Practical Automatic Assessment of L2 Learners
by
Andrew MELLOR

Department of Media Science
Faculty of Information Science and Technology
(Manuscript received September 30, 2009)
Practical Automatic Assessment of L2 Learners
by
Andrew MELLOR

Department of Media Science
Faculty of Information Science and Technology
(Manuscript received September 30, 2009)

Abstract
Automatic assessment of written essays has become a viable alternative to human rating in many testing situations in L1 and to a lesser extent in L2. This paper examines the challenge of making automatic assessment of L2 written essays a more accessible tool for teachers and test administrators in a wider variety of contexts. A background to the present state of automatic assessment of essays is presented along with a review of lexical statistics that may be useful as well as recommendations for the challenge of producing a more versatile assessment tool.

Keyword: second language acquisition, vocabulary, automatic assessment, language testing, CALL, lexical statistics
Introduction

Automatic assessment has a number of advantages over assessment by human judges. Firstly, automatic assessment is time and cost effective. It takes considerable time for a teacher to rate a stack of essays whereas automatic assessment can be done in a matter of seconds with the press of a button. A lack of time to mark written work may be one of the factors deterring teachers from setting more written assignments and assessments. Also, the cost of human judges if they are needed is usually prohibitively expensive. Secondly, the argument for automatic assessment is strengthened by the problems associated with human judges. These problems include inaccuracy, inconsistency and bias in individual judges as well as judges using different criteria and standards for rating the same essays. Reliability of human rating of essays is notoriously low. In order to increase inter-rater reliability, a number of judges can be employed. But the 3 or more judges needed to raise the inter-rater reliability are in practice not economically viable in most situations. In small scale classroom situations, using other judges is not an option. As a result, teachers often have to rate their own students which can raise questions of bias and objectivity. While there may be some people who feel uncomfortable about computers taking over the whole rating process, one possible role for automatic assessment is alongside human judges to improve reliability of rating. This could also provide a check in classroom settings against teacher bias when rating one’s own students. Also, in other situations, automatic assessment may preclude the need for teacher/rater input completely. Learners could submit their written work and get immediate rating and feedback from an automatic assessment computer application.

Automatic assessment has been used successfully in a variety of testing situations. However, there are still challenges involved in making it available in a more versatile and accessible desktop form that could be used for various assessment tasks or as a backup to human assessment. Such a tool may enable teachers and test administrators alike to use more writing in testing situations such as end-of-course evaluations or placement tests.

Automatic Assessment in large-scale testing

Automatic assessment has been used in a number of large-scale testing situations for some time. At present, most automatic assessment applications cater for first language (L1) writing while others also target second language (L2) writing.

Project Essay Grade

Project Essay Grade (PEG) uses computers to evaluate L1 writing of American high school students. In a PEG experiment (Page, 1994), 599 essays written by 12th grade students were rated by 6 human judges on a six point scale and the results were compared with ratings produced by an automatic assessment model based on multiple regression analysis. The automatic assessment ratings were found to be more reliable than ratings from individual human judges and as reliable as the pooled ratings of 2 or 3 human judges. The major criteria for selecting input variables for the multiple regression was to identify intrinsic qualities of writing, which Page calls trins, such as fluency, diction, grammar, punctuation and then find markers of these intrinsic qualities, called proxes, which can be measured. For the trins of diction, Page proposes a prox of word length since longer words are...
Practical Automatic Assessment

generally less common than shorter words. For the *trins* of complexity of sentence structure, *proxes* such as number of prepositions or the number of relative clauses could be considered. Unfortunately, Page doesn’t state exactly what input variables are actually used in the PEG model. As for the analysis, multiple regression demands a high number of datasets. In fact, Hatch & Lazaraton (1991) suggest 30 datasets for each independent variable. So if 5 variables were used, at least 150 essays would be necessary. The high number of essays, 600, in the PEG experiment is sufficient for a large number of independent variables. However, multiple regression might not be best suited to small-scale analysis which is often the case in many practical L2 testing situations where teachers often have far fewer essays that require rating.

**E-rater**

Another automated system is e-rater devised by Educational Testing Service (ETS). E-rater has been used to assess essays of both L1 and L2 writers. E-rater uses a stepwise linear regression model to find the best features to score a particular essay question. E-rater initially measures over 50 features from 4 types: syntactic, discourse, topical and lexical. It is trained on a set of essays graded by human judges and finds the 8 to 10 features to best score each question prompt. The score produced is rounded to the nearest whole number for the predicted score. In an experiment reported by Chodorow & Burstein (2004), e-rater01 was used to assess a large number of TOEFL written essays for 7 different essay prompts. For each prompt, e-rater was trained on a set of 265 essays marked by 2 human judges on a 6-point scale. A set of 500 essays for each prompt were then rated by e-rater and the ratings compared to those of the judges. E-rater01 scores correlated with the scores of 2 human judges with \( r = 0.6 \) which was a higher correlation than between the two judges themselves, \( r = 0.59 \). The following features were the ones most used in e-rater01 in four or more models out of seven

1) number of words of various lengths 5,6,7,8 characters (all models) (lexical)
2) number of different word types (all models) (lexical)
3) number of auxiliary verbs in essay (5 models) (syntactic)
4) score from topical analysis by argument (4 models) (topical)
5) number of argument development words in first paragraph (4 models) (discourse)

In contrast to the PEG methodology where variables are selected as proxy measures of important traits, in the ETS methodology, the selection of variables is driven purely by the statistical process. This favors the most statistically efficient variables and may have repercussions in terms of validity. Sheehan (2001) argues that e-rater depends on features influenced by essay length rather than writing quality. Chodorow & Burstein (2004) also found that there is a strong correlation between length of essay and ratings by humans and automatic rating systems. In this case, the e-rater model may indirectly reward length because three of the five most common features are totals that increase with essay length. Both the reliability and validity of automated systems like e-rater depend on the scores correlating highly with those of human judges. However, if the system is perceived as
overly rewarding essay length over quality, then the validity of the system itself is put into doubt.

**Latent Semantic Analysis**

A third method of automatic assessment is the Intelligent Essay Assessor (IEA) which is based on the process of Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) and aims to measure conceptual content of essays rather than grammar, style or mechanics. LSA is the mathematical representation of meaning relations among words and passages using statistical computations involving singular value decomposition. In LSA, each essay is transformed into a vector in semantic hyperspace and compared to other essay vectors. Two measurements are derived to help rate essays. Firstly, the cosine of each essay vector to other vectors is used to determine the quality of that essay. This is complemented by a second measurement, vector length, which measures domain relevance in relation to the semantic space created. Landauer et al. (2003) report on a number of experiments using IEA with standardized essays and classroom essays. In these experiments, the IEA scores agreed with human judges as well as human judges agreed with each other. Inter-rater reliability was checked between the human judges, IEA and individual judges and IEA and pooled judges. For over 1600 standardized essays, the values were 0.86, 0.85 and 0.88 respectively and for 845 classroom essays, 0.75, 0.73 and 0.78 respectively. IEA also requires a set of training essays. The effect of the number of these pre-scored essays on reliability was also investigated. Reliability values varied from 0.53 with 6 pre-scored essays to 0.69 with 25 essays to 0.75 with 400 essays.

LSA research has mostly focused on native language testing situations. Landauer & Dumais (1997) argue that LSA tells us about the writer’s conceptual knowledge of a particular topic by the degree of similarity to expert writers, subject material or other essay writers. It is likely that the words that are important in such an analysis may be those with particular relevance to that topic or field. In second language learning contexts, essay grading is usually not so much about the conceptual content of essays but about language ability. Question topics are likely to be more general and the vocabulary less topic specific. Whether LSA has a role to play in L2 testing situations is still unclear as there seems to have been little research in that area yet although Crossley et al. (2008) used LSA to track lexical development in L2 learners. In a longitudinal study of six L2 learners of English, the LSA scores of spoken utterances increased over the one year span of the experiment in line with other measures of lexical development.

In comparison with other methods of rating essays such as PEG and e-rater, LSA has a definite advantage in terms of validity if one buys into the linguistic model of LSA. But also the form of the analysis itself seems to provide more validity than the other 2 models. The methodology used by PEG or e-rater depends mostly on various types of word counts. In LSA the semantic relationships of words themselves are analyzed. So although the word order information is still lost in LSA, word meaning information is preserved. This seems to include an extra dimension of information compared with the other largely count-based analyses. However, one drawback to LSA is that despite a more complex model which incorporates more information, it doesn’t seem
to perform any better than the other models. It would be interesting to see if methods like LSA can be utilized alongside other analyses to give an enhanced methodology. In this respect, the other methods seem to have an edge in flexibility. Because they are not constrained by any theory, they are free to incorporate anything that works into their models. In fact, the latest version of e-rater seems to have incorporated an LSA-like feature (word-vector score and word-vector cosine correlation) as one part of its analysis (Lee et al. 2008). It will be interesting to see if this yields any improvement in performance. Another disadvantage that LSA shares with the other patented commercial applications is that it is difficult to find out exactly how it works. It is marketed for large-scale testing situations and opportunities to experiment with it are limited to online sites such as University of Memphis LSA site and the Coh-Metrix website (McNamara et al., 2005) which provide some simple LSA-like output. These drawbacks notwithstanding, LSA type procedures offer some hope for the creation of a more user-friendly and versatile testing instrument. One thing holding back automatic assessment being used in a greater variety of testing situations is the necessity of having pre-rated or model essays. Landauer et al. (2003) also report on a sub-experiment where essays were evaluated by proximity to other essays in the cohort rather than to expert material. This sub-experiment reported high levels of reliability albeit a little lower than experiments involving training sets of essays. This approach of evaluating essays solely by proximity to other test essays is worth exploring in more detail.

Although these 3 automatic assessment methods all perform impressively, they are all geared for large-scale testing situations. Also, they all require a large amount of training data as well as considerable computing power. For a more versatile application, perhaps simpler and more flexible methods are required.

**Lexical statistics and essay ratings**

The analyses of both PEG and e-rater involve a selection of lexical statistics. There has been much research on the role of lexical statistics and their correlation with evaluations of essays. In particular, a look at two detailed studies of lexical statistics and their relationships to ratings of L2 essays is worthwhile. Engber (1995) investigated the relationship between lexical proficiency and reader quality ratings of timed essays. Sixty six timed essays by L2 learners of English were rated by human judges and the scores compared with the following 4 measures of lexical richness based on the essays:

a) lexical variation
b) error free lexical variation
c) percentage of lexical error
d) lexical density

These four measures of lexical richness were calculated as follows:

a) lexical variation including error (LV1)

$$LV1 = \frac{\text{sum of different lexical items per segment}}{\text{sum of lexical items per segment}}$$

i) Lexical items were defined as full verbs, nouns, adjectives and adverbs with an adjectival base.
ii) The sum of different lexical items treats inflected forms of the same word as the same item.

iii) Lexical variation tends to decrease as sample size increases so each essay was divided into 126 word segments, based on 1/3 length of the longest essay, 378 words. LV1 was then calculated for each segment.

b) lexical variation excluding errors (LV2)

\[
LV2 = \frac{A - B}{C}
\]

where

- \(A\) = the sum of different lexical items per segment
- \(B\) = the sum of lexical errors per segment
- \(C\) = the sum of lexical errors per segment

i) grammatical and syntactic errors were ignored.

c) percentage of lexical errors (%LE)

\[
%LE = \frac{\text{total lexical errors}}{\text{total lexical items}} \times 100\%
\]

d) lexical density (LD)

\[
LD = \frac{\text{total lexical items}}{\text{total words in essay}}
\]

It was found that readers of essays were negatively affected by lexical errors. Error and lexical variation together accounted better for holistic score than either error or lexical variation individually which suggests that readers give higher scores to essays that use a variety of lexis correctly.

An important consideration with these measures is how practical they are. One element of this is whether the calculation can be automated and done by computer. Unfortunately, judging errors is difficult even for experienced teachers and examiners. Some of the problems with errors are:

- judging whether something is an error or not.
- judging whether an error is a lexical error or some other kind of error.
- judging to what degree complex compound errors are lexical errors.
- the counting of errors - a simple count of errors may treat all errors as equally serious.

Error analysis of an essay needs careful consideration by an experienced professional and is likely to take as least as much time as a holistic evaluation of the essay. Therefore, it seems difficult to integrate lexical error measurement into automatic assessment. However, one possibility is to devise an alternative predictor of error that is simpler to calculate: in the language of Page, a prox for the trins of lexical error.

In another experiment, Ferris (1994) analyzed L2 placement test essays for the occurrence of 28 quantitative lexical and syntactic features.

1. Number of words
2. Words per sentence
3. Word length
4. Present tense verbs
5. Past tense/perfect aspect
6. 1st/2nd person pronouns
7. 3rd person pronouns
8. Impersonal pronouns
9. Adverbials
10. Special lexical classes
A discriminant analysis was used to see how well the variables accounted for two identified proficiency groups. Results showed that the 28 variables divided learner essays into 2 groups with 82% accuracy. Higher level learners used more textual features than lower level learners. The data also supports earlier research that a factor involving passives, nominalizations, conjuncts and prepositions was positively correlated with holistic scores. All were used with greater frequency by learners of higher proficiency. The results also show that advanced learners have more lexical and syntactic tools available, use more cohesive devices, and devices which show pragmatic sensitivity than less advanced learners. A stepwise regression model showed the most influential variables were as follows with the percentage of variance accounted for by each feature in brackets:

1) Number of words (37.6%)
2) Synonymy / antonymy (6.0%)
3) Word length factor (3.3%)
4) Passives (2.5%)
5) 3rd person/impersonal pronouns (0.9%)

**Essay Length**

One striking feature of the Ferris experiment was the high percentage of variance accounted for by essay length in the stepwise regression analysis. It was the first factor to come out of the analysis accounting for 37.6% of the variance compared with 50.3% accounted for by the first 5 variables. Surprisingly Ferris doesn’t even mention this and concentrates the analysis on the other factors. This may reflect the generally uncomfortable relationship the assessment community has with essay length. It has long been recognized that human ratings are highly correlated with essay length, but many assessment organizations are at pains to emphasize that rating rewards quality rather than quantity. Accordingly, length is not included in the scoring guidelines for the TOEFL written test despite the fact that the assessment tool seems to be heavily influenced by it. The number of words either as the number of tokens or the number of types has consistently been shown to be a reliable indicator of grades in L1. Learners who write more tend to be scored higher. In L1, Page (1994) found that essay length was most important in short essays but less important in longer essays. Page recommended the fourth root of the number of words as the most efficient use of this variable. L2 essays have a tendency to be short and many studies have also suggested an association between essay length and holistic rating of L2 essays. We might also expect word total to be
a predictor of proficiency in a second language since fluency and lexical accessibility will have an effect on the number of words produced. Despite this correlation, it is not clear how much human judges are actually influenced by the length of an essay, if at all. It's doubtful that judges make a conscious decision related to essay length. Perhaps longer essays simply give the judge more chance to notice positive aspects of the essay. Or perhaps judges are not influenced by essay length at all and the correlation is due to the fact that longer essays just tend to be better ones. The case with automatic ratings is different. Here the influence of essay length is clear. Automatic systems are based on predicting the scores of human raters. Because essay length is the best single predictor of human scores, then just as lightning seeks the fastest way to earth, automatic systems will locate the most efficient features to predict human ratings. Lexical measures like the Type-Token Ratio (TTR) are often poor indicators of L2 essay scores because they have been designed to be independent of essay length. For example, the TTR is known to, in fact, have an inverse relationship with essay length and so tends to penalize learners who write more. Often, in order to get a purely lexical measure, TTR of a fixed sample can be calculated. This fails as a predictor of holistic scores by itself because although it doesn't penalize learners who write more, it doesn't reward them either. Total words (tokens), or the number of different words (types) are very strong indicators of learner proficiency and one of these should probably be included in any battery of measures used to assess written compositions.

**Other lexical features and statistics**

Essay length and the TTR have been mentioned but there are other lexical features and statistics of note. These include the Lexical Frequency Profile (LFP) (Laufer & Nation, 1999), P-Lex (Meara & Bell, 2001) and Lexical Signatures (Meara et al, 2002).

**Lexical Frequency Profile**

The LFP shows the percentage of words in a piece of writing belonging to different frequency levels. LFP profiles may be constructed according to learner proficiency. For less proficient learners, Laufer & Nation suggest 3 levels consisting of the 1000 most frequent words, the second 1000 most frequent words, and any other vocabulary may be suitable. However, for more advanced learners, they suggest 3 levels consisting of the second 1000 most frequent words, the University Word List (UWL), and words not in any of these 2 lists nor in the 1000 most frequent words may be more appropriate. Two drawbacks of the LFP are that it is both difficult to interpret and also difficult to use to compare learners. Most other lexical measures are unitary and so easy to compare. The LFP includes values for three or four categories and it becomes difficult to make judgments across these multiple categories. One way around this problem might be to convert these categories into a ratio of words in different levels of frequency.

**P-Lex**

Meara & Bell (2001) have developed P-Lex with an underlying assumption that learners with large vocabularies are more likely to use infrequent words than learners with smaller vocabularies. The method of P-Lex involves dividing an essay into 10-word segments and counting the number of easy and difficult words. Easy words are defined as words in the 1000 most frequent words plus
proper nouns and numbers. Other words are classified as difficult. The number of segments containing N difficult words is calculated. The distribution of N is usually skewed towards zero as the number of difficult words in each 10-word segment tends to be low. This distribution can be modeled using the Poisson distribution with mean lambda. Lambda values tend to vary from 0 to 4.5 with higher values indicating a higher use of infrequent words. Although an interesting idea, it has yet to be shown that P-Lex is superior to a simpler ratio of difficult words to easy words.

**Lexical Signatures**

Inspired by the work on LSA, Meara et al (2002) proposed Lexical Signatures to help assess lexical content of essays. Lexical Signatures are a way to investigate the unique lexical choices made by learners. This concept of uniqueness may have some potential in the evaluation of learner essays. It is likely that less proficient learners have more limited lexical choices than more proficient learners. More proficient learners might then be expected to use more lexical items not used by other learners in the group. These patterns may be picked up by a measure of uniqueness. However, there seem to be various ways one can measure uniqueness. For example, the simplest measure of uniqueness might be the number of words in an essay that are unique to that essay i.e. not present in any other essay in the cohort. The exact form and dynamic of a Lexical Signature approach needs more research. In particular, there are questions related to the reliability of scores (Mellor, 2003). As with P-Lex, it is not clear that the computationally complex Lexical Signature tells us anything more than a simple count of unique words.

**Distinctiveness**

Distinctiveness is a concept related to uniqueness (Mellor, 2006) but seems to have better mathematical properties. Distinctiveness measures the relative occurrence of lexical items across a cohort of essays. It was found to correlate with lexical proficiency more than uniqueness as measured by Lexical Signatures. The concept of distinctiveness is also superior to that of uniqueness for measurement purposes since it is more stable. The concept of uniqueness has a tendency to be unstable because the state of being unique can easily change. For instance, a word is unique if it only occurs in one essay but then if it is found in another essay, it is not unique. Also a word that occurs in only one other essay is treated the same way as one which occurs in all other essays. However, since distinctiveness is in effect the degree of uniqueness, it is more stable and a measure of distinctiveness is likely to be more reliable than one of uniqueness.

**Authorship Attribution**

The study of authorship attribution typically involves many lexical measures. Some of these may also have applications in automatic assessment. The following lexical features and statistics are often used in authorship attribution (Holmes, 1994).

**Lexical features**

1) word length
2) average number of syllables
3) sentence length
4) distribution of parts of speech - may be a percentage of parts of speech or as a ratio of certain classes, for example, adjectives to verbs. Studies of articles and pronouns have
shown particular promise.

5) function words – are context free and often show stability across work of an author.

Lexical statistics

6) Simpson’s Index D

\[ D = \sum r(r-1)V_r/(N(N-1)), \ (r = 1, 2, ...) \]

where \( V_r \) is the number of types occurring \( r \) times in a text length \( N \) words. D is sensitive to high frequency occurrences (usually a small number of function words) with low frequency occurrences (usually a large number of content words) having little influence.

7) Yule’s K

\[ K = 10^4(\sum r^2V_r - N)/N^2 \ (r = 1, 2, ...) \]

where \( V_r \) is the number of types occurring \( r \) times in a text length \( N \) words. K is directly related to Simpson’s Index D. Under certain assumptions, K is constant with respect to text length \( N \) (Yule, 1944).

8) Entropy H

\[ H = -100\sum p_i \log p_i / \log N \]

where \( p_i \) is the probability of the appearance of \( i^{th} \) word type. This gives a measure of diversity of any length text with absolute diversity as 100 and absolute uniformity as 0.

Authorship Attribution problems typically involve the analysis of long texts. Accordingly, some of the methods might be less appropriate in the analysis of shorter essays as found in L2 written assessment situations.

Word distributions

Various distributional properties of words have been investigated (Zipf, 1932) (Yule, 1944) (Baayen, 2001). Some of these were explored in Mellor (2008). In a comparison of native speaker essays and L2 learner essays, the proportion of hapax legomena (words appearing only once in an essay), Zipf rank and proportion of a and the, mean word length and fixed length TTR all were effective in distinguishing between native and learner essays.

Conclusion

Automatic assessment systems like PEG, e-rater or IEA have been very successful at reliably rating essays in large scale testing situations. However, they usually require a training set of essays rated by human judges on the same prompt. In creating a more versatile tool that can be used on a smaller number of essays and without a training set, a different method is required. An approach which judges essay quality in comparison to other essays rather than by some pre-judged standard is a possibility. Multivariate methods like Principal Components Analysis or clustering techniques could be employed to create a framework for such a comparison.

Many lexical measures such as essay length, the number of Hapax Legomena or TTR have been shown to correlate positively with holistic ratings but not well enough for measurement purposes by themselves. As a result, they have largely been discredited. However, perhaps they shouldn’t be discarded so quickly. In fact, some scores may be more accurate and reliable than others. Extreme values are likely to accurately identify relatively good or relatively poor essays while middling
values may be of less use in predicting proficiency. This type of measure may still have a role in assessment if used carefully as one of a range of measures. For example, a text with a very high or very low value on a particular lexical measure could receive a preliminary assessment of high or low quality which would be subsequently backed up or contradicted by other measurement evidence. Other measures such as lexical error which may be critical predictors of essay quality can also be incorporated into models where proxy measures can be identified.

References
Latent Semantic Analysis @ CU Boulder, [http://lsa.colorado.edu/](http://lsa.colorado.edu/)