Memoirs of Osaka Institute
of Technology, Series B
Vol. 57, No. 2 (2012) pp. 37～44

Automatic Essay Classification

by

Andrew MELLOR
Department of Media Science
Faculty of Information Science and Technology

(Manuscript received Sep 29, 2012)

Abstract

Two automatic algorithms, a clustering algorithm and a Bayesian classifier, are used to classify essays written by second language (L2) learners of English as higher quality or lower quality. The results of each algorithm are compared with each other and with ratings of human judges. The clustering algorithm is shown to agree with human ratings more than human raters agree with each other.

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

Introduction

This experiment investigates 2 possible algorithms for automatically assessing quality of English essays written by second language learners. The first algorithm is based on cluster analysis of lexical features. The second is based on Bayesian analysis of words used in the essays. In the latter algorithm, a number of basic lexical features are also used to identify a set of training essays. Both algorithms are used to classify the same essays as higher quality or lower quality and results are compared with ratings of native judges as well as between algorithms for insights into the reliability of automatic assessment methods.

Experimental design

One hundred essays written by first year Japanese university students were collected. Students were given 25 minutes to write an essay based on a cartoon strip (Oser, 1934). These essays were assigned to 2 groups of 50 essays according to quality by 2 native judges. One was a higher quality group and the other was a lower quality group. The essays were then also assigned to a higher or lower quality group by each of 2 algorithms, a clustering algorithm via the L_cluster computer program and a Bayesian algorithm via the L_Bayes computer program. Both L_cluster and L_Bayes are computer programs specially designed by the author.

A clustering algorithm

The clustering algorithm used a small set of lexical features to cluster essays according to similarity. This small set of features was refined from a larger set through a principal component analysis (PCA). Two initial clustering points were chosen for the clustering algorithm: one to indicate a likely high quality space and the other to indicate a likely low quality space.

Selection of features

The first stage of the analysis was to identify a set of input lexical features for the cluster analysis. The following simple features that have been associated with essay quality in previous research were considered:

1) Essay length in words (Larsen-Freeman & Strom, 1977; McNeill, 2006)
2) TTR(100), the number of word types in a 100 word sample (Malvern et al., 2004)
3) Hapax(100), the number of hapax legomena in a 100 word sample (Mellor, 2011)
4) An estimate of Advanced Guiraud (Guiraud, 1960; Daller et al., 2003; Mellor, 2011)
5) A distance measure of occurrences of *the* and *a* to typical occurrence in English (Evola et al., 1980; Mellor, 2008)
6) An estimate of mean sentence length (Mellor, 2008)
7) An estimate of mean clause length (Mellor, 2008)
8) Mean word length (Zipf, 1932; Mellor, 2008)
9) Entropy (Mellor, 2009)
10) Yule's K (Yule, 1944; Mellor, 2011)
11) An estimate of lexical error (Engber, 1995)

Some of these features are not easily calculated automatically and so estimates were calculated. Mean sentence length was estimated using the number of words in the essay divided by the number of sentence ending punctuation marks while mean clause length was calculated by dividing the number of words in the essay by the number of commas and sentence ending punctuation marks. Lexical error in this analysis was predicted by a small subset of error. This error estimation process involved checking all words in the essay against a list based on the first 1000 words of the JACET word list (Ishikawa et al., 2003) and against a list of proper nouns. Any words found outside these lists were flagged for human checking. Any judged to be non-words were tallied for an estimate of lexical error. An estimate of Advanced Guiraud involved comparing words in The JACET 1000 word list. Any words not in this list and also not appearing in a list of errors and proper nouns were considered advanced types.

Principal component analysis (PCA)

A PCA was carried out to identify a smaller set of features to use in the cluster analysis. PCA is a statistical technique which realigns multivariate data to provide a new set of variables which are ordered in terms of variance and are independent of each other. Each feature was calculated for each essay and the standardized z scores were subject to a PCA. Z scores were used to prevent features with large variances dominating the analysis. Table 1 shows the variance accounted for by each principal component (PC) and the cumulative variance for this experiment. The first 6 PCs accounted for 91.7% of the variance in the data.

Table 1: Percentage of variance by PCs

PC	Variance %	Cumulative variance %
1	34.2	34.2
2	20.1	54.3
3	12.9	67.2
4	12.3	79.5
5	6.2	85.6
6	6.1	91.7
7	3.9	95.6
8	2.2	97.8
9	1.4	99.2
10	0.6	99.9
11	0.1	100

The original z scores of features were then mapped against each PC to find original features that correspond closely to each PC to give a set of independent features to use in the cluster analysis (Jolliffe, 2002). In this analysis, the first PC which accounted for over 34% of the variance was highly correlated with the feature essay length. The 6 selected features for each of the first 6 PCs were as follows:

PC1: Essay length
PC2: TTR(100)
PC3: An estimate of mean sentence length
PC4: An estimate of lexical error
PC5: Mean word length
PC6: Hapax(100)

The use of PCA to ascertain the features that correspond with most of the variance in the essay data has at least 2 advantages. Firstly, a large number of features can be considered initially and the best features selected. Secondly, PCA makes the clustering algorithm more versatile as the optimum set of features can be selected by PCA for the algorithm according to the essay data in each case.

Clustering

Clustering was carried out using this set of 6 features and was initiated by using high values and low values of the selected features. High values should be indicative of high quality essays and low values should be indicative of lower quality essays. However, feature 4, lexical error, was opposite in its orientation. A low value of lexical error is likely to be indicative of a high quality essay while a high value is likely to be indicative of a low quality essay. Standardized z scores for features were used for clustering. Cluster locations were estimated for a high quality cluster and a low quality cluster. The high quality cluster was built around a point in 6-dimensional space based on the following parameters:

- Mean essay length + 1 standard deviation
- Mean TTR(100) + 1 standard deviation
- Mean sentence length +1 standard deviation
- Mean lexical error -1 standard deviation
- Mean word length + 1 standard deviation
- Mean Hapax(100) + 1 standard deviation

In a similar way, the initial cluster point for the lower cluster was set as:

- Mean essay length - 1 standard deviation
- Mean TTR(100) - 1 standard deviation
- Mean sentence length -1 standard deviation
- Mean lexical error +1 standard deviation
- Mean word length - 1 standard deviation
- Mean Hapax(100) - 1 standard deviation

Each feature was weighted in accordance with the percentage of variance accounted for by each corresponding PC. Essays were progressively added to each cluster according to relative Euclidean distance to the midpoint of each existing cluster until 2 clusters of 50 essays each were formed.

Results

The results of the analysis were compared with ratings of 2 native speaker judges and the decision agreements and Kappa statistics for agreement adjusted for chance agreement are shown in Table 2.

Table 2: Decision agreement (DA) and Kappa for clustering algorithm

	Rater 1		Rater 2	
	DA	Kappa	DA	Kappa
Clustering	.78	.56	.76	.52
Rater 1	-	-	.72	.44

The results of this clustering algorithm agreed with Rater 1 in 78 cases out of a 100 and with Rater 2 in 76 cases out of a 100. The Kappa statistic corrected for chance agreement is r = 0.56 for the clustering algorithm and Rater 1 and 0.52 for the clustering algorithm and Rater 2. The 2 raters agreed with each other in 72 cases out of a 100 which corresponds to a Kappa reliability of 0.44. Therefore the clustering algorithm agreed with each human rater more than the human raters agreed with each other.

One of the reasons for the relatively low kappa values may be the requirement that 50 essays be allocated to each group. Although this is a realistic assessment situation, it may cause problems in classification. It is unlikely that the 100 essays naturally fall into 2 sets of 50 essays according to proficiency. Very high quality and very low quality essays may be relatively easy to allocate but essays that are borderline are likely to prove more difficult. A good deal of this lost reliability may be due to raters and algorithms dealing with borderline essays in different ways. The performance of the algorithm on the essays the human raters agreed on is shown in Table 3.

Table 3: Clustering results for essays agreed by raters

		Raters	
		High	Low
Clustering	High	32	5
	Low	4	31

The 2 human raters agreed on 36 high quality essays and 36 low quality essays. Of these 36 high quality essays, 32 were also classified as high quality by the clustering algorithm but 4 essays that both raters identified as high quality were rated low quality by the algorithm. Out of 36 essays rated low by both human raters, 31 were also rated low by the algorithm but 5 were rated high. It could be that some essays are being mis-rated by the clustering algorithm. It could also be that these essays are, in fact, borderline essays that both raters just happened to rate the same way. For this smaller group of pooled ratings, there is decision agreement of 63 cases out of 72 (87.5%) or a Kappa statistic of 0.75.

A Bayesian algorithm

The second algorithm used a Bayesian classifier to categorize the essays. Fifty essays were identified as being high quality leaving the remaining 50 essays to be classified as low quality. The Bayesian classifier was trained by a sample of essays selected automatically from the whole set of essays. This selection was done by recognizing features that are highly likely to indicate high quality essays.

The basic premise of a Bayesian classifier is that classification of an item (in this case, an essay) can be guided by comparing the occurrence of features of the item (in this case, words in the essay) with the occurrence of features in groups of items (in this case, samples of high quality essays and other essays). The item can be classified as a member of the group to which there is most similarity in occurrence of features. Various essay features could be used in the analysis. Lexical statistical features such as those used in the clustering algorithm could be used but in this algorithm the lexical content of the essays was analyzed. Occurrence of particular words in each essay was compared to occurrence in a set of predicted high quality essays.

To carry out a Bayesian analysis, a number of essays are needed as training samples. In this experiment, these training essays were selected automatically from within the set of essays. The method of selecting these training essays was based on observations in other studies. A previous study (Mellor, 2008) suggested that some simple lexical features such as essay length or lexical diversity could be effective at identifying small numbers of very good essays or very poor essays.

Selecting the training set

A training sample of essays was selected from within the set of essays by using combinations of features which were highly likely to indicate good quality essays.

The 4 features chosen were essay length, TTR(100), Hapax(100) and an estimate of error. The error estimate was calculated in the same way as in the clustering algorithm and an error proportion in relation to essay length calculated.

A set of conditions for predictors of high quality essays was constructed. These conditions were considered in descending order of strictness until a sufficient number of essays (15-20) were identified. The conditions were as follows:

1) In the top quartile of each of 3 features, essay length, TTR(100) and Hapax(100) while also not including a high proportion of error (A high proportion of error was an error proportion in the top quartile)
2) In the top quartile of essay length and TTR(100) while being above average for Hapax(100) and not including a high proportion of error
3) In the top quartile of essay length and Hapax(100) while being above average for TTR(100) and not including a high proportion of error
4) In the top quartile of essay length while being above average for both TTR(100) and Hapax(100) and not including a high proportion of error
5) In the top quartile for both TTR(100) and Hapax(100) while also being above average for essay length and not including a high proportion of error
6) In the top quartile of TTR(100) while being above average for both essay length and Hapax(100) and not including a high proportion of error
7) In the top quartile of Hapax(100) while being above average for both TTR(100) and essay length and not including a high proportion of error
8) Above average for all 3 features of essay length, TTR(100) and Hapax(100) while exhibiting zero error

These 8 conditions identified 16 possible high quality essays in the proportions shown in Table 4.

Table 4: Essays identified by the training set

Condition	1	2	3	4	5	6	7	8
No. of essays	9	2	2	1	0	0	1	1
Total	9	11	13	14	14	14	15	16

Bayesian classification

These essays formed the training sample and were also initial members of the high quality group. The next stage of the analysis was to use a Bayesian classifier to assign a further 34 essays to the high quality group leaving the remaining 50 essays to form the low quality group.

Occurrences of all words in an essay that appeared in more than one essay in the set of essays were considered. For each word in each essay, the occurrences of that word in the essays in the high sample group and in the essays of the group that includes all other essays were tallied. For example, the first word of an essay under consideration is *once*. The occurrence of *once* is then checked in the 16 training essays. If *once* occurs in 8 out of these 16 essays but *once* only occurs 12 times in the remaining 83 essays, *once* has a higher relative occurrence in the high quality sample than in the group of other essays. Therefore, on the basis of the occurrence of *once*, the essay seems more likely to belong to the high quality sample than to the remaining group. Of course, on its own, this one word is not a reliable predictor, but when every other word in the essay is taken into account in a similar way, the Bayesian classifier will produce a more realistic probability of whether the essay belongs to the high quality group or the remaining group.

This procedure was done for all the remaining 84 essays and the 34 essays with the highest probability of belonging to the high quality group were allocated to that group. The remaining 50 essays were allocated to the low quality group.

Results

A comparison of decision agreements and Kappa statistics from the Bayesian algorithm with human ratings is shown in Table 5. This algorithm agreed with both Rater 1 and Rater 2 in 70 cases out of a 100. This gives a Kappa statistic of r = 0.40 for the reliability of the algorithm with each rater compared with r = 0.44 for raters with each other. This suggests that the performance of this algorithm is not as good as the performance of 2 human raters nor as good as the clustering algorithm.

Table 5: Decision agreement and Kappa reliability for Bayesian algorithm

	Rater 1		Rater 2	
	DA	Kappa	DA	Kappa
Bayesian	.70	.40	.70	.40
Rater 1	-	-	.72	.44

Table 6. shows the essays the human raters agreed on and how they were treated by the Bayesian algorithm.

Table 6: Automatic treatment of essays agreed on by raters

		Raters	
		High	Low
Bayesian	High	30	10
	Low	6	26

The number of agreements between the Bayesian algorithm and the pooled ratings was 56 cases out of 72 or 78% and the Kappa statistic was 0.56. Out of 36 essays rated as high quality by both human judges, 30 were also rated high quality by the Bayesian algorithm. This means that 6 essays rated high by both raters were rated low by the Bayesian algorithm. Similarly, out of 36 essays rated low by both human raters, 26 were also rated low by the Bayesian algorithm. This means that 10 essays rated low by both raters were rated high by the Bayesian algorithm.

The Bayesian algorithm allocated essays to the high quality group by 2 processes. The first process was by the training set conditions which identified 16 essays. The second process was by the Bayesian classifier which allocated a further 34 essays to the high quality group and the remaining 50 to the low quality group. Therefore, the agreement in high quality candidates cannot be credited entirely to the Bayesian classifier.

To check the reliability of the training set essays, the ratings of the 16 selected essays were compared with ratings for these essays by the 2 human raters. The number of training sample essays rated high by each rater by selection condition is shown in Table 7.

Table 7: Training set essays rated high by raters

	condition					
	1	2	3	4	7	8
No. of essays	9	2	2	1	1	1
Rater 1	9	2	2	1	0	1
Rater 2	9	2	1	1	1	1

The table shows that both raters rated all but one training set essay as being high quality. No essay was rated poor by both raters. The essay rated poor by Rater 1 was an essay selected by condition 7 and the essay rated poor by Rater 2 was an essay selected by condition 3. All the essays selected by the first 2 conditions were also rated high by both raters. Of the 16 essays identified by the training set, there were 14 whose rating was agreed on by both raters. These 14 were all rated as high quality by the training conditions.

Fourteen of the 16 training set essays were agreed on by both raters. Therefore, the essays agreed on by human raters allocated by the Bayesian classifier are shown in Table 8.

Table 8: Bayesian classifier treatment of essays agreed by raters

		Raters	
		High	Low
Bayesian classifier	High	16	10
	Low	6	26

There were 58 essays that the raters agreed on that were allocated by the Bayesian classifier. The agreement between the results of the classifier and the raters was 42 out of 58 cases which is agreement of 72% or a Kappa statistic of 0.47.

This suggests that the training set identification portion of the Bayesian algorithm may be more reliable than the Bayesian classifier portion of the algorithm.

Comparison of clustering and Bayesian algorithms

Results show the clustering algorithm was more reliable in classifying the essays than the 2 native judges but the Bayesian classifier was less reliable than both the clustering algorithm and the native judges.

Inter-algorithm reliability

Although one of the advantages of automatic assessment is that it eliminates some of the reliability concerns which plague human rating, it does come with some concerns of its own. Two important reliability concerns for humans are inter-rater reliability and intra-rater reliability. Inter-rater reliability refers to the consistency of score awarded to the same essay by different raters. It is problematic if 2 raters score the same essay differently. Intra-rater reliability refers to the inconsistency of rating by a particular rater. On different occasions the rater might award different grades to the same essay. While automatic assessment eliminates these 2 concerns, it does raise another reliability issue of its own: inter-algorithm reliability. There are potentially many different algorithms for automatic assessment. This experiment includes just 2 out of a great number of possibilities. Inter-algorithm reliability is concerned with the consistency of rating between algorithms. It checks to see if different algorithms rate the same essay in a similar way. The agreement between the 2 algorithms in this experiment is shown in Table 9.

Table 9: Agreement of 2 automatic algorithms

		Clustering	
		High	Low
Bayesian	High	39	11
	Low	11	39

Comparing the performance of the clustering algorithm and the Bayesian algorithm, there was agreement in 78 cases out of a 100. There was agreement in 39 high cases and 39 low cases. This translates to a Kappa inter-reliability measure of 0.56. This is higher than human inter-rater reliability but there is still considerable scope for improvement given that one of the arguments for automatic assessment is to improve reliability.

Essay length

Mellor (2009) summarized the evidence that essay length is a strong predictor of quality of second language learner essays. In order to investigate the relationship of essay length to the various ratings in this experiment, the essays were classified according to length only with the 50 longest essays classified as good quality and the 50 shortest essays classified as poor quality. The results of this classification were then compared with the classifications by human raters and algorithms as shown in Table 10.

Table 10: Agreement of essay length with raters & algorithms

	Rater 1	Rater 2	Clustering	Bayesian
Essay length	74	74	88	88
Clustering	78	76	-	78
Bayesian	70	70	-	-
Rater 2	72	-	-	-

The results show that essay length is a reliable predictor of rating for this set of essays. In fact, essay length correlates better with both raters than the raters do with each other. There were 74 agreements out of a 100 between an essay length model and either rater but only 72 agreements between the 2 raters. However, the clustering algorithm correlated more closely with both raters than essay length alone. The clustering algorithm matched Rater 1 in 78 cases compared with 74 for essay length alone and matched Rater 2 in 76 cases compared with 74 for essay length alone. This suggests that this algorithm may be an improvement over using only essay length as a predictor. However, essay length alone performs better than the Bayesian algorithm in terms of agreements with human raters with 74 matches compared with only 70 for the Bayesian algorithm.

The results also show that essay length is very strongly correlated with the results of the 2 automated algorithms with 88 out of a 100 matches. It is worth noting that both algorithms, although scoring 88 matches with essay length, showed different patterns of matches. This evidence suggests that these algorithms may be strongly influenced by essay length.

Conclusions

For assessing quality in the learner essays in this study, the clustering algorithm was more effective than the

Bayesian algorithm and it agreed with human ratings more than human raters agreed with each other. The Bayesian algorithm appeared less effective in comparison with human raters. Essay length again showed itself to be a strong predictor of essay quality in learners. The 2 approaches chosen in this experiment, cluster analysis and Bayesian analysis, are but 2 of almost limitless choices in the area of multivariate analysis. The experiment shows the promise of exploiting cluster analysis but there are many other approaches that could be considered. In addition, there are different types of information that can be used in the analysis. In this experiment, cluster analysis utilized statistical features of the essay while the Bayesian analysis was weighted heavily toward lexical content.

References

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in spontaneous speech of bilinguals. *Applied Linguistics 24* (2), 197-222.

Engber, C.A. (1995). The relationship of lexical proficiency to the quality of ESL compositions *Journal of Second Language Writing 4 (2)* 139-155.

Evola, J., Mamer, E., & Lentz, B. (1980). Discrete point versus global scoring for cohesive devices. In J.W. Oller & K. Perkins (eds), *Research in language testing* (pp. 177-181). Rowley: Newbury House.

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Dordrecht: D. Reidel.

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

Jolliffe, I.T. (2002). *Principal component analysis*. New York: Springer Verlag.

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning, 27* (1), 123-134.

Malvern, D.D., Richards, B.J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. New York: Palgrave MacMillan.

McNeill, B.R. (2006). *A comparative statistical assessment of different types of writing by Japanese EFL college students.* Unpublished PhD thesis, University of Birmingham, UK.

Mellor, A. (2008). A comparison of word distributions in L1 and L2 texts, 四日市大学総合政策学部論集、第 7 巻　第 1,2 合併号，91-98.

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.

Oser, E. (1934). *Vater und Sohn*. Konstanz: Sudverlag.

Yule, G.U. (1944). *The statistical study of literary vocabulary*. Cambridge: CUP.

Zipf, G.K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.