OMSCS 6460EduTech-Summer2019

Proposal

Automated Essay Scoring Investigation

Chenming Xue  
cxue34@gatech.edu

# 1. Introduction

The essay assessments are extremely time consuming when done manually. On the other hand, the judgement and scoring/grading are based on an individual’s subjective opinion and is complained to be unfair sometimes. This calls for the automated grading by computers that it will not only reduce the time for assessment but also remove human bias. The AES can quickly give students and teachers feedback of the essay and corresponding changes can be applied as a quick action. The benefits are time saving, cost/labor saving, reliability, fairness and generalizability in writing assessment. (Bereiter , 2003)

However, the text can be very complicated in the forms or resources of verbal interaction, written sample, or reading. With the previous research in this domain, implementing machine learning techniques with various features spaces, and large collection of data having different patterns offers human being a way to get reasonable outcomes. Currently, the AES systems are in the early stage and have been criticized for such problems which need to improve, e.g. lack of human interaction, vulnerability to cheating, requiring a large set of sample text to train the system. With the continuous research efforts and improvement of computer technology, the system is getting more advanced. (Ramalingam V., 2018)

There is a comprehensive discussion about this topic (Attali Y. 2013) and we can see this topic is important and there is big value, yet there is accompanying skepticism and criticism over the years. It is believed that AES should be constructed primarily as a complement to human scoring, limited in its ability to measure a subset of the writing construct. There are complicated factors and features: meaning of features, interpretation of features as measures of writing skills, quality of measurement, and validity of feature aggregations into scores for the validity process. Also it should not be judged exclusively against the criterion of human ratings. Therefore for my project a development plan is set based on the above results. The goals have two parts. For one thing, the information of essays will be collected and analyzed. This is datamining of essays which can provide students and teachers an overview of the essay and we can see how much valuable information the computers can provide us. Second, as mentioned above, the AES system is in early stage which has limit in understanding the meaning of the essay, there is much room to improve. With newly developed system in this project, the results will be compared to the existing system such as e-rater, and human scores.

**2. Related work**

Several methods have been implemented for AES technologies, and they have advantages and disadvantages. For the foundation, it is Natural language process (NLP) technology. Also, the performance of some of these methods is displayed, such as scale-up.

NLP can be considered as group of algorithms and methods that are created to analyze natural language and to give informative answers about this large dataset. (Noah Weber, 2018) The target of this research area is to find a way to parse natural language and use machine learning algorithms on train data so computers can learn and test them on test data. To solve this problem, there are techniques developed: statistical methods/algorithms including N-gram (Jurafsky D., 2014), Alternative with N-grams, Bag of words(BoW) (Mikolov T., 2013), Word2Vec (Mikolov T, online tool), and different neural network architectures for NLP, SVM, RF and NN (especially rNN). The effective connectedness features between sentences can be learned and the local coherence model with a state-of-the-art AES model is evaluated. (Farag Y. 2018)

Project Essay Grader (PEG) was firstly developed by Ellis Page in 1966 and planned to make the large-scale essay scoring process more practical and effective (Rudner, 2001). This method uses correlation to predict the intrinsic quality of the essays. This method has pros in predicting scores comparable to human raters and it can track wring errors. But this method is weak in evaluation in semantic aspect of essays and more focusing on the surface structures and failing to detect the content related features. (Kukich, 2000) Also, this method cannot provide instructional feedback to students.

Intelligent Essay Assessor (IEA)

analyzes and scores an essay using a semantic text-analysis method called Latent Semantic Analysis (LSA) (Lemaire, 2001). This LSA method is defined as “a statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information”. The idea can be summarized as:

“meaning of word1 + meaning of word2 + … + meaning of wordk = meaning of passage”

The advantage of IEA is this system needs smaller numbers of pre-scored essays to train, like 100 pre-scored training essays.

Electronic essay rater (e-rater) has been developed by ETS to evaluate the quality of an essay. This method identifies linguistic features in the text using natural-language processing (NLP) techniques, which identify specific lexical and syntactical cues. (Burstein, 2003) NLP is one of the most challenging areas of AI, which comprises a variety of fields. The empirical methods used in NLP employ statistical and machine learning techniques to train the system on large amounts of authentic language data. NLP contains several levels of processing and subtasks: speech recognition, syntactic analysis, discourse analysis, information extraction, and machine translation. Briefly, the sequence is the transformation of training essays into vectors of word frequencies, then into each word frequency and finally into word weight: training essaysword frequencies vectorsweight vectors. A scalable study has been executed investigated, this method has good potential and will provide better service in the future.(Andre R. 2019)

IntelliMetric is an AES system is the first essay-scoring tool that based on AI and relying on NLP. (Elliott, 2003) Similar to e-rater but having different architecture, IntelliMetric contains five key principles: modeled on human brain, learning engine, based on a complex system of information processing, inductive, multi-dimensional and non-linear. This is different from many other scoring systems based on General Linear Model. MyAccess! Is a web-based writing assessment tool based on IntelliMetric AES.

Bayesian Essay Test Scoring System (BETSY) BETSY is another AES but should be treated more as a research tool. (Rudner & Liang, 2002) It has several applications such as identifying spam and other unwanted e-mails. There are two Bayesian models widely used in text classification: Multivariate Bernoulli Model and Multinominal Model. This approach includes key concepts such as stemming, stop words, and feature selection.

Scale-up performance

While AES has been implemented in ETS for practical use, we only know the effectiveness on individual cases. In this research report, the design and empirical findings for a large-scale study of essay writing ability with approximately 2,500 high school students in Germany and Switzerland on the basis of 2 tasks with 2 associated prompts have been described, and for standardized writing assessment the scoring from both human and automated components were compared. (e.g., Rupp, 2018; see also Bejar, 2011; Bennett, Zhang, 2016;Williamson et al., 2012).

**3. Proposal**

**3.1 Design**

In my design, a AES system will be built. I will focus on a system that can analyze essay and provide constructive data for users. The scoring function will be built as well but may not be the most important target. During the exploration and testing, improvement could be achieved potentially and can contribute to this AES research topic.

For details, in first step information and feature extraction will be executed for essay analysis (NLP, data mining), then the modeling methods will be implemented (the idea is using NLP, e-rater, and others) for essays evaluation, and the last step will be scoring and comparison to human results or other AES systems. This is the version design plan, and a more comprehensive version will be improved based on further exploration and discussion. The next target is to improve the features used for scoring and devising new features that tap more qualities of writing performance.

In my project, the evaluation system will be working on English language only. The plan is to develop an AES system for classifying a corpus of textual entities into small number of discrete categories, corresponding to possible grades. Learning from literatures, linear regression technique will be utilized for training the model along with making the use of various other classifications and clustering techniques. For implementation, first of all, train classifiers on the training set will be executed to make it go through the downloaded dataset, and then measure performance of new dataset by comparing the obtained values with the dataset values. Here I will try to build the application with Python and Neural networks (deep learning) built on Pytorch or TensorFlow platform will be applied.

The AES system has two parts:

Data mining and analysis of essay based on NLP. This provides a statistical review of the essay for overview

Scoring and grading based on machine learning

The system architecture is:

Feature extraction. The key components are language fluency, grammatical and syntactic correctness, vocabulary and types of words used, essay length, domain information etc. The “text mining” library will be using the created library based on training corpuses documents.

Part-of-speech (POS) Tag. This is critical task to evaluate the quality of content in an essay. NLTK library will be used to for POS tag for each word in an essay so we will cover all most all English words.

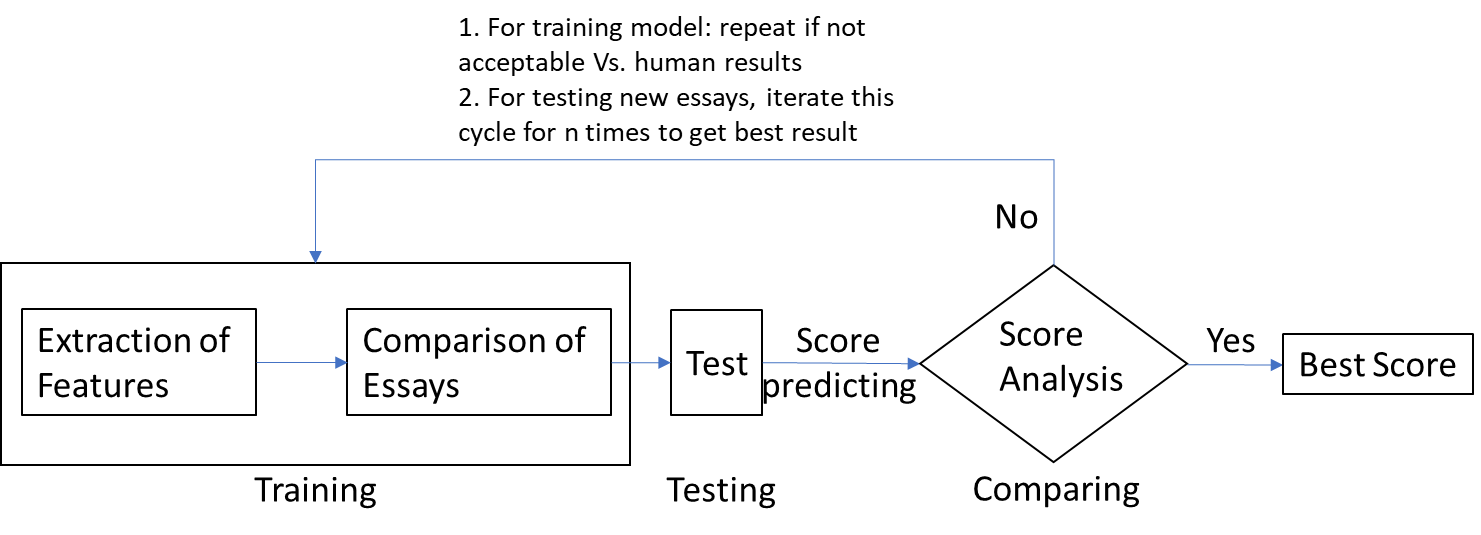
Spelling mistakes. This is a part of grading and will be included in the model. Spell checker ‘enchant’ will be implemented.

Domain Information Content. It is a marked feature in the model which makes the machine understand the content of the essay. First the best essays from each set will be picked and the nouns (keywords) will be pulled. These key words will be input into ‘WordNet’ and other equivalent words will be taken out and a pool will be formed, and most relevant words will be generated. The domain words in the essay will be counted.

Research features. Grammar, usage, mechanics, style microfeatures etc. Examples can be learned from literature. (Chen J. 2017)

**3.2 Evaluation**

For evaluation process: 1. The scoring model is made by analyzing as many essays as possible on a specific topic based on AES system and information of pre-scored by human raters is collected. 2. Then new scores are predicted using the scores already available and further analyzed and compared to human scores. 3. Eventually, scores extracted from all the features are combined which finally gives us a machine-generated score. 4. This is an option but interesting part that partial results will be compared with human results.



3.3 Resources

Harware & software: Mac, PC with GPU, Python, Jupyter Notebook, TensorFlow, Pytorch, and Linux (Ubuntu 18.04).

Database: <https://www.kaggle.com/c/asap-aes/data>

Open source resources: <https://github.com/chenmingxue/automated-essay-grading>

Commercial available services to refer to or compare: PEG (trial provided), e-rater (Attali Y. 2006); Free online essay grader (<https://analyze.academichelp.net/free-grade-my-paper.html>)

**4 Deliverables**

In the final project deliverable, I will provide code, data, report paper in PDF and video for the presentation. The deliverables contains:

Milestones 1: use NLP to analyze an essay, provide statistical data based on datamining from the essay.

Milestones 2: make the analysis of an essay more comprehensive, add information such as POS tag, spelling mistakes, domain information content and research features. In this milestone, the NLP system should be ready. May add scoring function depending on the analysis results. (If not ready, this part will be in final)

Final: Add scoring/grading function to the system. Comparing the results to the human rating and online open sources and free platform. Make improvements based on the results. Final report.

Task list:

Total: 120.5 hours

# 5 REFERENCES

1. André A. Rupp, Jodi M. Casabianca, Maleika Krüger, Stefan Keller, Olaf Köller, 2019, Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany, TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516
2. Attali, Y. (2004, April). Exploring the feedback and revision features of Criterion. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.
3. Attali, Y., Burstein J. (2006). Automated Essay Scoring With e-rater® V.2, The Journal of Technology, Learning, and Assessment.
4. Bereiter, C. (2003). Foreword. In Mark D. Shermis and Jill C. Burstein (Eds.), Automated essay scoring: a cross disciplinary approach (pp. vii–ix). Mahwah, NJ: Lawrence Erlbaum Associates.
5. Burstein, J. (2003). The e-rater scoring engine: Automated Essay Scoring with natural language processing. In M. D. Shermis and J. C. Burstein (Eds.), Automated Essay Scoring: A cross disciplinary approach (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.
6. Chen J., Zhang M., Bejar I., (2017) An Investigation of the e-rater® Automated Scoring Engine’s Grammar, Usage, Mechanics, and Style Microfeatures and Their Aggregation Model, ETS Research Report Series ISSN 2330-8516
7. Elliot, S. (2003). IntelliMetric: from here to validity. In Mark D. Shermis and Jill C. Burstein (Eds.). Automated essay scoring: a cross disciplinary approach. Mahwah, NJ: Lawrence Erlbaum Associates.
8. Huy H., Diane L., (2015). Argument Mining for Improving the Automated Scoring of Persuasive Essays. The Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
9. Kukich, K. (2000, September/October). Beyond Automated Essay Scoring. In M. A. Hearst (Ed.), The debate on automated essay grading. IEEE Intelligent systems, 27–31. Retrieved November 12, 2004,
10. Lemaire, B. & Dessus, P. (2001). A system to assess the semantic content of student essays. Educational Computing Research, 24(3), 305–306.
11. Ramalingam, A., Prateek C., Himanshu N. (2018), Automated Essay Grading using Machine Learning, Algorithm, Journal of Physics: Conf. Series 1000, 012030
12. Rudner, L. M. & Liang, T. (2002). Automated essay scoring using Bayes’ theorem. The Journal of Technology, Learning, and Assessment, 1(2). Retrieved September 18, 2005, from http://www.jtla.org
13. Rudner, L. & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer (ERIC Digest number ED 458 290).
14. Shermis, M. & Barrera, F. (2002). Exit assessments: Evaluating writing ability through Automated Essay Scoring (ERIC document reproduction service no ED 464 950).
15. Youmna F., Helen Y., Ted B., Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input., 2018.