Automated Essay Rater using Natural Language Processing

Shristi Drolia
Department of Computer Science and Engineering
PESIT Bangalore South Campus

Shrey Rupani
Department of Computer Science and Engineering
PESIT Bangalore South Campus

Pooja Agarwal
Department of Computer Science and Engineering
PESIT Bangalore South Campus

Abheejeet Singh
Department of Computer Science and Engineering
PESIT Bangalore South Campus

ABSTRACT

This paper proposes a regression based approach for automatically scoring essays that are written in English. We have used standard Natural Language Processing (NLP) techniques for extracting the features from the essays. We extensively evaluate our approach on a benchmark database and demonstrate that the result obtained is comparable to human evaluators while at a much faster rate. We also analyze how the essays are scored to get a better understanding about the proposed approach.

Keywords

ETS, NLP, BOW, NLTK

1. INTRODUCTION

This paper describes the development and evaluation of a prototype framework designed for the purpose of automatically scoring essay responses. The paper reports an evaluation result on 6 data sets obtained from kaggle.com.

Essays are crucial testing tools for assessing academic achievement, integration of ideas and ability to recall, but are expensive and time consuming to grade manually. Manual grading of essays takes up a significant amount of instructor’s valuable time, and hence is an expensive process. Automated grading, if proven to match or exceed the reliability of human graders, will significantly reduce costs.

The model built for AES can either be prompt specific or generic. In a generic model, the features extracted for all the essays which are to be assessed. The models that extract features based on NLP techniques alone are usually generic models. In a prompt-specific model, the features are dependent on prompt.

We consider the Automatic Essay Scoring as a regression based problem and not as a classification problem, as a marginal misgrading will not result in total misclassification. We utilize several NLP techniques to extract features from the essay and use regression based machine learning methods to allot a score to the particular essay.

2. RELATED WORK

There are many systems available which are either commercial or as a result of research in this field. Some of the systems will be discussed in this section.

Educational Testing Service (ETS)

This system uses lexical semantic techniques to build a scoring system, based on small data sets. It uses a domain specific, concept based lexicon and a concept grammar, both built from training data[2,1]. The training data essays are parsed by Microsoft Natural Language Processing (MsNLP) tool, any suffixes are removed by hand, and a list of stop words is also removed. This produces a lexicon. The list of words and terms in the lexicon remain constant whilst the features associated with each entry are modular, so can be replaced as necessary[1]. Some manual classification is necessary, such as specification of some words as metonyms of each other and so on. Grammar rules are then constructed, again manually, for each category of answer (each category should contain all the paraphrases for that possible answer) using syntactic parses of sentences from the training data along with the lexicon.

Electronic Essay Rater (E- Rater)

E-Rater uses a combination of statistical and NLP techniques to extract linguistic features from the essays to be graded. Essays are evaluated against a benchmark set of human graded essays[2]. With E-Rater, an essay that stays on the topic of the question, has a strong, coherent and well-organized argument structure, and displays a variety of word use and syntactic structure will receive a score at the higher end of a six-point scale. E-Rater features include the analysis of the discourse structure, of the syntactic structure and of the vocabulary usage (domain analysis)[3]. E-Rater adopts a corpus-based approach to model building by using actual essay data to analyze the features of a sample of essay responses[5]. The application is designed to identify features in the text that reflect writing qualities specified in human reader scoring criteria and is currently composed by five main independent modules[4]. Three of the modules identify features that may be used as scoring guide criteria for the syntactic variety, the organization of ideas and the vocabulary usage of an essay[6]. A fourth independent module is used to select and weigh predictive features for essay scoring. Finally, the last module is used to compute the final score.

3. PROPOSED METHODOLOGY

The workflow for our proposed approach is as follows: First we extract the features from each essay. We choose each feature such that the human graders may look while grading an essay. Then, Linear Regression was then used as our learning model. Scores were predicted for different essay sets. These scores were compared against human graded scores to get an error rate metric.

3.1 Features

We describe the features used in our approach in this section.

Bag of Words (BOW):

We used this feature as a basis for extracting words that are good predictors of essay score. We have extracted 100 words for each data set. This represents the most commonly used words in these type of essays.

For each essay set, the top words were extracted after removing the stop words like the, of, for. The stop words were removed using NLTK. The frequency of each word was measured for every essay.

Parts of Speech (POS) count:

Various parts of speeches such as Nouns, Pronouns, Adjectives are good proxies for the vocabulary of the writer. We have extracted the frequency count of various parts of speech in the essay. This feature is extracted using NLTK part of speech tagger. The essay is first tokenized into sentences before the tagging process.

Statistical Features:

We use a number of simple features such as word count, sentence count, average sentence length, paragraph count[3,4]. These features represent the fluency and dexterity of the writer. For this the essay is first tokenized and split using python utilities. Then individual token were then used to compute the word count, sentence count, average sentence length and paragraph count.

Orthography:

Correct word spelling indicates the command over the language and its correct usage. We have extracted the number of incorrect spellings from each essay to test these characteristics. We have used python's pyEnchant spellchecker to obtain the count of misspelt words in the essay.

$$\begin{array}{ccccccccc} & X & & & U & & \Sigma & & V^T \\ & (\mathbf{d}_j) & & & & & & & (\hat{\mathbf{d}}_j) \\ & \downarrow & & & & & & & \downarrow \\ (\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} & = & (\hat{\mathbf{t}}_i^T) \rightarrow & \begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} \dots \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \end{bmatrix} & \cdot & \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{bmatrix} & \cdot & \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix} \end{array}$$

Similarity:

Some datasets have a source essay based on which a question has been asked. As an essay cannot have one correct answer we have compared the similarity between the source essay and the answered essay. We have used Latent Semantic Algorithm for the comparison of similarity.

LSA is a technique in natural language processing of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text.

A term Document matrix which describes the occurrences of terms in documents is constructed. It is a sparse matrix where the rows represent the terms in the documents and the columns correspond to the documents.

Now in the term Document matrix we apply a mathematical technique called singular value Decomposition (SVD) to reduce the number of rows while preserving the similarity structure among the columns.

From Fig. 3.1 we see that X is decomposed which is represented by the equation:

$$X = U\Sigma V^T \qquad (3.1)$$

Where X is the term Document matrix. U and V^T are orthogonal matrices and Σ is a diagonal matrix.

We can compare the similarity between two documents $j \; and \; q$ by comparing the vectors $\Sigma_k \hat{d}_j$ and $\Sigma_k \hat{d}_q$ which is done using cosine similarity

3.2 Linear Regression

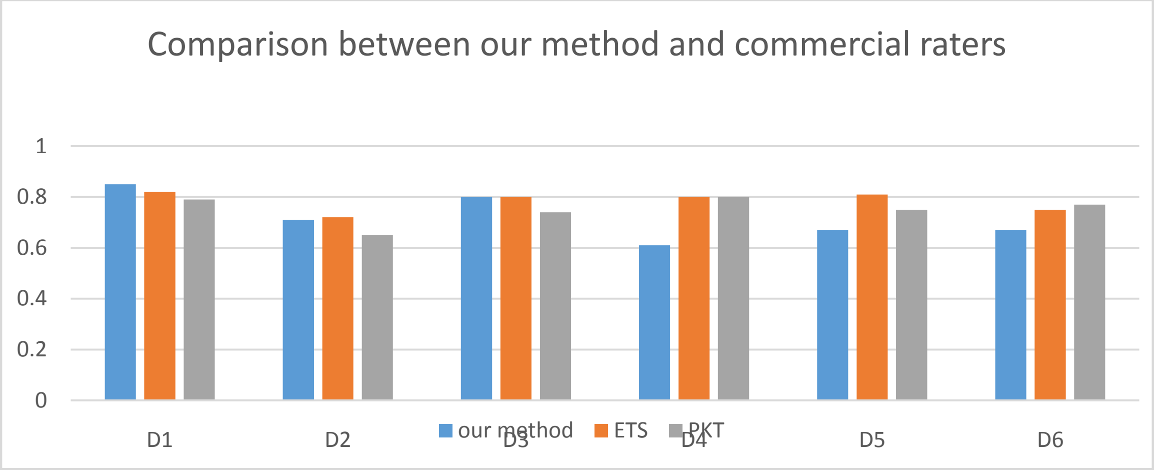
Linear Regression is an approach for modelling a relationship between a scalar dependent variable Y and one or more explanatory variables (or independent variables) denoted by X.

We have taken the features extracted from the essays as the X and the corresponding score as the Y for training the essay. The

After the model has been trained for finding the score of an unrated essay we extract the features from and give them as the X axis and the corresponding Y value is predicted which is the score for the corresponding essay. We have used scikit-learn library in python to implement the linear regression model.

4. EXPERIMENTAL RESULT

We also use Cohen's Kappa (Brenner et. al, 1996) which is a robust measure to quantify inter-rater agreement compared to percentage agreement as it also accounts for agreement occurring by chance. Quadratic weighted Cohen's Kappa is given by the formula.



$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$$

Where $O_{i,j}$ is the number of times the annotators assign the grade i and grade j respectively, $E_{i,j}$ is the is the expected number of times for the same event, given that both annotators randomly assign grades according to a multinomial distribution and $w_{i,j} = \frac{(i-j)}{(N-1)^2}^2$ and N is the number of possible grade levels. Cohen's Kappa is 1 when there is a perfect agreement and 0 when the agreement is random.

Essay Set	Kappa Score
1	0.85
2	0.71
3	0.71
4	0.8
5	0.61
6	0.67
Avg Kappa Score	0.725

5. CONCLUSION

In this paper, we propose relevant features for AES and integrate it with an existing improvised vector space model to achieve results comparable to expert raters. Although the proposed AES systems provides lucrative advantages such as saving time and better reliability in scoring, on some outliers, the absence of a human rater could result in missing out on inferential skills, critical thinking and abstract ideas. These form a scope for improvement for future essay evaluation systems to come.

6. REFERENCES

[1] Valenti, Salvatore, Francesca Neri, and Alessandro Cucchiarelli. "An overview of current research on automated essay grading." *Journal of Information Technology Education* 2 (2003): 319-330.

[2] Shermis, Mark D., and Jill C. Burstein, eds. *Automated essay scoring: A cross-disciplinary perspective*. Routledge, 2003.

[3] Burstein, Jill, Martin Chodorow, and Claudia Leacock. "Automated essay evaluation: The Criterion online writing service." *Ai Magazine* 25.3 (2004): 27.

[4] Rudner, Lawrence M., and Tahung Liang. "Automated essay scoring using Bayes' theorem." *The Journal of Technology, Learning and Assessment* 1.2 (2002).

[5] Powers, Donald E., et al. "Stumping E-Rater: Challenging the validity of automated essay scoring." *ETS Research Report Series* 2001.1 (2001).

[6] Dikli, Semire. "An overview of automated scoring of essays." *The Journal of Technology, Learning and Assessment* 5.1 (2006).