Automated Scoring of Chinese Engineering Students’ English Essays

Ming Liu, School of Computer and Information Science, Southwest University, Chongqing, China
Yuqi Wang, School of Computer and Information Science, Southwest University, Chongqing, China
Weimei Xu, College of International Studies, Southwest University, Chongqing, China
Li Liu, School of Software Engineering, Chongqing University, Chongqing, China

ABSTRACT

The number of Chinese engineering students has increased greatly since 1999. Rating the quality of these students’ English essays has thus become time-consuming and challenging. This paper presents a novel automatic essay scoring algorithm called PSO-SVR, based on a machine learning algorithm, Support Vector Machine for Regression (SVR), and a computational intelligence algorithm, Particle Swarm Optimization, which optimizes the parameters of SVR kernel functions. Three groups of essays, written by chemical, electrical and computer science engineering majors respectively, were used for evaluation. The study result shows that this PSO-SVR outperforms traditional essay scoring algorithms, such as multiple linear regression, support vector machine for regression and K Nearest Neighbor algorithm. It indicates that PSO-SVR is more robust in predicting irregular datasets, because the repeated use of simple content words may result in the low score of an essay, even though the system detects higher cohesion but no spelling error.

KEYWORDS

Computer Uses in Education, Language, Text Analysis

1. INTRODUCTION

Rating essays is a costly, laborious and time-consuming effort, which is especially true in China due to the large number of students. Statistics show that the number of college students in China has soared to twenty-six million in 2013 (Bureau of Statistics of China, 2013) including more than ten million engineering students, making up the largest proportion of English as Second Language (ESL) learners worldwide. Since 1987, the writing test has become an important aspect in the College English testing in China. Essay writing is the fourth part of these tests. Trained English teachers manually rate the essays due to the nature of subjectivity and creativity of essay writing. However, rating essays is a time-consuming effort and at the same time the ratings are prone to the subjective judgment of the trained English teachers leading to inconsistent and unreliable scores due to the impact of fatigue, deadlines or biases.

Computer aided assessment (CAA) has become an important educational technology (Clark & Byl, 2007) since it reduces teacher work-loads (Peat, Franklin, & Lewis, 2001), provides timely feedback to students (Sheard & Carbone, 2000), reduces in educational material development and delivery costs (Jefferies, 2000), and proliferate online education (White, 2000). Research in computer-based essay scoring, referred to as automatic essay scoring (AES), has been a real and viable alternative...
and complement to human scoring for more than 40 years (Shermis & Burstein, 2003). AES systems
do not actually read or understand essays as humans do. Whereas human raters may directly evaluate
various intrinsic features, such as diction, fluency and grammar, in order to produce an essay score,
the AES systems rely on a statistical scoring model, which combines these features and approximates
a final machine-generated score of the essay. In general, the task of automated grading can be viewed
as a regression problem in which the objective is to find a set of features that represent the essays
and serve as inputs of the regression methods. Regression algorithms are utilized to estimate the
weights of each term (i.e. feature) in the regression equation so that the prediction performance can
be optimized with regard to the actual values of the variable to be predicted/explained by the model.

Many AES systems, such as e-rater and PEG (Attali & Burstein, 2006; Page, 1966; Warschauer &
Ware, 2006), based on a multiple linear regression model with predefined textual features extracted by
using computational linguistic tools. Another approach to AES is based on Latent Semantic Analysis
technique (Landauer, McNamara, Dennis, & Kintsch, 2007) such as Intelligent Essay Assessor (Foltz,
Streeter, Lochbaum, & Landauer, 2013; Landauer, Laham, & Foltz, 2003) and IntelliMetric (Elliot,
2003; Rudner, Garcia, & Welch, 2006). But, this approach requires a large training corpus for a
specific essay prompt. More recently, McNamara et al. (2015) proposed a hierarchical classification
approach to automated essay scoring. In this study, we extend the traditional linear regression model
to the non-linear regression model for automated essay scoring since the qualities of ESL writing do
not linear relationship with the textual features.

With the advanced development of natural language processing techniques, many intelligent text
analysis tools (Biber, 1988; McNamara, Graesser, McCarthy, & Cai, 2014; Pennebaker & Francis,
1999) were developed to analyze and extract rich textual features for building automated essay
scoring models. The Biber Tagger (Biber, Conrad, & Reppen, 1998) automatically computes features
for lexical sophistication (e.g., word length), cohesion and rhetorical features (e.g. conjuncts and
emphatics), grammatical features (e.g. nouns and verbs), and clause-level features (e.g. subordinations
and passives). Similar to the Biber Tagger, Coh-Metrix (McNamara et al., 2014) calculates a number
of text-based linguistic features related to lexical sophistication (word frequency, word concreteness,
word familiarity, polysemy, hypernymy), syntactic complexity (incidence of infinitives, phrase length,
number of words before the main verb), and cohesion (word overlap, semantic similarity, incidence
of connectives). LIWC is an automated word analysis tool which reports the percentage of words in a
text that are in certain psychological categories (Pennebaker, Booth, & Francis, 2007). The categories
include linguistic processes (e.g. pronouns, past tense), psychological processes (e.g., social processes,
cognitive processes, perceptual processes), personal constructs (e.g., work, religion), and paralinguistic
dimensions (e.g., speech disfluencies). With these tools, the AES systems show correlations with
human judgments of essay quality that range between .60 and .85 (McNamara, Crossley, & Roscoe,
2013; Rudner et al., 2006). In this study, we used the Coh-Metrix for textual feature extraction since
it was commonly used in many AES research (Crossley & McNamara, 2011).

Although several automated systems are already available, they are not specifically developed
for most ESL learners (Burstein & Chodorow, 1999; Lonsdale & Strong-Krause, 2003). Burstein and
Chodorow (1999) evaluated the performance of the e-rater system TM on Test of Written English
(TWE) essay responses written by non-native English speakers whose native language is Chinese,
Arabic or Spanish. They found that the interaction of language group was significant
\[
\left( F_{[4,1128]} = 12.397, \ p < .001 \right), \text{ reflecting higher scores for human scores than for the system in some groups (e.g., Spanish) and lower scores for human scores than for e-rater in others (e.g. Chinese).}
\]
Chinese ESL Researchers (Ge, 2010; Liang, 2004; X. Liu, 2008) have identified some issues when
applying existing AES systems in Chinese ESL context. For example, ESL students have their own
OLnet: A New Approach to Supporting the Design and Use of Open Educational Resources
www.igi-global.com/chapter/olnet-new-approach-supporting-design/40730?camid=4v1a

Navigation and Visualisation Techniques in eLearning and Internet Research
Sue Fenley (2010). Looking Toward the Future of Technology-Enhanced Education: Ubiquitous Learning and the Digital Native (pp. 55-87).
www.igi-global.com/chapter/navigation-visualisation-techniques-elearning-internet/40727?camid=4v1a