AUTOMATED ESSAY SCORING LINGUISTIC FEATURE: COMPARATIVE STUDY

Soha M. EID Nayer M. WANAS
Electronics Research Institute – Cairo – Egypt
El-Tahrir St., Dokki, Giza, Egypt, sohaeid@eri.sci.eg, nwanas@eri.sci.eg

Abstract – *Automated Essay Scoring (AES) is the solution to a tedious and time consuming activity of manually scoring students' essays. AES is usually treated as a supervised machine learning problem where feature extraction plays an important role. In an attempt to investigate the importance of lexical features in AES systems, a new extended feature set is developed by combining popularly known features. The combined feature set contains 22 features that captures five different aspects of writing qualities. The importance of each feature in the combined feature set is tested by eliminating each feature separately. It was found that using the number of nouns in the essay slightly degrades the AES system performance. The significance of the combined feature set is compared against three state-of- the-art AES commercial systems and its performance was found comparable.*

Keywords*: Automated Essay Scoring; Linguistic Features; Linear Regression; Parallel Computing*

1. Introduction

Writing essays is crucial in the assessment of students' language skills. With periodic evaluation of the students' performance, manually scoring of essays becomes a stressful and time consuming process. Automated Essay Scoring (AES) has been suggested as a tool for evaluating and scoring of essays written in response to specific prompts. Automated essay scoring (AES) can be defined as the process of scoring written essays using computer programs. The process of automating the assessment process could be useful for both educators and learner, since it encourages the iterative refinement of students' writings.

Most of the existing AES systems start by analyzing written text to extract useful features that are related to intrinsic writing characteristics. Various features are weighted using a regression procedure to develop AES statistical model. AES systems consider different features according to the different scoring aspects they focus on. Basically the AES task has usually been regarded as a supervised machine learning problem. In this context, a set of essays, accompanied with their associated gold scores, constitute the training data is analyzed into its observable components such as average word length and spelling errors. These components represent the features of the machine learning model that can be generalized to unseen instances. The performance of the AES system depends on (i) the learning model and (ii) the extracted features. The learning model of AES system is usually built using either Regression [1-3] or classification [4-5]. The features extracted for AES can be based on content or style. Content-based features include words and phrases, which are usually sparse and prompt specific. [6]. On the other hand, style-based scoring systems depends on linguistic features, that may be either superficial features such as number of words or complex linguistic features such as spelling errors [1]. Alternatively, some suggested AES systems combine both kind of features to reflect more aspects of essays' qualities [7-8].

AES systems are developed in either commercial or research environment. A pioneer system has been suggested by Ellis Page in 1966 [9]. The Project Essay Grade (PEG) is a proprietary AES system that was developed at Measurement Inc. [1]. The system identifies a group of text features, proxies, that are considered approximations of the intrinsic variables of writing quality. Recently a hybrid feature methodology that incorporates variables derived statistically or extracted via NLP techniques [10].

Generally, there are three major commercial systems for Automated Essay Scoring: (i) Electronic Essay Rater (E-rater) developed by the Educational Testing Service (ETS) of America [7], (ii) IntelliMetric developed by Vantage Learning [11] and (iii) Intelligent Essay Assessor (IEA) developed by Pearson Knowledge Technology [12]. E-rater uses regression to model the combination of essay features to best correlate them with the ratings of the human experts [13]. There are 12 features used by E-rater v 2.0 to score essays [7]. These features include grammatical error proportions, word usage error proportions, number of word types divided by the number of word tokens, average length of words and total number of words. IntelliMetric considers 300 features that are extracted using Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies

978-1-5386-6407-0/17/$31.00 ©2017 IEEE

[11]. This feature set evaluates five different classes of writing quality, namely: (i) content, (ii) word variety, (iii) grammar, (iv) text complexity and (v) sentence variety. Intelligent Essay Assessor (IEA) is a completely different approach that is based on Latent Semantic Analysis (LSA) to analyze and score essays [14]. Latent Semantic Analysis is a machine-learning method that acquires and represents knowledge about meaning of words and documents by analyzing large bodies of natural text [15]. Although IEA provides scores for content and style, LSA-based procedures are most effective in assessing the content of essays [12].

Besides the above mentioned commercial AES, several approaches based on different feature sets have been suggested in the literature. Linguistic features sets were adopted to represent the quality of writing. McNamara et al. used Coh-Matrix tool to examine the degree of the proficiency of the essay [16]. They evaluated different aspects of writings such as the diversity of words used by the writer and other words characteristics (e.g. frequency and correctness). Liang clustered linguistic features according to the writing aspects they evaluate [17]. Language is approximated by number of word types, average word length and square derivation of word length. Singular Value Decomposition (SVD) similarity measure between documents estimates the writing content and paragraphing is an estimate of the organization of the text. Larkey proposed eleven heuristics dense features to represent the text complexity of the written essay [4] [18]. This feature set includes: number of characters in the document, number of words in the document, number of unique words, average word length and average sentence length. Muray and Orii used a smaller set of 7 features [19]. They included the number of spelling errors, difficult words and punctuation marks in their feature list. Mahana et al. used a larger dense features set that consisted of 14 features [20]. The features used estimate different characteristics of essay such as language fluency and dexterity. They regressed each feature individually in order to determine its importance in their model with discarding features that don't improve their model. The resultant feature set is not only carefully chosen, it is also ordered with number of word count and number of character count at the head of the list. The same procedure was also adopted by Ostling, et al for ranking features in Swedish essay scoring [21]. They produced a list of 46 features with the fourth root of the number of word tokens, number of word tokens and number of sentences at the head of their list.

In this paper, we aim to shed some light on the correlation between the linguistic features and students' grading performance in AES. This is conducted by implementing and comparing the performance of three different groups of features that were proposed in three different key research systems suggested in the literature. Also, the performance of combined feature set is investigated and compared with the performance of three AES commercial systems.

2. Linguistic Feature Comparative Study

Feature extraction plays a key role in any AES system. Similar to human raters, AES systems consider different aspects of written essays. Good features reflects the writing quality of essays and produces good prediction results. Linguistic dense features are used as an approximations for the intrinsic quality of the essays. In this work, 22 lexical features, that capture five essay aspects, are considered. These features have been suggested in three key non-commercial approaches [4] [19-20] and are popular within the research community. Table 1 summarizes the features that are considered in this study.

The Stanford POS-tagger was used to facilitate tokenization and sentence boundary [22]. Sentence boundary detection is essential in evaluating sentence count (SC) and average sentence length (ASL). Moreover, the various POS counts and punctuation marks' counts are obtained.

Richness of Essay Content

Essay length in character count and word count (WC) is a good estimation of language fluency and content richness. Word counts (WC) are estimated by adding 1 to the number of word separators (spaces). It is worth mentioning that a long essay is appreciated to a limit, beyond which it carries little additional weight. Moreover, the fourth root of essay length in words was found to be of a highly correlation with essay score [6]. Larkey included all three essay length estimations in his approach [4], while Murray and Orii excluded the fourth root of word count [19]. Mahana et al., considered only word count as their measurement of essay length [20].

Complexity of Term Usage

Essay's vocabulary richness is correlated with word's length. Mahana et al. used the average word length and long word counts to estimate richness of the essay [20]. A word is assumed a long word if its length is larger than average word length, Larkey substituted the number of long word with four word counts that are greater than 5, 6,7 and 8 characters [4]. On the other hand, Murray and Orii estimated the vocabulary richness by the number of difficult words (DWC). The difficult word count (DWC) is obtained by counting the number of words that exist in a list of 5000 words that frequently appear in SAT [19].

Word length is the number of character in each word and average word length of each essay (AWCessay) is estimated as the character count (CC) per word as shown in Equation (1).

$$AWC_{essay} = \frac{CC}{WC} \quad (1)$$

Table 1: Linguistic features used by the different approaches

Features' Groups	Feature ID	Features	Larkey	Murray and Orii	Mahana et.al.
Richness of essay content	1	Character Count	✓	✓	
	2	Word Count	✓	✓	✓
	3	Fourth Root of Word Count	✓		
Complexity of term usage	4	Average Word length	✓		✓
	5	Words Count > 5 Char	✓		
	6	Words Count > 6 Char	✓		
	7	Words Count > 7 Char	✓		
	8	Words Count > 8 Char	✓		
	9	Difficult word Count		✓	
	10	Long Word Count			✓
Orthography	11	Spelling Error		✓	✓
Text complexity	12	Unique Word Count	✓		
	13	Noun Count			✓
	14	Verb Count			✓
	15	Adjective Count			✓
	16	Adverb Count			✓
	17	Stop Words Count		✓	
Essay Organization	18	Sentence Count	✓		✓
	19	Average Sentence Length	✓		
	20	Exclamation Mark Count		✓	
	21	Question Mark Count		✓	
	22	Comma Count			✓

The average word length over the whole training data (AWC_{tot}) is given by Equation (2).

$$AWC_{tot} = \frac{\sum_{essays} CC}{\sum_{essys} WC} \quad (2)$$

Long word Count (LWC) is the number of words whose length is greater than average word count of the training data (AWCtot). It is worth mentioning that the word count is performed without stop-word removal.

Orthography

Command over language can be estimated with correct word spelling. Mahana et al. included the number of spelling errors in his approach they used the PyEnchant 1.6.5 spell checker with the aspell dictionary [23]. The spelling error count (SEC) is estimated by the number of the distinct misspelled words per essay.

Text Complexity

Text that avoids word repetition is found to be difficult and complex. Number of unique words is an estimate of the words repetition. Although Mahana et al. proposed lexical diversity feature, there was no further explanation of its measure [20]. Moreover, they included the various POS counts in their feature set that may indicate the diversity of words' usage. Murray and Orii included the number of stop words [19]. The stop word count (SWC) is estimated by counting the number of words that belong to a stop-word list. The stop-word list was obtained from the natural Language Tool Kit which contains 127 words [24].

Organization

Punctuation is a good indication of a well-structured and organized essay. Murray and Orii included the number of exclamation and question marks in their features set [19]. Larkey's feature set included sentence count and average sentence length that can be an estimation of essay structure [4]. Mahana et al. included comma counts and sentence counts in their feature set [20].

2.1 Experimental Data

The data used in this study is the publically available Automated Student Assessment Prize (ASAP) dataset [25]. It contains around 13,000 essays written by students in grades 7, 8, and 10. This dataset contains 8 prompts of different genres. The average length of the essay differs for each prompt ranging from 150 to 650. Each prompt represents a different context to ensure variability of the domain. The essay sets can mainly be either a Source Dependent Responses or Persuasive/ Narrative responses. Source Dependent Responses are prompts based upon a passage that the students first has to read. On the other hand, Persuasive/Narrative responses ask students for stories or formal arguments to persuade the reader in agreeing with the student's opinion on a particular topic. Essays were manually graded by at least two different

Table 2: Detailed Description of the dataset

Prompt Id	# essays	Grade level	Essay Type	Score Range	Average length
1	1783	8	Persuasive/Narrative	2-12	350
2	1800	10	Persuasive/Narrative	1-6, 1-4	350
3	1726	10	Source Dependent	0-3	150
4	1772	10	Source Dependent	0-4	150
5	1805	8	Source Dependent	0-4	150
6	1800	10	Source Dependent	0-4	150
7	1569	7	Persuasive/Narrative	0-30	250
8	723	10	Persuasive/Narrative	0-60	650

graders with different scales for each set. Essay Set 2 has two different scales for two different domains of writing: applications and language conventions. Table 2 gives an overview of the datasets used in this study.

2.2 Experimental Work

The comparative study consists of multiple phases. A scoring model based on word count and character count is implemented and considered as a baseline model. Then, scoring models based on lexical feature sets under study are implemented separately. We proposed an aggregated model that implement a combined feature set. The extended feature set combined all the features that was suggested by the non-commercial approaches under study without repetition. This feature set contains the twenty-two features listed in Table 1. The combined feature set enables more features as an estimate for the five aspects of writing qualities. The performance of the extended list is compared against the approaches that previously suggested them. The importance of each feature is evaluated by eliminating each feature individually. Obviously, as the importance of the feature increases the evaluation metric decreases when such feature is eliminated. As a final step, the importance of the final feature set is examined by comparing its performance against the performance of three commercial AES systems.

2.3 Learning Model

An overview of related prior work indicates that linear regression works well for essay grading applications [1-3]. Moreover, the range of scores is very large in two prompts and the classification will not work well in these cases. Hence, we chose linear regression as our learning model. The linear regression is used to predict a y based on features x extracted from a given essay. For a given feature vector $x \in \mathbb{R}^m$, an output $\hat{y} \in \mathbb{R}$ is predicted using a linear model with weight of β

$$\hat{y} = \beta_\circ + x^T\beta \quad (3)$$

To learn the values of (β0, β), the model minimizes the sum of the squared errors for a set of training set containing an n pairs of essays and scores (xi, yi), where xi $\in \mathbb{R}$m and yi $\in \mathbb{R}$ for $1 \leq i \leq n$. The grades are produced by rounding the prediction ŷ to the nearest integer and thresholding at the minimum and maximum grades

2.4 Evaluation

The performance of the learning model was evaluated using the Quadratic Weighted Kappa (QWK) [25]. Quadratic Weighted Kappa is a robust error metric that takes into account the possibility of agreement occurring by chance as the baseline. This metric typically varies from 0 (only random agreement) to 1 (complete agreement). In the event that there is limited agreement between the raters than expected by chance, this metric may go below 0. The QWK is calculated between the automated scores for the essays and the resolved score for human raters on each set of essays. For N possible essay ratings, an $N\times N$ matrix O is constructed where $O_{i,j}$ represents the number of essays receiving grade I from the first grader and j from the second grader. Moreover, the matrix E is constructed the same way, but assuming that there is no correlation where E_{ij} is given by Equation 4.

$$E_{i,j} = \frac{\sum_i O_{i,j} \sum_j O_{i,j}}{\sum_{i,j} O_{i,j}} \quad (4)$$

An $N\times N$ matrix w is also calculated where

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \quad (5)$$

The Quadratic weighted kappa is calculated by:

$$k = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \quad (6)$$

The mean of the QWK is then taken across all sets of essays. This mean is calculated after applying the Fisher

Table 3: Results of the average 5-folds on the essay sets

Prompt id	Baseline	Lackey	Murray & Orii	Mahana et.al.	Combined approach
1	0.64642	0.71782	0.66108	0.68544	**0.7256**
2(WA)	0.551892	0.551892	0.61408	0.57586	**0.65556**
2(LC)	0.36962	0.36962	0.54808	0.52578	**0.6305**
3	**0.64394**	0.64336	0.63876	0.64246	0.6438
4	0.64284	0.66474	0.64828	0.64342	**0.67708**
5	0.7581	0.77878	0.76778	0.76756	**0.78118**
6	0.59666	0.63538	0.61562	0.64258	**0.66656**
7	0.57014	0.57014	0.596	0.62356	**0.66538**
8	0.52746	0.52746	0.5965	0.57084	**0.6509**
Average k	0.613	0.662	0.642	0.645	**0.68409**

Transformation [25] to the kappa values. The Fisher Transformation is approximately a variance-stabilizing transformation and is defined by:

$$z = \frac{1}{2}\ln\frac{1+k}{1-k} \quad (7)$$

The mean of the transformed kappa values is calculated in z-space. For essay set 2, which has scores in two different domains, each transformed kappa is weighted by 0.5. Finally, the reverse transformation is applied to get the average kappa value:

$$k = \frac{e^{2z}-1}{e^{2z}+1} \quad (8)$$

3. Results

The results of the average five-folds Quadratic Weighted Kappa on each essay prompt is presented in Table 3. The results presented for the baseline, Larkey, Murray & Orii, Mahana et. al., and our proposed combined approach. The overall resultant QWK measure is presented as average k. The table shows that all the considered approaches performs better than the baseline. The dominance of the word and character counts on the performance of automated AES approaches is obvious. Larkey feature set performs better than both Murray & Orri and Mahana et. al. approaches. Meanwhile, the aggregation of features in the combined approach results in an enhanced performance due to using more increasing the number of features that estimate each writing aspect.

The significant of each feature is experimented by eliminating each feature separately. Fig (1) shows the performance evaluation of the AES system by eliminating the features in a descending order of their significance. The detailed description of the feature ID is given in Table 1. The QWK evaluation shows that the fourth root of the number of words has a remarkable influence on the performance of the AES system. Its influence surpass other text length features. On the other hand, it is obvious that the number of nouns has a harmful effect on the AES performance since removing it slightly enhances the QWK metric.

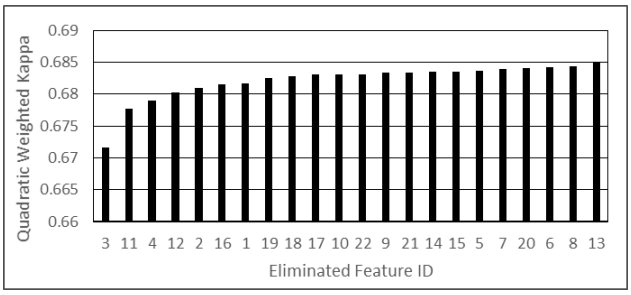

Fig (1): The influence of eliminating features separately

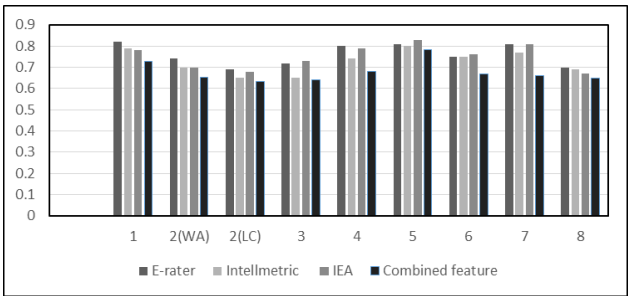

Fig (2): performance comparison between combined approach and commercial AES systems

To demonstrate the significance of the produced lexical feature set, its performance evaluation is compared to the commercial AESs. Fig (2) shows the quadratic weighted kappa results for the combined approach and three commercial AES systems namely, e-rater, Intellimetric and IEA. The results are presented over the eight essay

prompts. It is obvious that the combined approach performs comparably to the commercial vendors.

4. Conclusion

In this work, we investigate the importance of the linguistic features in Automated Essay Scoring Systems. A 22 lexical features gathered from three non-commercial approaches. These approaches are compared against the proposed combined feature set. The proposed feature set outperforms the three approaches with QWK of 0.68409 against QWK of 0.662 which is the best of the three approaches. This leads to believe that incorporating more features measuring the same quantity boost the performance of the AES system. This study also investigate the importance of each feature by eliminating each feature separately. It shows that noun count, number of words more than 6 and 8 degrade the performance of the scoring system. Accordingly, the feature set could be shortened by removing these features without degrading the system performance. Moreover, the combined feature set performs comparably with three state-of- the-art AES commercial systems that imply its high significance.

References

[1] Page, E.: Computer Grading of Student Prose Using Modern Concepts and Software, In: Journal of Experimental Education (1994), 62(2), p. 127–142

[2] Persing, I., Ng, V.: Modeling Prompt Adherence in Student Essays. In: Proceedings of the 53rd Annual meeting of the Association of the Computational Linguistics and the 7th International Joint Conference on Natural Language Processing 2014, Vol 1: Long Papers, p. 543-552

[3] Phandi, P., Chai, K. and Ng, H.: Flexible Domain Adaptation for Automated Essay Scoring using Correlated Linear Regression. In: Proceedings of the Conference on Empirical Methods in Natural Language processing 2015, p. 431-439

[4] Larkey, L.: Automatic Essay Grading Using Text Categorization Techniques. In: Proceedings of 21st International conference of the Association for Machinery Computing-Special Interest Group on Information Retrieval ACM-SIGIR 1998., Melbourne, Australia

[5] Runder L. and Liang, T.: Automated Essay Scoring using Bayes' Theorem. In: The Journal of Technology, Learning and Assessment 1(2), 2002

[6] Shermis, M. and Burstein, J.: Automated essay Scoring: Cross-disciplinary Perspective. Computational linguistics. 30(2), 2004. p. 245-246

[7] Attali, Y. and Burstein, J.: Automated Essay Scoring with e-rater V. 2. In: The Journal of Technology, Learning and Assessment. 4(3), 2006, p. 3–29.

[8] Li, X. and Liu, J.: Automated essay Scoring Based on Coh-Matrix Feature Selection for Chinese English Learners. LNCS 10108, 2017, p. 382-393

[9] Page, E.: The Immense of Grading Essays by Computer. In: Phi Delta Kappan, 48. 1966, p. 238-243

[10] Page, E.: Project Essay Grade: PEG. In: Automated Essay Scoring: A Cross- disciplinary Perspective. 2003, p. 43-54

[11] Elliot, S: IntelliMetric: From Here to Validity. In: Automated Essay Scoring: A Cross- disciplinary Perspective. 2003, pp. 71-86

[12] Miller, T. Essay Assessment with Latent Semantic analysis. In: Journal of Educational Computing Research, 28(3), 2003.

[13] Yang, Y., Buckendahl, C., Juszkeiwicz, P. and Bhola, D.: A Review of Strategies for Validating Computer Automated Scoring. In: Applied Measurement in Education, 154, 2002, p. 391- 412

[14] Landauer, T., Laham, D. and Foltz, P.: The Intelligent Essay Assessor. IEEE Intelligent systems.15 (5). 2000, p. 27–31

[15] Landauer, T., Foltz, P. and Laham, D.: Introduction to Latent Semantic Analysis. In: Discourse Processes. 25. 1998, p.259-284

[16] McNamara, D., Crossley, S. and McCarthy, P.: Linguistic Features of Writing Quality, In:Written Communication, 27(1), 2010, p. 57-86

[17] Liang, M.: Constructing a Model for Automated Scoring of Chinese EFL Learners' Argumentative Essays. PhD dissertation, Nanjing University, 2005

[18] Preston, D., Goodman, D.: Automated Essay Scoring and the Repair of Electronics. Technical Report, 2012, http://snap.stanford.edu/class/cs341-2012/reports/03-preston_cs34_-_Dan_and_Danny_-_Final.pdf

[19] Murray, K. W. and Orii, N.: Automatic Essay Scoring. 2013, http://www.cs.cmu.edu/~norii/pub/aes.pdf

[20] Mahana, M., Johns, M. and Apte, A.: Automated Essay Grading Using Machine Learning. In: Mach Learn Session, Stanford University, 2012

[21] Ostling,R., Smolentzov, A., Hinnrich, B. and Hoglin, E.: Automated Essay Scoring for Swedish. In: Proceedings of the 18th Workshop on Innovative use of NLP in building Educational Applications, 2013.

[22] Toutanova, K., Klein, D., Manning C., and Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proc. of HLT-NAACL 2003, p. 252-259

[23] Kelly, R.""PyEnchant. http://packages.python.org/pyenchant

[24] NLTK, www.nltk.org

[25] Hewlett Prize. 2012. http://www.hewlett.org/newsroom/press-release/hewlett-foundation-sponsors-prize-improve-automated-scoring-student-essays