Memoirs of
Osaka Institute of Technology
Vol. 66, No. 2（2021） pp. 1～5

A Clustering Algorithm for Use in the Automatic Classification of L2 Spoken Output

by
Andrew MELLOR
Department of Media Science
Faculty of Information Science and Technology
(Manuscript received December1,2021)

Abstract

An automatic clustering algorithm was used to classify the spoken output of second language (L2) learners of English as being higher or lower quality. The results of this algorithm were compared with the ratings of two human judges. The clustering algorithm was shown to agree with one of the human raters more than the two human raters agreed with each other.

Keyword; second language acquisition, vocabulary, automatic assessment, testing

Introduction

This experiment trialed a clustering algorithm for use in the automatic assessment of English spoken output of second language (L2) learners. The algorithm was used to classify the spoken learner output as higher or lower quality. These results were compared with quality ratings determined by native judges. The algorithm is designed to be used with linguistic data resulting from speech recognition processing of spoken output. Such processing is outside the scope of this study.

Experimental design

Sixteen audio recordings of Japanese L2 learners of English were made. Learners spoke for two minutes on a simple topic. These recordings were independently assessed by two native judges and assigned, according to quality, to two groups, a higher and a lower quality group. The recordings were then transcribed to be assigned to a higher or lower quality group by an automatic clustering algorithm.

The clustering algorithm

The clustering algorithm used a small set of lexical features to classify the output by finding clusters of similar content in spaces that suggest high or low quality. This small set of lexical features was based on the features used in Mellor (2012) to cluster essays written by L2 learners. The following features were employed by the algorithm in this experiment:

1) length of spoken output

2) number of standard verb collocations

3) standardized type-token ratio (STTR)

4) sampled mean word length (MWL)

5) sampled estimate of lexical error

6) sampled number of hapax legomena

In Mellor (2012), essay length in words was used. This time, the length of spoken output was calculated as the number of words in the written transcript of the spoken output. In Mellor (2012), the type-token ratio (TTR) of a 100-word sample was used. In that study, a sample was used to preclude any length of transcript effects on the value of the TTR. This time, the standardized type-token ratio (STTR) was preferred because it is independent of text length but facilitates the use of all the words contained in each transcript. Mean word length was also used in the previous study and in this study it was sampled from each transcript to preclude any length of transcript effects. As in the previous study, an estimate of lexical error was included. This estimate was calculated as the number of words not appearing in the JACET word list (Ishikawa et al., 2003). Also, the number of hapax legomena, words appearing only once in any output, was calculated for a sampled size. An additional measure, the number of standard verb collocations was included. This was calculated by taking a number of common verbs from the output of each learner. The sampled verbs were those appearing in the most common 100 verbs in the JACET word list (Ishikawa et al., 2003). Credits for collocations identified as standard collocations according to the iWeb corpus (Davies, 2018) were then awarded. The magnitude of this credit was calculated in direct relationship with the frequency of the collocation in English according to the corpus. Not only were the number of common verbs per output standardized but also the number of credited collocations. Where features were sampled, as in feature 4), 5) and 6), the sample size was equal to the length of the shortest output.

Implementation of the Clustering Algorithm

Clustering was carried out using this set of six features and was initiated by using high and low values of each selected feature with the exception of the estimate of lexical error. With most features, high values should be indicative of higher quality output and low values should be indicative of lower quality output. However, in the

case of an estimate of lexical error, a low value is likely to be indicative of high quality output while a high value is likely to be indicative of low quality output. Standardized z scores for features were used for clustering. Initial cluster locations were estimated for a high and a low quality cluster. The high quality cluster was built around a point in 6-dimensional space based on the following parameters:

- Mean length of spoken output + 1 standard deviation
- Mean number of standard verb collocations + 1 standard deviation
- Mean STTR + 1 standard deviation
- Mean sampled mean word length + 1 standard deviation
- Mean sampled estimate of error - 1 standard deviation
- Mean sampled number of hapax legomena + 1 standard deviation

Similarly, the initial cluster point for the lower quality cluster was set as follows:

- Mean length of spoken output - 1 standard deviation
- Mean number of standard verb collocations - 1 standard deviation
- Mean STTR - 1 standard deviation
- Mean sampled mean word length - 1 standard deviation
- Mean sampled estimate of error + 1 standard deviation
- Mean sampled number of hapax legomena - 1 standard deviation

Output was progressively added to each cluster according to relative Euclidean distance to the midpoint of each existing cluster until two distinct clusters were formed.

Results

The results of the cluster analysis were compared with ratings of two native speaker judges. The decision agreements and Kappa statistics for agreement adjusted for chance are shown in Table 1.

Table 1: Decision agreement (DA) and Kappa (K) statistics for clustering algorithm and raters

	Rater 1		Rater 2	
	DA	K	DA	K
Clustering	.63	.26	.94	.88
Rater 1	-	-	.69	.38

The results of this clustering algorithm agreed with Rater 1 in 10 cases out of 16 and with Rater 2 in 15 cases out of 16. The Kappa statistic corrected for chance agreement is r = 0.26 for the clustering algorithm and Rater 1 and 0.88 for the clustering algorithm and Rater 2. The two raters agreed with each other in 11 cases out of 16 which corresponds to a Kappa reliability of 0.38. The clustering algorithm agreed with Rater 2 more than the human raters agreed with each other.

Comparing these results to those in Mellor (2012), the inter-rater decision agreement is slightly lower in this study, 0.63 compared to 0.72 in the previous study. But in both studies, a cluster algorithm often agreed with human raters more than raters agreed with each other.

Allocation of borderline cases is particularly difficult in these classification problems so it may be instructive to see the performance of the algorithm applied to the 11 cases that the human raters agreed on as shown in Table 2.

Table 2: Clustering results for output agreed by raters

		Raters	
		High	Low
Clustering	High	5	0
	Low	1	5

Here we can see that the clustering algorithm was very good at identifying the spoken output the raters agreed on. The algorithm correctly assigned all the low quality output the raters agreed on and only failed to identify one high quality output the raters agreed on. This agreement corresponds to a Kappa coefficient of K = 0.82. These results suggest that the algorithm is clearly good at identifying some high quality and low quality output but less effective at classifying more borderline cases. However, this also seems to be the case for the human judges themselves.

Individual features and quality

Mellor (2009) and Mellor (2011) showed that essay length by itself may be a strong predictor of quality of L2 learner essays. In order to investigate the relationship of length of spoken output and other features to the various ratings in this experiment, the output was classified according to each feature using the same clustering technique. The results of this classification were then compared with classifications by human raters and the clustering algorithm and the results are shown in Table 3.

Table 3: Decision agreement of features with raters and algorithm

	Rater 1	Rater 2	Clustering
Length	.69	.88	.94
Collocations	.69	.63	.63
STTR	.56	.75	.81
Error	.43	.63	.69
Hapax	.43	.63	.69
MWL	.56	.63	.69

Not surprisingly, all features show a relatively high correlation with the clustering algorithm probably because the features are incorporated into the algorithm. Output length shows the greatest correlations with the human raters but collocations also show a relatively strong correlation with Rater 1 and STTR shows a strong correlation with Rater 2. To investigate the correlation of length with human raters, the output was ranked according to length from the longest to the shortest and displayed in terms of whether they were classified as high or low by each rater. For Rater 1, we get the following sequence:

H H H L H L H H H H H L L L L L

This representation suggests that output length is very good at predicting the highest and the lowest quality output. The three longest learner outputs were rated as high quality by Rater 1 and the five shortest outputs were rated as low quality. However, some relatively long output was rated low by Rater 1. If we similarly rank the output according to length in terms of whether they were classified as high or low by Rater 2, we get the following sequence:

H H H H H H H L L L H L L L L L

Again, output length is very good at predicting the highest and the lowest quality output. The seven longest learner outputs were rated as high by Rater 2 and the five shortest were rated as low quality.

However, some relatively short output was rated high by Rater 2 In the same way, if we rank the output according to number of collocations and display them in terms of whether they were classified as high or low by the two raters, we get the following sequences:

H L H H H L L H H L L H L H L H (Rater 1)

H L H H L H L L H L H H L L L H (Rater 2)

This time, we see that although collocations have a relatively high decision agreement with the raters, they are not as good as output length at identifying high or low quality output. When viewed in terms of STTR, the following sequences result:

H H H H L L H L H L L H L H H L (Rater 1)

H H H H H L H L L H L H L L L L (Rater 2)

High values of STTR clearly identify high quality output according to the decisions of both raters and low values of STTR are also able to identify low quality output according to the ratings of Rater 2. These results support previous research showing the strong relationship between length and quality of student output.

Conclusions

For assessing quality in L2 spoken output in this study, the clustering algorithm agreed with human ratings more than the human raters agreed with each other. This suggests that even a simple model may have a role to play in automatic assessment. Length of the output showed itself to be a strong predictor of output quality and may play a major role in the performance of the clustering algorithm. Among the other features, the number of common verb collocations and the STTR showed a strong relationship to output quality in certain situations and may also have a role to play in automatic assessment of spoken output.

References

Davies, M. (2018) *The iWeb Corpus*. Available online at https://www.english-corpora.org/iWeb/.

Ishikawa, S., Uemura, T., Kaneda, M., Shimizu, S., Sugimori, N., & Tono, Y. (2003). *JACET8000: JACET list of 8000 basic words.* Tokyo: JACET.

Mellor, A. (2009). Practical Automatic Assessment of L2 Learners, *Memoirs of the Osaka Institute of Technology*,Series B, Vol. 54, No. 2, 15-26.

Mellor, A. (2011). Essay Length, Lexical Diversity and Automatic Essay Scoring, *Memoirs of the Osaka Institute of Technology*, Series B, Vol. 55, No. 2, 1-14.

Mellor, A. (2012). Automatic Essay Classification, *Memoirs of the Osaka Institute of Technology*, Series B,Vol. 57, No. 2, 37-44.