A Ranked-based Learning Approach To Automated Essay Scoring

Hongbo Chen
School of Computer and Control Engineering
Graduate University of the Chinese Academy of Sciences
Beijing, China
chenhongbo11@mails.gucas.ac.cn

Ben He
School of Computer and Control Engineering
Graduate University of the Chinese Academy of Sciences
Beijing, China
benhe@gucas.ac.cn

Tiejian Luo
School of Computer and Control Engineering
Graduate University of the Chinese Academy of Sciences
Beijing, China
tjluo@gucas.ac.cn

Baobin Li
School of Computer and Control Engineering
Graduate University of the Chinese Academy of Sciences
Beijing, China
libb@gucas.ac.cn

***Abstract*—Automated essay scoring is the computer techniques and algorithms that evaluate and score essays automatically. Compared with human rater, automated essay scoring has the advantage of fairness, less human resource cost and timely feedback. In previous work, automated essay scoring is regarded as a classification or regression problem. Machine learning techniques such as K-nearest-neighbor (KNN), multiple linear regression have been applied to solve this problem. In this paper, we regard this problem as a ranking problem and apply a new machine learning method, learning to rank, to solve this problem. We will introduce detailed steps about how to apply learning to rank to automated essay scoring, such as feature extraction, scoring. Experiments in this paper show that learning to rank outperforms other classical machine learning techniques in automated essay scoring.**

***Keywords*-Automated essay scoring; Learning to rank; Feature extraction; Machine learning;**

I. Introduction

Automated Essay Scoring (AES) is defined as the computer techniques and algorithms that evaluate and score essays automatically [3]. In this paper, we mainly discuss automated essay scoring on English essays. In recent years, with the growing need of essay scoring in large-scale English test and in the teaching of English writing skill, automated essay scoring has become a hot issue in the research of natural language processing. Nowadays, large-scale English test has been widely spread in the world such as GRE, GMAT, TOEFL [1]. On one hand, the essay scoring task in such test costs huge human resources but the efficiency is low. On another hand, the essay score given by human rater is mostly determined by rater's personal will, emotion and energy. An essay scored highly by one rater may recieve a low score from another rater. Even the same rater probably gives different scores for the same essay at different times. Thus, the fairness of essay scoring cannot be guaranteed. What's more, automated essay scoring is also needed in the teaching of English writing skill. Usually, it is a challenging task for one teacher to finish the essay scoring of all student essays in a short time. Thus, students cannot get feedback on their essays in time, leading to the situation that it is hard for them to improve their writing skill [4] [5]. In such requirement background, researchers proposed automated essay scoring techniques. Automated essay scoring techniques has the advantage of fairness, less human resource cost and timely feedback [3].

In general, automated essay scoring is a machine learning problem [2]. More specifically, it is a supervised learning problem. Scored essays can be seen as labeled training data and unscored essays can be seen as unlabeled test data. The main process of automated essay scoring is to learn a scoring function or model from the training data and then use the scoring function or model to score essays in the test data. Previous solutions can be divided into mainly two categories: classification and regression [2]. When automated essay scoring is regarded as a classification problem, the score is seen as the class label. Classical classification algorithms like KNN are applied to solve this problem. When it is regarded as a regression problem, the score is seen as a comparable value. Classical regression algorithms like multiple linear regression are applied to solve this problem. In this paper, we regard automated essay scoring as a ranking problem and plan to solve this problem by learning to rank algorithms. Learning to rank is a family of supervised learning algorithms that automatically construct a ranking model or function to rank objects such as the retrieved documents [8]. The major advantage of learning to rank is its flexibility in incorporating diverse kinds of features into the process of ranking [8].

The major contributions of this paper are two-fold. Firstly, we are the first to apply learning to rank algorithms to automated essay scoring. Secondly, we make a comparison between the performance of learning to rank algorithms and other classical machine learning algorithms in automated

978-0-7695-4864-7/12 $26.00 © 2012 IEEE
DOI 10.1109/CGC.2012.41

essay scoring through extensive experiments.

The rest of this paper is organized as follows. In Related work and Background, we introduce previous work on automated essay scoring and give an briefly introduction to learning to rank. The Method section gives a detailed description of how we applied learning to rank algorithms to automated essay scoring. The Experiment Setup part describes the experimental data, main experiment steps and the evaluation measure of our experiments. The Experimental Result section gives the experimental result and the analysis on the result. Finally, we conclude this research and discuss the future trend of automated essay scoring.

II. Related Work and Background

A. Automated Essay Scoring

The research of automated essay scoring began in the 1960s. There are four mature automated essay scoring systems after 40 years' research. The main approach of these systems is to define a large number of objective measurable essay features and use machine learning algorithms to predict the score of essay, such as multiple linear regression, KNN [2].

The first automated essay scoring system, Project Essay Grading (PEG), is developed by Ellis Page in 1966 upon the request of the College Board of American. It is also the first automated essay scoring system widely used in testing companies, universities, and public schools. The PEG system extracts shallow text features from essays and uses multiple linear regression to learn the scoring function. The accordance between scores given by the PEG system and scores given by human raters is proved to be high [3]. However, PEG system only considers shallow text features while ignores the essay content, leading that it is easy to be cheated by students. For example, students who write long essays and long sentences tend to get a high score [1].

Another automated essay scoring system, Intelligent Essay Assessor (IEA), developed in the last of 1990s, scores essay by measuring the semantic feature [3]. The IEA system computes the semantic matrix of scored essays and unscored essays by a semantic text analysis method named Latent Semantic Analysis (LSA) [9]. Firstly, the system calculates out the cosine similarity between unscored essays and scored essay. Then, for each unscored essay, the system picks out the most n similar scored essays. The parameter n is defined by the system and can be adjusted. Finally, the system computes the mean weighted sum of scores of these n picked out scored essays as the final score for the unscored essay. The weight of each essay score is set as the respective similarity degree [1].

E-rater, developed by Educational Testing Services (ETS) in America in the last of 1990s, is one of the only two mature commercial automated essay scoring systems and has been currently put into use by ETS for essay scoring in the Graduate Management Admissions Test (GMAT) and the Test of English as a Foreign Language (TOFEL). The E-rater system extracts both shallow text features and deep text features, and then makes a linear regression like PEG [3] [4] [5].

Another mature commercial automated essay scoring system is Intellimetric, which is developed by Vantage Learning company in 2003. It is the first automated essay scoring system based on artificial intelligence. Like E-rater, it extracts more than 300 text features, including both shallow text features and deep text features. The method of feature extraction is complicated. It is a blend of artificial intelligence, natural language processing and statistical techniques [1] [3] [5]. However, Intellimetric learns the scoring function by complicated nonlinear mathematical model while E-rater learns the scoring function by linear regression. The accuracy of both E-rater and Intellimetric in essay scoring is higher than 97%. However, the details of how these two systems are developed are commercial secrets.

B. Introduction to Learning to Rank

Learning to rank, namely machine-learned ranking, is proposed to solve the ranking problem in information retrieval. In recent years, with the rapid development of Internet and fast growth of information, searching for the information we need in the Internet has become more and more difficult. Effective information retrieval system has become more and more important. The key problem in modern information retrieval system is how to rank the retrieved documents [8]. In traditional solution to this problem, a scoring function or model is constructed based on some measurement such as relevance degree. Then retrieved documents are ranked according to the output of the scoring function or model. There are many classical retrieval models to solve this problem, such as vector space model, boolean model, language model. Nowadays, with the development of machine learning techniques, researchers try to apply these techniques to solve the ranking problem and have proposed many innovative and effective ranking models. We call this research area learning to rank [8].

Learning to rank is a family of supervised learning algorithms that automatically construct a ranking model or function to rank objects such as the retrieved documents. The main framework of learning to rank includes four parts: input space, output space, hypothesis space, loss function [8]. The input space contains objects to be ranked and represented by feature vectors. The output space contains objects that have been ranked by the ranking model or function. The hypothesis space defines the class of ranking functions or models. Loss function, like other machine learning algorithms, measures the difference between the prediction generated by the ranking model or function and the ground truth label. In training process, a ranking function or model is learned from the labeled training data. In test process, the ranking function scores documents and then ranks documents according to the output score. The main advantage of learning to rank is the ability of integrating variety kinds of features of documents into the process of ranking [8].

It is widely accepted that there are three kinds of learning to rank algorithms, that is, the pointwise, pairwise, listwise approaches. We will introduce these three approaches below respectively [8].

In pointwise approach, ranking algorithms process a single document and the relation between documents is not considered. The input space contains feature vector of every single document. The output space contains the relevance degree of each single document. The hypothesis space contains functions that predict the relevance degree of the document in input space. In pointwise approach, ranking can be modeled as regression, classification and ordinal regression. So the loss function can be regression loss, classification loss and ordinal regression loss [8].

In pairwise approach, ranking algorithms process pairs of documents. The input space of the pairwise approach contains pairs of documents both represented by feature vectors. The output space contains the pairwise preference between each pair of documents. The hypothesis space includes the ranking function that predicts the preference relation between single pair of documents [8] [11] [12]. Ranking is usually modeled as a pairwise classification problem in many pairwise ranking algorithms. Thus, the loss function is always a classification loss.

In listwise approach, ranking algorithms process a list of documents. The input space of the listwise approach contains a list of documents and the corresponding query. The output space contains a list of ranked documents that correspond to the same query. The hypothesis space includes the ranking function that predicts the ranking of the list of documents. The loss function aims at measuring the accordance between predicted ranking list and the ground truth label [8] [10] [12].

III. Method

In this section, we will firstly give a detailed description of how to apply learning to rank algorithms to build an automated essay scoring system. Then, we discussed the details of the key step in learning to rank, that is, feature extraction.

A. Learning to Rank for Automated Essay Scoring

There are two key steps in automated essay scoring: extracting essay feature, learning a scoring function or model by machine learning algorithms that can score essays by comprehensively considering these essay features. The major advantage of learning to rank is its flexibility in incorporating diverse kinds of features into the process of ranking. Thus, we can try to apply learning to rank to incorporate various kinds of essay features into the process of essay ranking. Then what we get from the output of learning to rank algorithms is a ranked list of essays. We can see the ranked list of essays as the ranking of essays' writing quality. If essay A is ranked in front of essay B, we can judge that essay A is written better than essay B. Finally, we should transform the score output by ranking function into practical score in some way.

There are many literatures have proved that listwise approach and pairwise approach perform better than pointwise approach. In this paper, we mainly discuss the application of learning to rank with listwise approach and pairwise approach on automated essay scoring. There are several kinds of pairwise ranking algorithms and listwise ranking algorithms. In this paper, we choose Support Vector Machine for ranking (SVMrank) as the pairwise ranking algorithm to be used and LambdaMart as the listwise ranking algorithm to be used [12].

Ranking SVM, proposed by Joachims (2003), is an application of the conventional support vector machine, which is used to learn a ranking function. Ranking SVM algorithm includes three main steps, document feature extraction and representation, partial-order relationships defined on document pairs, conventional SVM optimization, namely minimizing the sum of empirical loss and regularizer [8] [11] [12].

LambdaMart, proposed by Wu et al. (2008), is a combination of the ranking model LambdaRank based on RankNet and the boosted tree optimization method MART. It won Track 1 of the 2010 Yahoo! Learning to Rank Challenge. LambdaMart is a listwise ranking algorithm and aims at optimization IR evaluation measures, such as Normalized Discounted Cumulative Gain (NDCG). As these evaluation measures are not continuous and thus non-differentiable, LambdaMart solves this problem by defining a smooth approximation to the gradient of the cost function [8] [10] [12].

B. Feature Extraction

Feature extraction is the key step in learning to rank. When applying learning to rank to automated essay scoring, good features that can reflect the writing quality of essays well produce good prediction result.

In essay scoring, human raters usually consider several aspects of essay, such as term usage, sentences quality, the fluency and richness of content. Then they score the essay by comprehensively considering these writing features. Similarly, in automated essay scoring system we must extract measurable features of an essay from different aspects like what human raters do. We divide extracted essay features in this paper into three categories, that is, term usage, sentence quality, content fluency and richness. Table 3.1 lists these features and detailed description of them are given below.

- Term usage:
  - a. Statistics of variety kinds of part-of-speech. We use the Stanford Natural Language Processing

Table I
TABLE 3.1 EXTRACTED ESSAY FEATURES IN THIS PAPER

No.	Feature	Description
Term usage		
1	prep	#Prepositions
2	modal verb	#Modal verbs
3	gerund	#Gerunds
4	wordlength4	#Words with length in characters>4
5	wordlength6	#Words with length in characters>6
6	wordlength8	#Words with length in characters>8
7	wordlength10	#Words with length in characters>10
8	wordlength12	#Words with length in characters>12
9	wordlevel1	#Words in level 1
10	wordlevel2	#Words in level 2
11	wordlevel3	#Words in level 3
12	wordlevel4	#Words in level 4
13	wordlevel5	#Words in level 5
14	wordlevel6	#Words in level 6
15	wordlevel7	#Words in level 7
16	wordlevel8	#Words in level 8
17	spellingerror	#Spelling errors
Sentence quality		
18	sentencelength5	#Sentences with length in words>5
19	sentencelength10	#Sentences with length in words>10
20	sentencelength15	#Sentences with length in words>15
21	sentencelength25	#Sentences with length in words>25
22	attributive	#Attributive clauses
23	adverbial	#Adverbial clauses
24	prepositional	#Prepositional phrases
25	grammarerror	#Grammar errors
content fluency and richness		
26	lsa1	Mean similarity degree to essays graded 1
27	lsa2	Mean similarity degree to essays graded 2
28	lsa3	Mean similarity degree to essays graded 3
29	lsa4	Mean similarity degree to essays graded 4
30	lsa5	Mean similarity degree to essays graded 5
31	lsa6	Mean similarity degree to essays graded 6
32	essaylength	Essay length
33	conjunction	#Conjunction words

  tools to parse the essay. These tools are open-source software developed by Stanford University. The tools will assign parts of speech to each word and phrase. We mainly count the number of preposition, modal verb, gerund. Good sentence tend to use more prepositions modal verbs and gerund [13].

  - b. Statistics of the length of terms. The change of term length reflect term using. We count the number of word whose length is more than 4, 6, 8, 10, 12 respectively [6]. The changing of term length reflects the complexity of term usage.
  - c. Statistics of variety levels of words. All words in Webster dictionary are divided into 8 parts according to the College Board Vocabulary Study conducted by Hunter M. Breland, Robert J. Jones and Laura J. [7]. Each part corresponds to an English level. Words in level 8, which are always used by professional writers, correspond to the highest English level. Words in level 1, which always occur in essays written by English beginners, correspond to the lowest English level. We count the number of words belonged to each part.
  - d. Statistics of spelling errors. We count the number of word which is not appeared in the Webster dictionary.

- Sentence quality:
  - a. Statistics of sentences' length. The change of sentence length reflects the sentences' quality. We count the number of sentence whose length is more than 5, 10, 15, 25 respectively. The changing of sentence length reflects the complexity of sentences.
  - b. Statistics of variety kinds of phrases and clauses. Good sentence always contains various kinds of phrases and clauses [13]. We count the number of them respectively. We mainly focus on attributive clause, adverbial clause, prepositional phrase, etc. The recognition of clauses and phrases can be accomplished by Stanford NLP tools.
  - c. Statistics of grammar errors. We count the number of grammar errors by Linked Grammar which is a Natural Language Processing tool with grammar checking function [13]. It is an open-source software developed by Carnegie Mellon University.
- Content fluency and richness:
  - a. Semantic vector similarity degree. We construct a essay-semantic matrix by latent semantic analysis (LSA) techniques [9] and then calculate the cosine similarity between unscored essays' semantic vectors with scored essays' semantic vectors. We calculate the mean value of the similarity degree of essays in each score level. For example, if there are 6 score level 1, 2, 3, 4, 5, 6, we calculate the mean value of the similarity degree of essays scored 1 to 6 respectively.
  - b. Essay length. Essay length reflects the richness of essay content.
  - c. Statistics of conjunction words. The number of conjunction words reflects the fluency of essay [13]. We count the number of conjunction words by Stanford NLP tools.

C. *Score transformation*

The score output by the ranking model is a real number, even may be a negative number, and it is not in the interval of practical score. We must transform it into correct form and correct interval, such as integer range from 1 to 6. Firstly, we take training set as the test set and get a list of scores of essays in training set. Then, we apply the ranking model to the test set and get a list of scores of essays in test set. Here we call score given by ranking model relative score. Both scores in the first list and the second list mentioned above are relative scores given by ranking model. Of course, every essay in the training set has a practical score and every essay in the test set need a practical score. Finally, for each relative score in the second list, we insert them to the first list and pick out the ten most nearest relative scores and also the ten corresponding practical scores. We compute the average of the ten corresponding practical scores as the final practical score for the essay in the second list. Thus, the transformation of score is finished.

IV. The Experimental Setup

In this section, we give out the details of the major components in our experiment, that is, experimental data, algorithms in experiment, evaluation measurement and cross-validation.

A. Dataset

The dataset we used in our experiment comes from the Automated Essay Scoring Competition held by the Willian and Folra Hewlett Foundation. The William and Flora Hewlett Foundation is a private foundation, established by Hewlett-Packard co-founder William Redington Hewlett and his wife Flora Lamson Hewlett in 1966. The Hewlett Foundation awards grants to support educational and cultural institutions and to advance certain social and environmental issues. It is one of the largest grant-giving institutions in the United States, with assets of over $7 billion. This year the Hewlett Foundation is sponsoring the Automated Student Assessment Prize. Hewlett is appealing to data scientists and machine learning specialists to help solve an important social problem, automated essay scoring. They need fast, effective and affordable solutions for automated grading of student-written essays [14].

Datasets in this competition consist of eight essay sets. Each essay set was generated from a single prompt. Selected essays range from an average length of 150 to 550 words per response. Some of the essays are dependent upon source information and others are not. All responses were written by students ranging in grade levels, from Grade 7 to Grade 10. All essays were hand graded and were double-scored. Each of the eight datasets has its own unique characteristics [14]. The whole datasets are divided into three parts, training set, validation set, test set. Training set accounts for 50% of the whole dataset. Every essay in the training set contain ground truth label. Validation set accounts for 30% and test set accounts for 20%. Every essay in both valid set and test set does not contain ground truth label. Each of the eight essay sets are distributed into these three datasets proportional.

B. Algorithms in Experiment

In this section, we will introduce the four algorithms in our experiment: Ranking SVM, LambdaMart, KNN, multiple linear regression.

1) LambdaMart: LambdaMart is a listwise ranking algorithm. The algorithm processes a list of documents and the loss function of LambdaMart is defined at the query-level. It aims at the optimization of common information retrieval measurements, such as NDCG [10]. However, there is no concept of query in automated essay scoring as every essay title is independent. We just make a comparison between the performance of LambdaMart and SVMrank when there is no concept of query. The parameter trees.num-leaves is set as 7. Trees.min-instance-percentage-per-leaf is set as 0.25. Boosting.learning-rate is set as 0.05. Boosting.sub-sampling is set as 0.3. Trees.feature-sampling is set as 0.3. Boosting.num-trees is set as 2000. The learning algorithm is set as LambdaMart-RegressionTree and the evaluation-metric is set as NDCG.

2) Ranking SVM: Ranking SVM is a pairwise ranking algorithm. The algorithm processes a pair of documents. Ranking SVM is proposed based on conventional SVM. The main difference is that the constrains in Ranking SVM are defined on partial-order relationships within document pairs [12]. The kernel function of the Ranking SVM here is linear kernel function. And the parameter C, which controls the trade-off between empirical loss and regularizer, is set as 20.

3) KNN: When we apply KNN to automated essay scoring, the first thing we should do is to extract essay features and transform them to feature vectors. The features in the feature vector need not be what we defined in the last section. They can be occurrences of each word appeared in the essay or other kinds of features. Here we adopt the features defined in the last section. Then, KNN algorithm select out k scored essays in the training set which are most similar to the unscored essay. We measure the similarity degree between essays by cosine similarity between feature vectors. Finally, the unscored essay receives a score which is the similarity-weighted mean of the scores of the k nearest essays. The parameter k, the number of the nearest essays we select out, can be tuned by the training set. KNN is a common machine learning algorithm for classification and also a common algorithm used in automated essay scoring.

4) Multiple Linear Regression: Multiple linear regression is the most commonly used machine learning algorithm in automated essay scoring [1]. Many famous automated essay scoring systems, such as PEG, E-rater, adopted this algorithm [1]. Some literatures show that in automated essay scoring multiple linear regression performs better than some classical machine learning algorithms which perform very well in many applications, such as conventional support vector machine [6]. In our experiment, we regard the performance of multiple linear regression as the baseline.

When applying multiple linear regression to automated essay scoring, we take the extracted features (numeric value) of a scored essay in the training set as independent variables and take the score of the essay (numeric value) as dependent

Table II
TABLE 4.1 A BRIEF INTRODUCTION TO EIGHT ESSAY PROMPTS

Essay set No.	Student grade level	Training set size	Average length of essays	number of section scores	section score range	final score range
1	8	1785	350	2	1-6	2-12
2	10	1800	350	2	1-6,1-4	1-6,1-4
3	10	1726	150	2	0-3	0-3
4	10	1772	150	2	0-3	0-3
5	8	1805	150	2	0-4	0-4
6	10	1800	150	2	0-4	0-4
7	7	1730	250	8	0-3	0-24
8	10	918	650	12	1-6	10-60

Table III
TABLE 5.1 MEAN QUADRATIC WEIGHTED KAPPA ON TEST SET

algorithm	KNN	MLR	LambdaMart	SVMrank
test set	0.63035	0.72108	0.72031	0.74956

variable. Then we run multiple linear regression algorithm on the training set to learn a scoring function. Finally, we use the scoring function to calculate the predicted score of the essay in the test set.

C. Evaluation Metrics

After essay scoring, we measure the performance of automated essay scoring system by the quadratic weighted Kappa error metric, which is used to measure the agreement between two raters. This measuring method is adopted by the Automated Essay Scoring Competition. The quadratic weighted Kappa is calculated between the scores given by automated essay scoring system and the scores given by human raters. The mean of the quadratic weighted Kappa is then taken across all subsets of essays. This mean is calculated after applying the Fisher Transformation to the Kappa values. This metric typically varies from 0 (only random agreement between raters) to 1 (complete agreement between raters). In the event that there is less agreement between the raters than expected by chance, this metric may go below 0 [14].

D. Cross-validation

We take a ten-fold cross-validation on the training set to validate the prediction accuracy of these four algorithms in automated essay scoring. The training set, which has been introduced in this section, includes all the essays that contains ground truth label and accounts for 50% of the whole dataset.

V. THE EXPERIMENTAL RESULT

A. Performance On the whole essay set

Table 5.1 summarizes the performance of four kinds of algorithms on test set in our automated essay scoring experiment. Table 5.2, concludes their performance on training set in cross-validation. The performance is measured by the quadratic weighted Kappa error metric mentioned in the last section.

From these tables, we can see that the performance of KNN is far worse than other three algorithms. The reasons are two-fold. One reason is that KNN regards automated essay scoring as a classification problem. However, it is not appropriate to see score as undiscriminated class label in automated essay scoring. Another reason is that KNN select out essays similar to the unscored essay for final score calculation. But in the view of human raters, one essay is evaluated in many aspects and the final score is a comprehensive evaluation for these aspects. For example, there two essays, essay A and essay B. Essay A performs well in term using while essay B performs well in content. Although these two essays are different from each other, they may get the same score.

We can also find that SVMrank out performs LambdaMart although both of them are learning to rank algorithms. But SVMrank is a pairwise ranking algorithm while LambdaMart is a listwise ranking algorithm. LambdaMart aims at the optimization of common information evaluation measures, such as NDCG, MAP, which is not suitable for evaluating the performance of an automated essay scoring system [10]. What we concerned in automated essay scoring is the accordance between human rater and automated essay scoring system which is measured by mean quadratic weighted Kappa. What's more, essays with different titles are completely independent from each other. Thus, the advantage of LambdaMart compared with pairwise ranking algorithms disappeared as its cost function is defined in query-level, in this paper we can call it title-level. However, SVMrank aims at the optimization of partial-order relationship between each pair of documents [11]. In automated essay scoring, we can regard partial-order relationship between a pair of essays as the comparison of the two essays' written quality. Thus, the optimization goal of SVMrank is consistent with the goal of automated essay scoring system.

Finally, we can find from these figures that SVMrank out performs multiple linear regression. Even LambdaMart has almost equal performance with multiple linear regression. However, as we mentioned in section IV, multiple linear regression is the most commonly used algorithm in automated essay scoring system, such as famous E-rater, PEG. Thus, the performance of learning to rank algorithms in

Table IV
TABLE 5.2 TEN-FOLD CROSS-VALIDATION ON TRAINING SET

cross-validation	1	2	3	4	5	6	7	8	9	10	mean
KNN	0.59274	0.61431	0.60666	0.59759	0.60311	0.5 8751	0.62005	0.60856	0.63995	0.62484	0.60893
MLR	0.67775	0.71244	0.68551	0.69164	0.68136	0.69241	0.70044	0.70561	0.71376	0.70785	0.69688
LambdaMart	0.68049	0.70300	0.65186	0.67065	0.69 590	0.69295	0.69296	0.68674	0.71723	0.69154	0.6 8833
SVMrank	0.71537	0.74980	0.71963	0.72834	0.72507	0.72454	0.74066	0.74547	0.75066	0.74514	0.7360 6

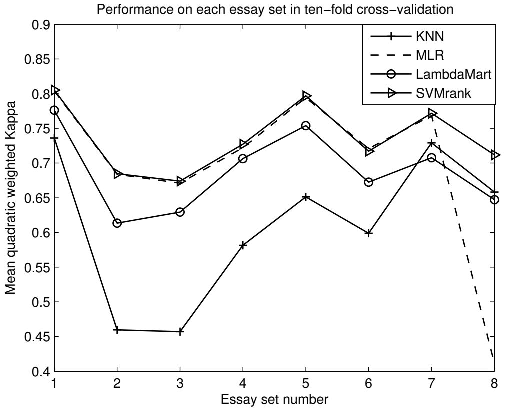

Figure 1. Performance on each essay set in ten-fold cross-validation

automated essay scoring worth affirmation. Further research on applying learning to rank algorithms to automated essay scoring, especially SVMrank, is worthy doing.

B. Performance On each essay set

Table 5.3 shows the performance on each essay set in ten-fold cross-validation. Figure 1 is the corresponding figure. There are eight essay sets and essays in the same set was generated from the a single prompt. We can see that the performance of SVMrank is better than other algorithms on all essay sets. Actually, the performance of multiple linear regression is almost the same with SVMrank on essay set 1 to 7. But its performance is terrible in essay set 8. The essay set 8 has 12 section scores and the final score is calculated based on these 12 section scores. After 12 times multiple linear regression, the accuracy came down rapidly. All algorithms performs well in essay set 1,5,7 and relatively bad in essay set 2,3,6. From Table 4.1 we can summarize that essays in set 2,3,6 are written by students whose grade level is 10 while essays in set 1,5,7 are written by students whose grade level is 8. Thus the reason lies in the selection of essay features. The selected essay features are appropriate for the rating of essays written by grade 8 students, but not appropriate for the rating of essays written by grade 10 students.

VI. CONCLUSIONS AND FUTURE WORK

From the experimental result, we can see that learning to rank performs well in automated essay scoring compared with the most commonly used algorithm multiple linear regression. As we mentioned in section 3, good features that can reflect the writing quality of essays well produce good prediction result. If the extracted features changed, such as more high quality features added into the feature set, or if the method of feature extraction improved, what the result will be remains to see. More good features need exploration and further experiments need to be conducted.

From the progress that has been achieved by E-rater and Intellimetric in practical use, we can also see that the accuracy of our system is not as good as these two existed commercial automated essay scoring systems. One reason is that the quantity of the extracted features is less and the quality of extracted features is also poorer. As the developers of these two commercial automated essay scoring systems have large-scale useful data produced by the large-scale examinations held every year, they can get more useful information and the scoring model can be more accurate. Another reason is that they adopted artificial intelligence techniques to the process of feature extraction and the building of final scoring model [1].

Nowadays, mature commercial automated essay scoring systems have been developed and even have been put into use in practical large-scale test with a high accuracy more than 97%. Thus the goal of accuracy of developing automated essay scoring has been solved. However, the two commercial automated essay scoring systems E-rater and Intellimetric just achieve a high accuracy in the English essays written by writers whose mother tongue is English. When it comes to the essays written by people whose mother tongue is not English, especially by people whose English level is poor, the accuracy of these two systems reduced greatly [4] [5]. Thus, automated essay scoring system for non-English native speakers has become the hot issue in this research area. We plan to go further study on how well learning to rank performs on essays written by non-English native speaker.

ACKNOWLEDGEMENTS

This work is supported in part by the National Natural Science Foundation of China (61103131/F020511), the President Fund of GUCAS (Y15101FY00/Y25102HN00), and the CAS R&E Project (110700EA12). The authors would like to thank all fellow students and colleages for their assistance and valuable advice on this research work, and more thanks here for the reviewers' constructive comments.

Table V
TABLE 5.3 PERFORMANCE ON EACH ESSAY SET IN TEN-FOLD CROSS-VALIDATION

Algorithm	1	2	3	4	5	6	7	8	mean
KNN	0.73612	0.45960	0.45703	0.58143	0.65121	0.59883	0.72911	0.65814	0.60893
MLR	0.80384	0.68386	0.67101	0.72208	0.79417	0.72074	0.76835	0.41097	0.69688
LambdaMart	0.77618	0.61342	0.62939	0.70642	0.75393	0.67243	0.70766	0.64723	0.68833
SVMrank	0.80522	0.68457	0.67398	0.72719	0.79697	0.71711	0.77175	0.71167	0.73606

REFERENCES

[1] Mark D. Shermis and Jill C. Burstein, *Automated Essay Scoring: A Cross-disciplinary Perspective*, 1st ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2002.

[2] Larkey and Leah S., *Automatic essay grading using text categorization techniques*. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 90-95, 1998.

[3] Semire D., *Automated Essay scoring*. Turkish Online Journal of Distance Education, Volume:7 Number:1 Article:5, 2006.

[4] Attali, Y. and Burstein, *Automat Essay Scoring With e-rater V.2.*. Journal of Technology, Learning, and Assessment, 4(3), 2006. Available from http://www.jtla.org

[5] Marti A. Hearst, *The debate on automated essay grading*. IEEE Intelligent systems (5):25, 2000.

[6] Lawrence M. Rudner and Tahung L., *Automated essay scoring using Bayes' Theorem*. The Journal of Technology, Learning, and Assessment, (2): 3-21, 2002.

[7] Hunter M. Breland, Robert J. Jones and Laura J., *The College Board Vocabulary Study*. College Entrance Examination Board, New York, 1994.

[8] Tie-Yan Liu, *Learning to Rank for Information Retrieval*. Foundations and Trends® in Information Retrieval Vol. 3, No.3(2009), 225-331.

[9] Susan T. Dumais, *Latent Semantic Analysis*. Annual Review of Information Science and Technology, Volume 38, Chapter 4, 189-230, 2004.

[10] Qiang Wu, Chris J.C. Burges, Krysta M. Svore and Jianfeng Gao *Ranking, Boosting, and Model Adaptation*. Microsoft Research Technical Report MSR-TR-2008-109.

[11] T.Joachims, *Training Linear SVMs in Linear Time*. Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), 2006.

[12] Qin, Tao and Zhang, Xu-Dong and Tsai, Ming-Feng and Wang, De-Sheng and Liu, Tie-Yan and Li, Hang, *Query-level loss functions for information retrieval*. Information Process and Management: an International Journal, v.44 n.2 p.838-855, March, 2008.

[13] Christopher D. Manning and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

[14] Ben Hamner, *Description - The Hewlett Foundation: Automated Essay Scoring - Kaggle*. http://www.kaggle.com/c/asap-aes, 2012.