Classification of Speaking Proficiency Level by Machine Learning and Feature Selection

Brendan Flanagan[1], Sachio Hirokawa[2], Emiko Kaneko[3] and Emi Izumi[4]

[1] Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan
b.flanagan.885@s.kyushu-u.ac.jp
[2] Research Institute for Information Technology, Kyushu University, Japan
[3] Center for Language Research, Aizu University, Japan
[4] Center for General and Liberal Education, Doshisha University, Japan

Abstract. Analysis of publicly available language learning corpora can be useful for extracting characteristic features of learners from different proficiency levels. This can then be used to support language learning research and the creation of educational resources. In this paper, we classify the words and parts of speech of transcripts from different speaking proficiency levels found in the NICT-JLE corpus. The characteristic features of learners who have the equivalent spoken proficiency of CEFR levels A1 through to B2 were extracted by analyzing the data with the support vector machine method. In particular, we apply feature selection to find a set of characteristic features that achieve optimal classification performance, which can be used to predict spoken learner proficiency.

Keywords: Foreign Language Proficiency; Machine Learning; Feature Selection; SVM

1 Introduction

At present there are many machine readable data that are publicly available, and this has increased the application of machine learning to the task of supporting language learning. In this paper, we analyze the NICT-JLE corpus [1] to investigate which words describe and discriminate different speaking proficiency levels by applying a method of machine learning called SVM (Support Vector Machine) to the classification task. The corpus consists of 1280 transcribed recordings of the Standard Speaking Test[1–3] (herein referred to as SST) English language learner exam. Each exam contains 3 different tasks and the transcriptions are made up of the dialogue between the examiner and examinee. The proficiency level for each examinee was determined by an expert examiner and ranked on a scale from 1 to 9, from beginner to advanced respectively. In this paper, the focus of the classification analysis will be on the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) (Council

[1] http://alaginrc.nict.go.jp/nict_jle/index_E.html

Table 1. Equivalent levels of CEFR, CEFR-J, and SST

CEFR	-	A1			A2		B1		B2		C1	C2
CEFR-J	Pre A1	A1.1	A1.2	A1.3	A2.1	A2.2	B1.1	B1.2	B2.1	B2.2	C1	C2
SST	1	2/3	3	4	4	5	6/7	8	9	9	9	9

of Europe, 2001)[4] which is utilized internationally, rather than the SST proficiency levels that are applicable only within Japan. The equivalent proficiency levels of SST, CEFR, and CEFR-J (a version of the CEFR that has been tailored to the needs of Japanese learning English) as defined by Tono et al. [5] are shown in Table 1. It should be noted that SST level 4 can be assigned to either CEFR level A1 and A2, and we will refer to these as CEFR1 and CEFR2 respectively. In this paper, the evaluation of the classification method was performed with SST level 4 included in the CEFR level A2. The classification of SST level 4 included in the CEFR level A1 should be investigated in future work. SST level 9 is included only in CEFR level B2.

For each of the 1280 examinee's in the SST data there are 5 stages of the interview that have been transcribed. In this paper, the results for each examinee were represented as one document, and there were 1280 sample documents for which the proficiency level classification problem was analyzed. Examinees who have an SST proficiency level of 1 were excluded as it would be equivalent to Pre A1 CEFR level. A total of 9,626 words were analyzed along with 11 parts of speech (POS) from Lancaster University's CLAWS5 and CLAWS7 tag sets[2].

Automated language scoring using a computer was first proposed by Page in 1968 [6]. Since then research into the prediction of foreign language proficiency has focused on a number of different approaches. Supnithi et al. [7], analyzed the vocabulary, grammatical accuracy and fluency features of the NICT-JLE corpus. SVM and Maximum Entropy classifiers were trained to automatically predict the proficiency level of the learner, with SVM achieving the best prediction accuracy of 65.57%. There has also been research into extracting features that can be useful in classifying proficiency levels in the NICT-JLE corpus [8, 9]. In this paper, analysis by SVM and feature selection is used to not only improve the accuracy of proficiency classification, but also identify optimal sets of characteristic features that can describe learners from different proficiency levels.

2 Proficiency Level Classification by SVM and Feature Selection

The occurrence frequency (tf) of each word was used to vectorize each of the transcripts. This was realized by creating a term document matrix of the exam transcripts using GETA[3].

[2] http://ucrel.lancs.ac.uk/claws5tags.html, http://ucrel.lancs.ac.uk/claws7tags.html
[3] http://geta.ex.nii.ac.jp

To evaluate the performance of classifying documents into two classes of proficiency levels, the documents of level X were represented as positive examples, while the documents of level Y were represented as negative examples to train a machine learning model. SVM^{perf}[10] was use to train and test models on the data of the corpus. The experiment process can be broken down into 3 main steps. All features (words, POS tags) are used to train a model in step 1. The ranking $weight(w_i)$ scores for each feature are then extracted from the model in step 2. These feature weights are then ranked in step 3 were the classification performance of models trained and evaluated using feature selections of increasingly larger sets of $N = 1, 2, ...10, 20, .., 100$ is analyzed. The optimal feature selection is the best performing model trained on N features. The classification performance of each model was evaluated using 5-fold cross validation.

The feature $weight(w_i)$ score extracted in Step 2 represents the distance from the SVM hyperplane that separates the positive and negative classes on which the model was trained. Models were trained with the upper proficiency level learner data as the positive class, and the lower level learner data as the negative class. Features that have positive $weight(w_i)$ are characteristic of upper level learners, and a negative feature $weight(w_i)$ are characteristic of lower level learners.

2.1 Feature Selection Measures

The classification performance of a model trained using all features for A1 and A2 were: Precision 0.8923, Recall 0.8117, F-measure 0.8491, and Accuracy 0.7830. Although the classification performance is quite high, we do not know which grammar items are effective for discriminating between different proficiency levels. In this paper, we apply the method from Sakai and Hirokawa [11] to the problem of feature selection to find a set of optimal discriminating features.

The feature score $weight(w_i)$ extracted in Step 2 was calculated using 6 different evaluation measures as shown in Table 2. $df(w)$ is the number of documents in which the word w occurs, and abs returns the absolute value of the enclosed value.

Table 2. Measures used for Feature Selection.

Symbol	Measure	Symbol	Measure
$w.o$	$weight(wi)$	$w.a$	$abs(weight(wi))$
$d.o$	$weight(wi) * df(wi)$	$d.a$	$abs(weight(wi) * df(wi))$
$l.o$	$weight(wi) * \log(df(wi))$	$l.a$	$abs(weight(wi) * \log(df(wi)))$

In the case of measures that do not take the absolute value of the score: the top N positive $weight$ features are selected along with the top N negative $weight$ features for vectorization. For measures that do take the absolute value of the score the top $2N$ positive $weight$ features are selected for vectorization.

3 Proficiency Classification Performance

This section explains the results of the proficiency classification performance by accuracy that are shown in Table 3, and plots of feature selection results for all measures shown in Figure 1. The x-axis in these plots represents $2N$ number of features selected. The results for A2 vs B1, shown on the left of Figure 1, and B1 vs B2 show that as the number of selected features increases, the accuracy increases following a curved line, suggesting that as the number of features increases the accuracy will steadily get higher. In other words, it is not possible to classify these classes with few features. Conversely, in the results of A1 vs A2, A1 vs B2, and A2 vs B2, shown on the right of Figure 1, the accuracy rises quickly at around $N = 10$. This indicates that a decent level of classification performance can be achieved using a small number of features.

Table 3. Classification performance when using all feature words.

Acc	A2		B1		B2	
A1	all	0.8188	all	0.9675	all	0.9966
	$N = 20$	0.7607	$N = 20$	0.8099	$N = 20$	0.7763
	$N = 50$	0.7837	$N = 50$	0.8615	$N = 50$	0.9333
	$N = 100$	0.8171	$N = 100$	0.9269	$N = 100$	0.9760
A2			all	0.8673	all	0.9879
			$N = 20$	0.6292	$N = 20$	0.8774
			$N = 50$	0.7750	$N = 50$	0.9657
			$N = 100$	0.8393	$N = 100$	0.9846
B1					all	0.8512
					$N = 20$	0.4493
					$N = 50$	0.6809
					$N = 100$	0.8405

The baseline classification performance of a model that was trained using all of the features is shown in Table 3. It can be seen that the classification performance of adjacent proficiency levels is low. The classification accuracy of feature selection is shown in Figure 1, where for the plot on the left A1 is the positive class and A2 is the negative class, and the x-axis is $2N$ top ranking features selected. When $N = 200$ of greater the accuracy of the model is slightly better than a model trained using all features. The two measures: $l.o$ and $d.o$ outperform the baseline at $N = 9$ which indicates that classification can be achieved with a small number of features.

4 Characteristic Features of Level A1

The top 10 characteristic features of level A1 are compared to other levels in Table 4. The feature "jp" represents a Japanese word that was said in the exam

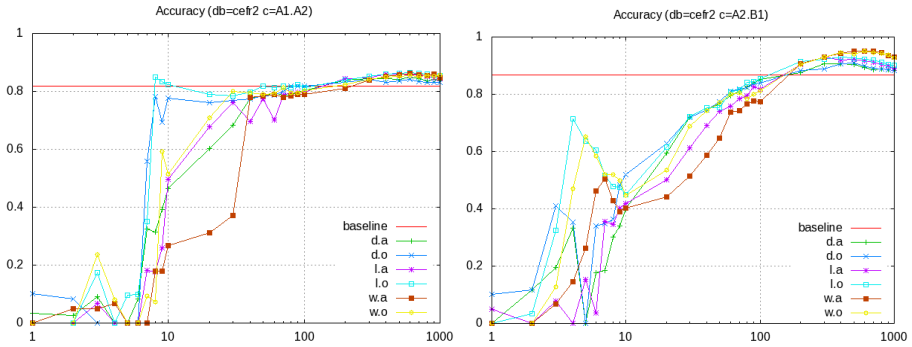

Fig. 1. Performance accuracy of feature selection

Table 4. Characteristic features of A1 and other comparative levels.

Characteristic Features of Level A1	Comparative level characteristics
look, please, jp, first, work, just, picture, what, friend, cat	(A2) home, find, when now, will, ask, other eat, think, if
ten, c7=RGQ, story, speak, theater, boy our, bring, anonym., favorite	(B1) really, also, ask call, actually, different c7=RRR, your, stay c7=DA
theater, pardon, cold color, zoo, lion, monkey shinjuku, recently, tv	(B2) an, into, drive brother, anything, club fun, once, teacher explain

and has been replaced during transcription. Regardless of which level A1 is compared to, the nouns: cat, theater, boy, zoo, lion, and monkey are frequent. This is most likely effected by the contents of picture cards on which conversations are based in certain SST tasks. On the other hand, other levels have higher numbers of verbs, adverbs, and adjectives. However, more high level parts of speech features such as VERB and ADJ are not seen as characteristic features. Therefore, discrimination between levels is not possible using simple parts of speech. Even though the POS tag information was analyzed, looking at the top ranking features when comparing A2 and B1, only 3 POS tags appear as characteristic features: C7=RGQ (adverb expressing a degree) for A1, and C7=RRR (comparative adverb) and C7=DA (adjective used as pronoun) for B1. Also in Table 4 is can be seen that different characteristic features are chosen when comparing A1 to different levels. An unexpected result is that classification can be achieved with just 20 features.

5 Conclusion and Future Work

In this paper, we analyzed the transcripts of a speaking test corpus by applying SVM machine learning to the problem of classifying the differences between CEFR proficiency levels. Feature selection was used to find an optimal feature set by evaluating the model accuracy. It was found that a set of about 20 fea-

tures produced the same performance as a model trained using all words and an accuracy of greater than 90%. For adjacent levels the classification accuracy was around 10% less. Classification of levels A1 vs B1 and B1 vs B2 were difficult and decent accuracy could not be achieved using small numbers of features. The characteristic features of level A1 contained numerous Japanese words, proper nouns, and simple nouns. In this paper, when assigning the equivalent levels of SST and CEFR we made the assumption that SST level 4 was contained within CEFR level A2. The analysis of SST Level 4 as CEFR A1 should be carried out in future work. Also the analysis of CEFR-J levels which are more detailed than CEFR level should be undertaken in future work.

6 Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 15H02778, 24242017, and 15J04830.

References

1. Izumi, E., Uchimoto, K., Isahara, H.: The NICT JLE corpus, ACL Publishing, 2004. (in Japanese)
2. Izumi, E., Uchimoto, K., Isahara, H.: The NICT JLE Corpus: Exploiting the language learner 's speech database for research and education. International Journal of the Computer, the Internet and Management 12(2), 119–125 (2004)
3. Izumi, E., Uchimoto, K., Isahara, H.: The Overview of the SST Speech Corpus of Japanese Learner English and Evaluation through the Experiment on Automatic Detection of Learners 'Errors 4th International Conference on Language Resources and Evaluation, 1435–1438 (2004)
4. Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press (2001)
5. Yukio Tono (Ed.): The CEFR-J handbook : a resource book for using CAN-DO descriptors for English language teaching, Taishukan Publishing (2013) (in Japanese)
6. Page, E. B.: The use of the computer in analyzing student essays, International Review of Education 14(2), 210–225 (1968)
7. Supnithi, T., Uchimoto, K., Saiga, T., Izumi, E., Virach, S., Isahara, H.: Automatic proficiency level checking based on SST corpus, Proc. RANLP, 29–33 (2003)
8. Abe, M.: Frequency Change Patterns across Proficiency Levels in Japanese EFL Learner Speech, Apples: Journal of Applied Language Studies 8(3), 85–96 (2014)
9. Flanagan, B., Hirokawa, S.: The Relationship of English Foreign Language Learner Proficiency and an Entropy Based Measure, IEE 1(3), 29–38 (2015)
10. Joachims, T.: Training Linear SVMs in Linear Time, Proc. ACM-KDD, 217–226 (2006)
11. Sakai, T., Hirokawa, S.: Feature words that classify problem sentence in scientific article. In Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, 360–367 (2012)