers to have a TOEIC test score, it did not matter when they took the test. The reason of this is also not to limit the diversity of test takers. Since many people do not take a language-proficiency test after they graduate from universities, it is easily speculated that limiting the time when the learners took a proficiency test would lead to limiting the diversity of test takers.

4. Analysis using Test Theory

4.1. Notation

First, let us introduce some notations. Let us consider the case in which test takers respond to problems. Problems to be solved are called *items* in psychology or psychological statistics. Let the set of test takers (learners) be I and the set of problems be J. Each test taker $i \in I$ answers item $j \in J$. Each problem can be scored in binary format: let y_{ij} be how i answers j: $y_{ij} = 1$, i.e., correctly or $y_{ij} = 0$, i.e., incorrectly.

4.2. Chronbach's Alpha

Chronbach's alpha is a measure of *test reliability*. Test reliability implies whether the test can be used to measure uni-dimensional hidden characteristics of test takers to be estimated, which we usually call "ability". Another interpretation of Chronbach's alpha is that it measures *internal consistency*. That is, it measures how well items that can be used to measure similar characteristics of test takers would result in similar response patterns of test takers.

Using the notations explained above, t is defined as follows:

$$\alpha = \frac{|J|}{|J|-1}\left(1 - \frac{\sum_{j=1}^{|J|} P_j(1-P_j)}{\sigma_X^2}\right) \qquad (1)$$

[1] http://yoehara.com/esl-vocabulary-dataset/

[2] https://www.ets.org/toeic

[3] https://www.ets.org/toefl

where σ_X is the variance of all responses, and P_j represents the proportion of correct answers to the item (i.e., problem) j.

In our dataset, the Chronbach's alpha was 0.91, which is regarded as "excellent" (George, 2011; Kline, 2013; DeVellis, 2016). This means that our dataset is highly reliable.

4.3. Item Response Theory

Item response theory (IRT) (Baker and Kim, 2004) is usually used for analyzing test-result data including language-test data. Item response theory is used like a name of a field rather than a specific models. However, a two-parameter model (**2PL**) and one-parameter model (**1PL**) are frequently used for analysis. With both models, it is assumed that each problem is independent: whether a test taker answered a problem correctly has no influence on whether he/she answered the other problems correctly. The **1PL** is sometimes called the Rasch model (Rasch, 1960).

The **2PL** is a generalization of **1PL**. It models the probability that test taker i correctly responds to a problem j in the following equation. Let σ denote the logistic sigmoid function, i.e., for $t \in \mathcal{R}; \sigma(t) = \frac{1}{1+\exp(-t)}$.

$$P(y_{ij} = 1|i, j) = \sigma\left(a_j(\theta_i - b_j)\right) \quad (2)$$

The model has two item parameters a_j and b_j, and one test-taker parameter θ_i, which is called the *ability* parameter and denotes the ability of i. Parameter b_j is called the *difficulty* parameter and denotes the difficulty of j. Since the logistic sigmoid function is a monotonously increasing function, the larger θ_i is, the more likely that i will correctly answer j. On the other hand, the larger b_j is, the less likely that i will correctly answer j. Since $\sigma(0) = 0.5$, i is more likely to respond correctly to j if and only if $\theta_i > b_j$ and vice versa.

Parameter a_j is called a *discrimination* parameter and has a more complicated definition. Briefly speaking, it shows how well j discriminates low-ability test takers from high-ability test takers. In other words, if a_j becomes larger, the difference in the probability to respond correctly between high and low test takers becomes larger.

Equation 2 can be graphically drawn. The curve showing probability (i.e., y-axis) against θ_i (i.e., x-axis) is called an *item characteristic curve* (ICC). It shows how difficult a j is for an i whose ability is the value on the x-axis. We show an example of the ICCs for some of the calculated words in Figure 1.

4.4. Test Information Function

A test information function is an important concept in IRT. It shows how reliable a test is in the form of a function against *ability*. Intuitively, a test is unreliable for test takers too skilled or un-skilled for whom the test is designed. For example, an elementary school math test cannot reliably measure the abilities of computer-science math students. Therefore, the informative-ness of the test results can be given as a function of the ability of test takers. This is the underlying idea behind the test information function. Figure 2 shows the test information function of our dataset

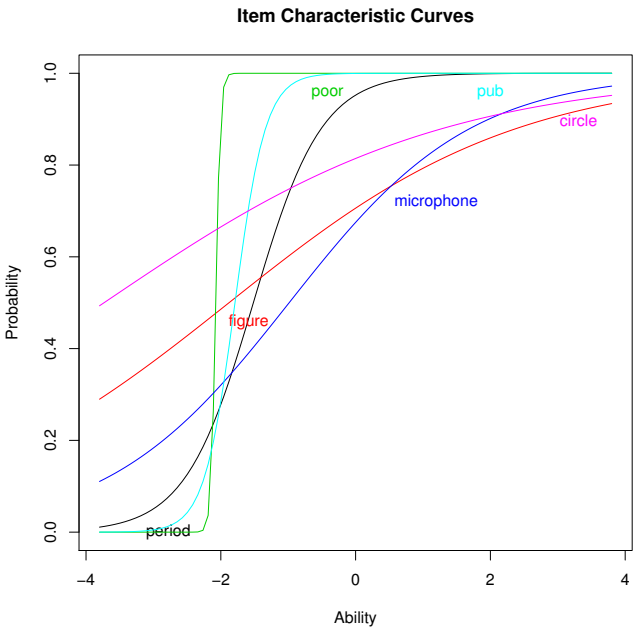

Figure 1: Examples of ICCs in our dataset

[4] We can see that it contains information in a wide range of abilities.

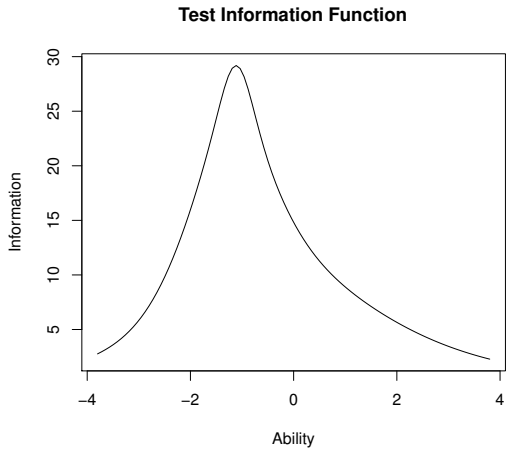

Figure 2: Test information function of our dataset

4.5. Evaluation of dataset's reliability using TOEIC scores

In this dataset, we limited test takers to those who had previously taken the TOEIC test. The TOEIC test thoroughly examines test takers's ability. It takes about 2 hours to take the test; thus, it imposes a heavy burden on test takers. The TOEIC test is mainly used as a proficiency test of English in Japan and Korea. We evaluated the reliability of our

[4] When drawing Figure 2, we removed four words that were outliers in the data, namely, "poor", "pub", "octopus", and "puma".

dataset by analyzing the correlation between the test-takers' θ_is and their TOEIC scores.

In Figure 3, we show the TOEIC score against the calculated θ_i of each test-taker. A linear-regression analysis shows that θ_i is a good estimator of TOEIC score. A TOEIC score is estimated by $86.50 \times$ ability parameter + 703.08 ($p < 0.001$).

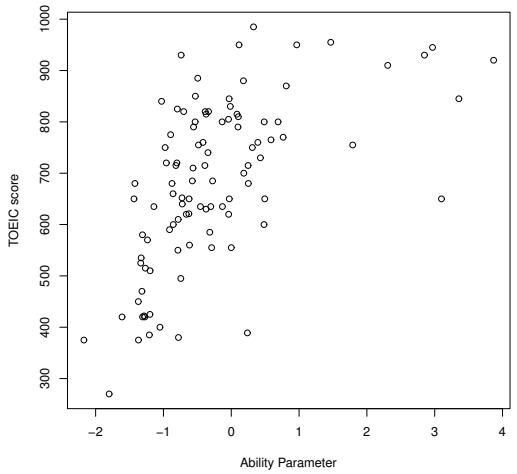

Figure 3: TOEIC scores against calculated ability parameters

A high correlation with the TOEIC score also decreases test-taker burden: taking the VST we used takes about 20 minutes. This high correlation means that we can practically measure learners' English proficiency over the Web using only 20 minutes of vocabulary testing.

5. Prediction Experiment

A developer typically needs to know the difficulty level of many words to develop educational software. For example, to support readers, the reading support system by Ehara et al. (2013) requires the difficulty parameters of the words appearing throughout a given text.

To obtain the values of the difficulty and discrimination parameters of the words in a large vocabulary set, however, testing all the words in the set imposes too much burden on the test takers, who are potential users of educational systems. A workaround for this problem is as follows: we first calculate the values of difficulty parameters and discrimination parameters for a small vocabulary set using IRT. Then, using these calculated parameters as training data, we regress the parameters with features such as *word embeddings*, or vector representations of words, to predict the parameters of the words outside this small vocabulary set.

We conducted a prediction experiment. Among the 100 words used in the vocabulary test, 92 were available in Wikipedia. We further divided these 92 words into 70 words for training and 22 words for testing. We prepared word embeddings using the word2vec toolkit (Mikolov et al., 2013) over the entire English Wikipedia. We used skip-gram as the algorithm and the size of the vectors was set at 300. To predict the parameters from the word embeddings, we applied support vector regression (SVR) with a linear and radial basis function (RBF) kernel.

As a result, the difficulty parameters were predicted with a root mean squared error (RMSE) of 3.632, and the discrimination parameters were predicted with an RMSE of 1.020 using SVR with a linear kernel. Presumably, due to the small training dataset, the SVR with an RBF kernel was overfitted, and its RMSE scores were worse than those in the case when linear kernels were used. Considering that difficulty values typically range from -5.0 to 5.0 and discrimination values range from 0.0 to 5.0, our results suggest that predicting difficulty and discrimination only from word embeddings without using data from learners is difficult. These results also suggest that our dataset is valuable because important statistics calculated from the data cannot be easily predicted from typical word features such as word embeddings.

6. Conclusion

We introduced a dataset of vocabulary-test results. Our dataset is freely available to the public and has high reliability. The calculated vocabulary ability has high correlation with the TOEIC English proficiency test ($p < 0.001$). We analyzed the reliability of our dataset using Chronbach's alpha and IRT.

We also conducted an experiment to predict the IRT parameters using word embeddings. Future work includes making this prediction more reliable.

7. Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15K16059.

8. Bibliographical References

Baker, F. B. and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques*. Marcel Dekker, New York, second edition.

Culligan, B. (2015). A comparison of three test formats to assess word difficulty. *Language Testing*, 32(4):503–520.

DeVellis, R. F. (2016). *Scale development: Theory and applications*, volume 26. Sage publications.

Ehara, Y., Shimizu, N., Ninomiya, T., and Nakagawa, H. (2010). Personalized reading support for second-language web documents by collective intelligence. In *Proceedings of the 15th international conference on Intelligent user interfaces (IUI 2010)*, pages 51–60, Hong Kong, China. ACM.

Ehara, Y., Sato, I., Oiwa, H., and Nakagawa, H. (2012). Mining words in the minds of second language learners: learner-specific word difficulty. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, December.

Ehara, Y., Shimizu, N., Ninomiya, T., and Nakagawa, H. (2013). Personalized reading support for second-language web documents. *ACM Transactions on Intelligent Systems and Technology*, 4(2).

Ehara, Y., Miyao, Y., Oiwa, H., Sato, I., and Nakagawa, H. (2014). Formalizing word sampling for vocabulary prediction as graph-based active learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1374–1384, Doha, Qatar, October. Association for Computational Linguistics.

Ehara, Y., Baba, Y., Utiyama, M., and Sumita, E. (2016). Assessing translation ability through vocabulary ability assessment. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI*, pages 3712–3718.

George, D. (2011). *SPSS for windows step by step: A simple study guide and reference, 17.0 update, 10/e*. Pearson Education India.

Kline, P. (2013). *Handbook of psychological testing*. Routledge.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119.

Nation, P. and Beglar, D. (2007). A vocabulary size test. 31(7):9–13.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.

Webb, S., Sasao, Y., and Ballance, O. (2017). The updated vocabulary levels test. *ITL - International Journal of Applied Linguistics*, 168(1):33–69.