http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (24): 7977-7982, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.7977.7982

Combination of Multiple Regressoion and Text Categorization in Automated Essay Scoring of College English Writing

Shili Ge
School of English for International Business, Guangdong University of Foreign Studies, Guangzhou, China

Abstract: In order to determine the joint effect of two methods, multiple regression and text categorization, in Automated Essay Scoring (AES) of English writing by Chinese college students, a joint model involving both methods is constructed, applied and evaluated. A corpus of college English writing containing 660 compositions is used as the subject and it is further divided into the training set and testing set. First, a multiple regression model is constructed and testing set compositions are automated scored as the baseline for this research. Then, a joint model of multiple regression and text categorization is constructed mainly using phrasal features. At last, final results are achieved and evaluated. The experimental results show that with the help of the joint model, all indexes including precision, recall, total accuracy and total wrong ratio of the AES model are improved. Though this model cannot be applied into practical use yet, it lays a solid foundation for further research in AES field.

Key words: Automated essay scoring, college english writing, multiple regression, text categorization, phrasal features

INTRODUCTION

In Automated Essay Scoring (AES), the most popular computational method is multiple regression while other methods, such as text cagetorization, are not discussed in depth, not to mention the joint effect of methods. Therefore, this study aims to explore the joint effect of two scoring methods, multiple regression and text classification in AES. After training on a training set of 440 compositions, the joint model is applied in the scoring of a testing set of 220 compositions to test the validity of the mixed model in AES for Chinese college students' English writing.

Traditionally, people consider "language text as the result of a very large number of complex choices. (…) This is probably the normal way of seeing and describing language. It is often called a 'slot-and-filler' model, envisaging texts as a series of slots which have to be filled from a lexicon which satisfies local restraints" (Sinclair, 1991). These restraints are mainly grammatical. This is the so called "open-choice principle" in linguistic communication. This priciple is constructed around lexicon and grammar, but Sinclair points out that words "do not occur at random in a text", "The choice of one word affects the choice of others in its vicinity. Collocation is one of the patterns of mutual choice and idiom is another. The name given to this principle of organization (of language) is the idiom principle (Sinclair, 1991)." In other words, "the language user has available to him a large number of preconstructed or semi-preconstructed phrases that constitute single choices, even though they appear to be analysable into segments (Partington, 1998)." First langauge acquisition studies show that through the acquisition, holistic memorization and application of prefabricated phrases in communciation, children can summarize and then acquire the construction rules of these phrases, so as to form grammatical ability and store the whole prefabricated expressions in their mental lexicon (Nattinger and DeCarrico, 1992). The cornerstone of "idiom principle" is linguistic segments that combine different lexical and grammatical functions. There are still disagreements in the theoretical definition, identification and classification of this linguistic unit, even different terms are adopted in different researches, such as holophrases, lexical phrases and chunks. In this research, they are called "phrases" but different names are adopted in citation of different literature.

EXPERIMENT OF AES WITH JOINT METHODS

Preparation of scoring experiments: In order to test the effect of joint methods in AES, the testing materials, namely, the compositions have to be prepared first. Then, phrases are extracted from the compositions automatically and the results are evaluated. Finally, together with lexical

features, phrasal features are applied in the automated scoring of compositions. These are the steps in any machine learning algorithm such as Zhang *et al.* (2012).

First, composition set should be prepared.

The Chinese college students' compositions in this research come mainly from the sub-corpus, st3, of Chinese Learner English Corpus (CLEC) and partly from the composition collection by Guan and Chen (2004). There are totally 660 compositions including more than 106,000 words. Word spelling errors and simple grammar errors, such as subject verb agreement, are identified and corrected in the phase of pre-processing. All compositions are scored by two raters with inter-rater coefficient of 0.761. With the help of a third rater, the discrepancy of scoring is resolved. In pre-processing, compositions are also POS tagged with Stanford NLP Group Part-of-Speech tagger (Toutanova, 2009), in which the tag set of Penn Treebank is adopted. The accuracy of POS tagging is 97.6%, which is higher than the acceptable POS tagging accuracy. The 660 compositions include 60 compositions with a score level of 1 and 150 compositions with a score level of 2 to 5, respectively. All these compositions are randomly divided into training set, which includes 2/3 of all compositions in each score level (440 compositions in sum) and testing set, which includes 1/3 of all compositions in each score level (220 compositions in sum), for model construction and testing of AES.

Second, phrasal features have to be extracted and evaluated.

The program of phrase identification is keyword based. When a POS tagged sentence is input, words are scanned one by one from the beginning and checked in the base. Once a word is found in the base as an entry, the matching program is started.

Every group of patterns in the base are listed in sequence and matched from the first to the last when needed. Once a pattern is matched, the matching program reports the match result and starts from the next word behind the matched string until the end of the sentence. The detailed algorithm is as following:

```
1   for all compositions do
2       for all sentences in a composition do
3           for all words in a sentence do
4               if a word in entry of ruleset and the rule
match the sentence:
5                   output the phrase and skip
to the next word behind the phrase
6               else:
7                   skip to next word
8           end
9       end
10  end
```

The evaluation criteria of phrase identification are precision and recall, which are estimated using counts of hits, false positives and misses. A hit occurs when the program identifies an actual phrase, no matter correct usage or wrong usage. A miss occurs when the program fails to identify a phrase. A false positive occurs when a non-phrase is identified as a phrase, a wrong used phrase is not identified, or a correct used phrase is identified as wrong use.

Precision is the proportion of identified items that are, in fact, phrases, no matter correct or wrong use:

$$\text{Precision} = \frac{\text{Hits}}{\text{Hits} + \text{False\_Positives}} \quad (1)$$

It measures how often the program is correct when it reports that a phrase has been found. Recall is the proportion of actual phrases that have been identified:

$$\text{Recall} = \frac{\text{Hits}}{\text{Hits} + \text{Misses}} \quad (2)$$

It measures the program's coverage, i.e., the fraction of phrases that the program has identified.

The program identified 7384 phrases in the experimental composition set including 6662 correct usages and 722 wrong usages. In order to gain an accurate evaluation, all 660 compositions were read through by two experienced English teachers to count hits, false positives and misses. The kappa statistic (Chodorow and Gamon, 2010) calculated is 0.71, which means that a substantial agreement exists between the two raters. With careful discussion, two raters settled the disagreement and the final results achieved are as follows: hits, 7109; false positives, 275; and misses, 1457. Then, the precision and recall calculated are 96.28 and 82.99%, respectively.

At last, the experiments of AES are conducted and before the expriments the evaluation criteria have to be set first.

In this study, the results of automated scoring are compared with the calculation of the following parameters: Precision and recall of each score level and total precision and total wrong ratio. They are defined as following:

$$\text{Precision of each score level} = \frac{\text{No. of score} \times \text{Compositions which are scored as X}}{\text{Total no. of compositions which are scored as X}} * 100$$

$$\text{Recall of each score level} = \frac{\text{No. of score X compositions which are scored as X}}{\text{Total no. of compositions with score X}} * 100$$

$$\text{Total accuracy} = \frac{\text{No. of accurately scored composition in the testing set}}{\text{Total no. of compositions in the testing set}} * 100$$

$$\text{Total wrong ratio} = \frac{\text{No. of compositions with an AES score which is 2 points larger or smaller than is real score}}{\text{Total no. of compositions in the testing set}} * 100$$

In the above formulae, X= 1, 2, 3, 4, or 5.

The higher the precision for each score level, the more reliable the score assigned by the AES model to the composition is. For example, the precision of a certain score level X is 50%. While composition A is scored X, then the probability that A really has a score X is 50%.

The higher the recall of each score level, the more contribution they have to the total accuracy.

The higher the total accuracy, the more accurate the model is and thus, the more practical and empirical to put into real application.

The lower the total wrong ratio, the less unacceptable error the AES model possesses. Thus, the acceptability of the AES model in real application is increased.

Scoring experiments and evaluation: The experiment of AES using multiple regression is first carried out.

There are usually 5 steps in an AES experiment: feature selection, algorithm design, model training, automated scoring and model evaluation.

The most often used method in AES researches and systems is multiple regression, such as PEG, e-rater and Liang (2012) Therefore, this method is first adopted to train a scoring model as the basis of automated scoring for this research. The features involved in multiple regression model include both lexical and phrasal features. Lexical features involved in this research are lexical distribution, average word length and composition length. Lexical distribution (LD) has long been believed to be closely related to writing quality (Laufer and Nation, 1995). LD is measured based on a certain word list. There are many different word lists and they have different validity for different types of writing. In this research, a specifically revised word list from Laufer, B. *et al* (1995) is adopted, which includes 4 categories of words, wd1 to wd4, from the most frequently used words to rare words. Average word length (AWL) is a key lexcial feature adopted in e-rater (Enright and Quinlan, 2010). It is also an important index for lexical complexity. Another lexical feature is the total number of words in a composition, that is, the composition length (CL). "Across many studies, the length of an essay strongly predicts the quality score a human rater will assign it" (Enright and Quinlan, 2010). Phrasal features involved in this research are the total number of phrases, the number of verb phrases and the number of wrongly used phrases in a composition.

The steps of multiple linear regression are as following:

- Input compositions from the training set one by one
- Segment composition into single words
- Count total word number, namely, CL
- Calculate the average word length (AWL)
- Count the number of words in different word level according to word list by Li and Ge (LD) (Li and Ge, 2008)
- Count the number of phrases (ttlphrase), verb phrases (verbphrase) and wrong phrases (wrongphrase)
- Output the composition score (CS), LD, AWL, CL, ttlphrase, verbphrase and wrongphrase
- Take CS as dependent variable and LD, AWL, CL, ttlphrase, verbphrase and wrongphrase as independent variables and input them into multiple linear regression to get the regression formula as AES model

After extract all the values of above variables from the training set with self-coded Python programs, the regression equation is achieved as following:

$$\text{Score} = -1.392 - 0.027 * wd2 - 0.015 * wd3 + 0.024 * wd4 + 0.125 * AWL + 0.032 * length + 0.060 * ttlphrase + 0.017 * verbphrase - 0.144 * wrongphrase$$

The key elements of the multiple regression model are listed in Table 1.

After applying this model to the testing set, each composition receives a machine assigned score and the score is evaluated based on its real score. The results are also listed in Table 2.

Table 1: Elements of regression equation of the training set

Elements	Feature
Coefficient of determination	0.445
Adjusted coefficient of determination	0.423
ANOVA F	46.745
ANOVA P	0.000
Maximum partial correlation coefficient	Average word length
Outlier	Yes

Table 2: Accuracy of multiple regression model

Score level	Evaluation criteria	Multiple regression method
1	Precision	83.33
	Recall	25.00
2	Precision	47.62
	Recall	40.00
3	Precision	45.78
	Recall	76.00
4	Precision	35.71
	Recall	50.00
5	Precision	78.95
	Recall	30.00
Total accuracy	43.18	
Total wrong ratio	13.18	

Table 2 shows that the total accuracy is a little over 43% and the total wrong ratio is about 13%. These results provide the basis for the combinaitonal experiment of different AES methods.

Before the AES experiment using text categorization method, a general review is needed.

There are many different approaches to the problem of text categorization as mentioned in (Xu *et al*., 2012), but they usually contain three steps: the preprocessing of text, namely, transforming text into computerized models including VSM and bag of words models; machine learning and generating classifiers; and applying classifiers in text classification.

The generation of classifiers and classification procedures are as following:

- Define feature space F = { f_1, f_2,..., f_n} and text d_{new} is represented as d_{new} = (s_1, s_2,..., s_n). n refers to the dimensions of the feature space. The ith dimension is corresponding to the feature f_i. s_i represents f_i if f_i presents in the text. If presents, s_i is 1, otherwise 0
- Assume category set C = (c_1, c_2,..., c_m). For a category unknown text d_{new}, the similarity of d_{new} and text category c_i can be represented as:

$$T_i(d_{new}) = \prod_{j=1}^{n}\left(s_j P(f_j \mid c_i) + (1-s_j)(1-P(f_j \mid c_i))\right) \quad (3)$$

Here, $P(f_j|c_i)$ is the probability of a certain feature presents in category c_i.

$P(f_j|c_i)$ can be estimated from training set according to the following Eq:

$$P(f_j \mid c_i) \approx \frac{1+\sum_{k=1}^{|D|} in(f_j, d_k) y(d_k, c_i)}{2+\sum_{k=1}^{|D|} y(d_k, c_i)} \quad (4)$$

Here d_k refers to one of the text in the training set. in $(f_j,d_k)\in\{0,1\}$ means if feature f_j presents in text d_k, the value is 1, otherwise 0. $y(d_k, c_i)\in\{0,1\}$ means if text d_k belongs to category c_i, the value is 1, otherwise 0.

In the categorization of composition texts, based on Eq. 3, we can calculate $T_i(d_{new})$, namely, the probable estimate of each ungraded composition belonging to any score category. Based on the scoring results using multiple regression, we can calculate the probability of each composition belonging to every score category c_i while it is scored as c_j', namely:

$$S_i(d_{new}) = P(c_i \mid c_j') \quad (5)$$

In Eq. 5, i, j$\in${1, 2, 3, 4, 5}. c_j' is the scoring results based on the previous mentioned features. $P(c_i|c_j')$ is the probability that a composition is scored as different levels from 1 to 5.

The combination of regression results and text categorization is fulfilled this way:

$$C(d_{new}) = \underset{i}{\mathrm{ArgMax}}(\alpha S_i(d_{new}) + (1-\alpha)T_i(d_{new})) \quad (6)$$

In the Equation, parameter α is the weight in the combination of multiple regression and text categorization. $S_i(d_{new})$ is already known; after calculating $T_i(d_{new})$ using text categorization method and adjusting parameter α, the best scoring of each composition can be achieved. The adjusting of α can be based on total accuracy or teaching requirements. What should be noticed is that α is training set related, namely, the best α value in this research may not be the best value for other composition set, especially compositions with very different linguistic features. However, the method of adjusting α can be generalized and applied for different text materials. With the optimal α value achieved by the training, a desired scoring result can be achieved.

Finally, the experiment of AES combining multiple regression and text categorization is conducted as following:

Feature extraction: To classify texts, the texts should be analyzed first and distinguishing features should be extracted. Text categorization usually uses lexical features, especially content based text categorization. In this research, in order to score compositions, phrasal features are adopted for text categorization. The phrases used in college English writing, especially those high frequency verb phrases and most non-verb phrases present in many compositions have little relavancy with composition content but strong relavancy with the English proficiency of the writer. Therefore, these phrases can be adopted as distinguishing features.

Preparation of categorization: The procedures of preparation are as following:

- Construct feature space F = {f_1, f_2,... f_{616}}, in which F consists of 616 phrases selected
- For every composition in the 440 compositions of training set, construct text vector d = {s_1, s_2, ..., s_{616}} If phrase f_1 presents in the composition, s_1 is 1, otherwise 0. In this way, 440 vectors containing 0 and 1 are formed
- By far, the preparation for categorization has finished and all the 220 compositions in the testing set are classified, namely, automated scored

Table 3: AES performance of the joint model

Score level	1	2	3	4	5
Accuracy	100.00	83.33	78.72	65.08	75.81
Recall	30.00	70.00	74.00	82.00	94.00
Wrong ratio	0.00	0.00	4.26	15.87	16.13
Total accuracy	75.45				
Total wrong ratio	10.00				

All phrasal features are extracted from the 220 compositions in the testing set and the procedures of preparation are followed to construct 220 vectors that represent all the 220 compositions in the testing set. Then, the probable estimates $T_i(d_{new})$ of every composition belonging to each of the 5 score level are calculated.

After calculating $S_i(d_{new})$ and $T_i(d_{new})$, the parameter α are set from 0.1 to 0.9 and increased by 0.1, the values of $C(d_{new})$ are calculated. When α is set to 0.4, the values of $C(d_{new})$ have the highest accuracy, which are listed in Table 3.

Comparying Table 3 and 2, we can see that the accuracy and recall for all score levels are highly raised except for the accuracy of score level 5. The total accuracy is increased over 75% and total wrong ratio is decreased below 10%. Another important phenomenon can be observed from Table 3 is that the central tendency of recall, namely, the recall of middle score level is high while the high and low level are low, is changed to another pattern, that is, the recall raises high for high score levels. However, there is another tendency that the wrong ratio is increased along with the score level. Combining these two tendencies, the combination of two methods raise the total recall, but there is also a tendency that the score of every composition is raised to some extent, which leads to the decrease of total wrong ratio and scoring accuracy for low score level compositions, but high wrong ratio for high score level compositions. That means the compositions that are automated scored low scores are real poor compositions, but not all the compositions that are automated scored high scores are really good compositions.

CONCLUSION AND DISCUSSION

The review of related literature shows that the methods used in automated essay scoring have a significant role for the performance of the constructed model or system. With the help of college English writing corpus and large English corpus, phrase identification rules are coded, including verb phrases, non-verb phrase and wrongly used phrases. Based on the rule set, phrases can be accurately identified from college English compositions, which form the solid base for furth automated essay scoring research. And then, two frequently used methods, multiple regression and text categorization are employed to construct a mixed model and the scoring results are evaluated.

The AES experimental research in this study includes two stages. The first stage is based solely on multiple regression method and the second mixed with text categorization mainly based on phrasal features. Experiments show that when text categorization method is adopted and work together with multiple regression, the precisions and recalls for almost all score levels are increased and certainly the total accuracy is increased, too. Though the total accuracy is not high enough for practical use in daily English teaching, this study laid a solid foundation for further AES research.

ACKNOWLEDGMENT

This study was financially supported by the National Social Science Fund (No. 13BYY097).

REFERENCES

Chodorow, M. and M. Gamon, 2010. Automated Grammatical Error Detection for Language Learners. Morgan and Claypool Publishers, USA., ISBN: 9781608454709, Pages: 122.

Enright, M.K. and T. Quinlan, 2010. Complementing human judgment of essays written by English language learners with e-rater scoring. Language Test., 27: 317-334.

Guan, X. and J. Chen, 2004. College English Writing. Jilin University Press, Jilin.

Laufer, B. and P. Nation, 1995. Vocabulary size and use: Lexical richness in L2 written production. Applied linguistics, 16: 307-322.

Li, Y. and S. Ge, 2008. The validity of word list in automated essay scoring for college students. Foreign Languages Their Teach., 10: 48-52.

Liang, M., 2012. The Research and Development of Automated Essay Scoring System for Large Scale English Test. Higher Education Press, Beijing, China.

Nattinger, J.R. and J.S. DeCarrico, 1992. Lexical Phrases and Language Teaching. Oxford University Press, Oxford, ISBN: 9780194371643, Pages: 218.

Partington, A., 1998. Patterns and Meanings: Using Corpora for English Language Research and Teaching. John Benjamins Publishing, Amsterdam, ISBN: 9789027222718, Pages: 162.

Sinclair, J., 1991. Corpus, Concordance, Collocation. 3rd Edn., Oxford University Press, Oxford, ISBN: 9780194371445, Pages: 179.

Toutanova, K., 2009. Stanford long-lines part-of-speech tagger. http://nlp.stanford.edu/software/tagger.shtml

Xu, B., X. Guo, Y. Ye and J. Cheng, 2012. An improved random forest classifier for text categorization. J. Comput., 7: 2913-2920.

Zhang, Q.Y., P. Wang and H.J. Yang, 2012. Applications of text clustering based on semantic body for Chinese spam filtering. J. Comput., 7: 2612-2616.