Correlation Between an Entropy Based Measure and English Language Learner Proficiency

Brendan Flanagan  
Graduate School of Information Science and Electrical Engineering, Kyushu University  
Fukuoka, Japan  
b.flanagan.885@s.kyushu-u.ac.jp

Sachio Hirokawa  
Research Institute for Information Technology, Kyushu University  
Fukuoka, Japan  
hirokawa@cc.kyushu-u.ac.jp

Abstract—It is important for education systems to analyze and provide an appropriate level of support to meet the needs of learners. An example of this is how the effectiveness of automatic language learner error detection and correction can vary depending on the learner’s proficiency level. Covering a wide range of language complexity makes the task of error detection difficult. By predicting the learner’s proficiency level, different error models can be applied for different proficiency levels. In this paper, we propose a measure based on the frequency of words in the sentences produced by learners during speaking exams to predict the learner’s language proficiency. The proposed measure is compared to the learner’s vocabulary size by correlation analysis. The results suggest that there is a stronger correlation between the proposed measure and the proficiency of the learner than the learner’s vocabulary size.

Keywords—Learner Proficiency; Proficiency Prediction; Speaking Errors.

I. INTRODUCTION

Foreign language learners at different levels of proficiency are faced with different needs and problems. It is important to provide appropriate support and feedback that matches these needs. In a traditional classroom environment a teacher would estimate the progress and proficiency of the learner and provide suitable support. However, as language learning increases due to globalization and the use of the Internet as a multi-national multi-lingual platform, the demand for language teaching outpaces the supply and availability of such services.

In the last decade or so, the use of online language learning services has increased, with services being provided in the form of SNS (Social Networking Service) platforms, which bring together language learners from all around the world. These SNS allow learners from different native language backgrounds to engage in language exchange by correcting each other’s writings. While this method has its benefits, autonomous learners are not able to measure their proficiency and receive appropriate support to meet their learning needs. In recent years, the research of error detection and correction, and other automated support have increased in an attempt to fill this void.

In previous research, we have investigated the automatic detection of foreign language writing errors on language learning SNS [1]. However, as the proficiency level of learners on these SNS is often broad, it is difficult to determine errors, as machine classification has to deal with a wide range of writing complexity and increases the chance of false positive classification. Prediction of a learner’s foreign language characteristics could be analyzed to provide tailored tools that can focus on the particular errors, support, and feedback needs of learners at specific language proficiency levels.

In this paper, we propose a measure based on the entropy of word occurrences in the sentences of learner discourse during a speaking exam. The measure is then compared to the vocabulary size and also the entropy of word occurrences at the learner level, to examine which has a stronger correlation with the learner’s language proficiency.

II. RELATED WORK

Previous research on the prediction of foreign language proficiency has focused on a number of different approaches, including: lexical, grammatical, and errors and fluency of discourse.

Supnithi et al. [2], analyzed the vocabulary, grammatical accuracy and fluency features of learners in speaking exams. These features were then used to train Support Vector Machine and Maximum Entropy machine-learning algorithms to automatically predict the proficiency level of the learner. Vocabulary features, such as: bi-grams, words expressed by both the examiner and learner, words only expressed by the learner, and words from a list of twelve different levels of proficiency. Maximum prediction accuracy of 65.57% was achieved using an SVM classifier. In this paper, we analyzed the same corpus and propose a different measure that could be used in the prediction of learner proficiency.

Crossley et al. [3], examined the importance of different lexical features that could be analyzed to create a model of learner proficiency. Human raters based on standardized lexical criteria evaluated a corpus of 240 foreign language writings. It was reported that lexical diversity, word hypernymy values and content word frequency accounted for around 44% of the variance in the lexical proficiency evaluations. In further research, Crossley and McNamara [4] further developed their model of predicting learner proficiency by incorporating features relating to cohesion and linguistic sophistication. They argue that learners with high proficiency don’t necessarily produce writing with more cohesion, but
instead use less frequent and familiar words to increase lexical diversity.

Other research has analyzed learner corpora to extract features that identify characteristics of certain proficiency levels. Yoon et al. [5], investigated the distribution of syntactical patterns in the form of parts of speech (POS). A large learner corpus that had been classified into four different levels of proficiency was parsed to extract POS tags, which were then indexed to create vector space models. The cosine similarity of the test vectors and corpus vectors were then compared. The proficiency prediction was based on the proficiency of the most similar corpus vector. Zechner et al. [6], analyzed 1,400 speaking tests using automatic speech recognition and feature extraction for fluency, pronunciation, prosody, and grammatical accuracy. Different linear regression models were built for each of the 21 speaking items in the test and were used to predict the proficiency level of the learner. Their system achieved a correlation of 0.73 with the human rater scores. In this paper, we analyze a corpus of transcribed speaking test without extracting features concerning the production of utterances, and propose a measure for the prediction of learner proficiency.

III. DATA

The data analyzed in this paper is based on a collection of recorded oral proficiency interview exams conducted as a part of the ACTFL English Standard Speaking Test (SST) [7]. This corpus is commonly known as the National Institute of Information and Communications Technology Japanese Learner English (NICT-JLE) Corpus and is made up of transcripts from 15 minute speaking exams. There are nine different proficiency levels in the SST exam, with level 1-3 representing elementary proficiency, level 4-8 as intermediate, and level 9 representing learners who have advanced proficiency. Professional examiners determined the SST proficiency level grade for each exam. This provides a reliable insight into the proficiency level of the learner, as opposed to other corpora that rely on learner experience, such as: length of study [8]. The corpus is split into two main sets of tagged data: learner original, and learner error tagged transcriptions. Error tagged transcripts of the same learner were also included in the learner original dataset, and we removed duplicates across the two datasets. A total of 1114 original learner transcriptions were analyzed to build an index upon which a system was constructed using GETA1. The transcripts are marked up with a custom tag set that includes non-lexical tags associated with discourse events such as: long pauses, non-verbal sounds, etc. The transcripts also contain the dialog spoken by the interviewer in the exam. The information provided by these tags was not used for analysis in this paper. The transcripts were preprocessed to remove non-lexical information and dialog by the interviewer. Each of the learners utterances were indexed as individual documents within the search engine, and tagged with the SST proficiency level as provided in the header of the transcripts.

IV. CORRELATION BETWEEN PROFICIENCY AND LEARNER TRANSCRIPT CHARACTERISTICS

A. An Entropy like measure of Language Learner Transcripts

In 1948, Shannon [9] introduced the theory of information entropy to determine the average amount of information contained in an event. In this paper, we propose that a measure based on the entropy of learner transcripts can be used in the analysis of learner proficiency. We propose that the information entropy formula in Equation 1 can be used to calculate the information in the transcript of a learner’s exam.

\[
E(u_i) = - \sum_{j \in W} P(w_j, u_i) \log P(w_j, u_i)
\]  

(1)

Where \( u_i \) represents a learner, \( W \) is the set of all words, and \( w_j \) represents a word contained within the corpus. In Shannon’s theory, \( P(w_j, u_i) \) is the probability of occurrence of the word \( w_j \).

\[
P(w_j, u_i) = \frac{tf(w_j, u_i)}{\sum_{k \in U} tf(w_k, u_i)}
\]  

(2)

The formula in Equation 2 would usually be used to calculate this probability, where \( tf(w_j, u_i) \) represents the occurrence frequency of the word \( w_k \) in the transcript of learner \( u_i \).

![Fig. 1. Learner proficiency versus entropy for all learner transcripts.](image)

A scatter plot of the entropy of the transcripts for each learner versus the SST proficiency level are shown in Figure 1. A relation between the proficiency and entropy exists, however it is broad. We propose an alternate formula for as seen in Equation 3 for the calculation of this term.

\[
P(w_j, u_i) = \frac{df(w_j, u_i)}{|U|}
\]  

(3)

1 http://geta.ex.nii.ac.jp/
Where $d_f(w_j, u_i)$ is the number of sentences in the transcript of learner $u_i$ which contain the word $w_j$, and $U$ represents the set of all learner transcripts in the corpus.

As seen in Figure 3, the vocabulary size of learner transcripts increases steadily as the learner’s proficiency (SST level) increases until SST level 6, at which point the increase is not as great. It suggests that vocabulary size is a strong determiner of proficiency from elementary to intermediate levels. However, at higher proficiency level the use of similar size vocabularies might have an affect on the perceived proficiency level scored.

We examined the differences in word usage for learners with SST levels from 4 to 9 by analyzing the corpus using a part of speech parser, TreeTagger [10], to divide the vocabulary into subsets. The vocabulary size and proposed measure was calculated for each of these subsets. These were then analyzed to determine the strength of the correlation between the POS subsets and SST proficiency. Scatter plots of SST Level versus vocabulary size of the top 9 POS subsets are shown in Figure 4. It should be noted that the granularity is coarse because the total count of some POS subsets is small, and therefore it increases the possibility that multiple results occur in the same position in the graph.

In Figure 5, scatter plots of our proposed measure versus SST level show a finer level of granularity when compared to the vocabulary size plots.
C. Correlation Analysis

The Pearson product-moment correlation coefficient can be used to measure the linear correlation between two variables. In this section, the correlation between learner proficiency and vocabulary size, entropy, and our proposed measure are compared.

<table>
<thead>
<tr>
<th>SST Level</th>
<th>Entropy</th>
<th>Proposed Measure</th>
<th>Vocabulary Size</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – 9</td>
<td>0.8021619</td>
<td>0.843445</td>
<td>0.840757</td>
<td>1114</td>
</tr>
<tr>
<td>4 – 9</td>
<td>0.7464927</td>
<td>0.7816754</td>
<td>0.7810395</td>
<td>890</td>
</tr>
</tbody>
</table>

In Table 1, the correlation coefficient $r$ of all SST proficiency levels is higher than that of the intermediate and advanced levels. This is most likely due to greater variation in word usage rather than vocabulary at higher proficiency levels.

<table>
<thead>
<tr>
<th>Parts Of Speech</th>
<th>Proposed Measure</th>
<th>Vocabulary Size</th>
<th>Non-zero Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN (preposition/subord. conj.)</td>
<td>0.7448775</td>
<td>0.6893652</td>
<td>890</td>
</tr>
<tr>
<td>DT (determiner)</td>
<td>0.7150026</td>
<td>0.5642325</td>
<td>890</td>
</tr>
</tbody>
</table>

The Pearson correlation coefficient $r$ for each of the relations is shown in Table 2. Both the proposed measure and vocabulary size data contained 890 sample pairs, except transcripts that did not contain a particular POS tag. All correlations are significant at $p < 0.01$, except for the vocabulary size of the part of speech “TO” which is marked in red. The table is sorted by strongest correlation to weakest, with the strongest correlation for each part of speech bolded. The correlation between the proposed measure and the learner’s proficiency is stronger than vocabulary size for the majority of parts of speech. This confirms that the proposed measure of transcripts is a stronger indicator of learner proficiency than vocabulary size for SST proficiency equal to or higher than level 4.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a measure based on the entropy of the sentence occurrence frequency of words in transcripts of English speaking proficiency exams. The proposed measure was compared with the vocabulary size and entropy of the same transcripts. It was found that the proposed measure has a stronger correlation with SST learner proficiency than both vocabulary size and entropy. The correlations were then compared on parts of speech subsets. It was found that the proposed measure has a stronger correlation with proficiency in a majority of subsets. In future work we will undertake a
comparison of prediction with other speaking and writing learner corpora, and assess it usefulness in the enhancement of learner error detection.

REFERENCES


