ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

A Survey on Automated Essay Scoring (AES)

Poornima M[1], Mrs Shibily Joseph[2]

PG Student, Dept of Computer Science & Engineering, Government Engineering College, Palakkad, Kerala, India[1]

Associate Professor, Department of Computer Science and Engineering, Government Engineering College, Palakkad, Kerala, India[2]

Abstract: Automated Essay Scoring (AES) is a challenging task which assigns grades to essays written in an educational institute. It reduces human errors, inequality problem, time consuming and so on. There are diverse approaches for this task like natural language processing, machine learning, deep learning etc. The overall performance of such systems is tightly bound to high-quality features. The essential purpose of this paper is to evaluate these strategies of AES in both in-domain and cross domain settings.

Keywords: Convolution Neural Network (CNN), Long Short-Term Memory (LSTM), Deep Learning, Natural language Processing

I. INTRODUCTION

Essay writing is used for assessing academic achievements of a student which is an expensive, time-consuming. Manual grading of essays takes up substantial amount of instructor's time; therefore it's an expensive method. Automating Essay Scoring (AES) is the tool which is cost effective and fastest method for grading the essay without inequality problems. Education institutes are developing new sorts of testing and grading strategies, to assess the new common core standards. Essays are a crucial expression of educational action, however they're costly and time overwhelming for states to grade them by hand. So, we have a tendency to be often restricted to multiple-choice standardized tests [1]. The tendency is to believe that automatic evaluation systems will yield quick, effective and cheap solutions that will enable states to introduce essays and different refined testing tools. Since the 1960s, there have been different approaches in creating AES methods. Different sorts of algorithms and models based on NLP, machine learning and deep learning techniques have been proposed to actualize AES systems. The AES system often disagrees with the first human rater, a second human rater may be needed in most instances. Therefore, the creation of the AES system does not bring a whole lot advantage in lowering the human workload. It is consequently applicable to reduce the disagreement between the machine and human raters.

Automatic Essay Scoring (AES) is the task of automatically assigning grades to student essays. It can be highly challenging, requiring not only knowledge on spelling and grammars but also on semantics, discourse and pragmatics. Traditional models use sparse features such as bag-of words, part-of-speech tags, grammar complexity measures, word error rates and essay lengths, which can suffer from the drawbacks of time-consuming, feature engineering and data sparsity there are various applications used for Automated Essay Scoring (AES) methods. Automatic essay grading is a very useful machine learning application. There are various methodologies for Automated Essay Scoring (AES) they are String kernels and word embedding, Rank based Algorithm, RNN CNN, Correlated Linear Regression etc. This survey aims to discuss different ways of grading the essays and thus highlight the advantages and disadvantages of each. Such a comparative study is very important as there is wide range of applications using Automated Essay Scoring. This survey will provide some insights for choosing the right methodology to develop Automated Essay Scoring (AES) methods. This paper is organized as follows: Section II gives a formal definition of the Automated Essay Scoring (AES). Section III discusses about various approaches used for grading the Essays and provides a comparison between the methodologies. Section IV discusses the future scope of AES system. Section V gives a brief conclusion of Automated Essay Scoring.

II. AUTOMATED ESSAY SCORING

Event Just several decades before the arrival of automatic essay scoring technology, the release of the IBM 805 automatic selected-response scoring machine in 1938 heralded a period of great expansion in the use of multiple-choice tests. From the beginning, this development was seen by some as an advance in the objectivity, economy, and speed of

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

standardized testing, and by others as a restrictive, inauthentic, and ultimately unfair means of measuring examinee knowledge and abilities [2]. The earliest example in the literature of the development of an automatic essay scoring tool is Project Essay Grade (PEG), developed by Ellis Page in the mid 1960s. Just as with the arrival of the IBM 805, PEG was met with both excitement and scepticism.

In general, existing solutions consider AES as a learning problem. Based on a large number of predefined objectively measurable features, various learning techniques, including classification, regression and preference ranking, are applied. The problem of Automated Essay Scoring can be formally defined as: Given an input essay of students, the goal of the AES system is to give score to essays based on the vocabulary, Grammar and Spelling Mistakes

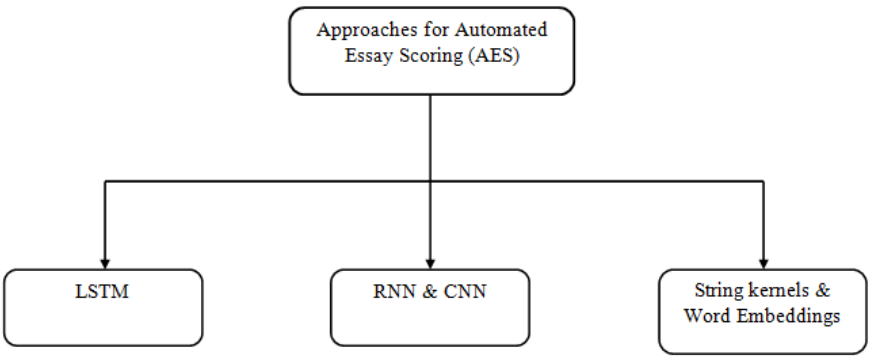

Fig. 1 Classification of Methodologies based on Approach Used in Automated Essay Scoring

III. DIFFERENT METHODOLOGIES OF AUTOMATED ESSAY SCORING

A. LSTM Based Approach

Feature extraction plays a significant role in any of the machine learning task and therefore here also we will utilize the concept of it. To build a robust yet effective essay scoring algorithm is to bank on model attributes like language fluency, grammatical and syntactic correctness, vocabulary and types of words used, essay length, domain information etc. At present, the model is making use of the following set of features extracted from the ASCII text of the essays. These are trivial features of any text document but it influences the scoring of the document as well. The features is extracted using text mining library, the extracted term-document matrix is to provide data to train corpus [3]. In this library a list of 276 common stop words in English language is also provided. Now, since these stop words are of not much use, we can skip them while we are calculating the word count of each document from the term document matrix.

To evaluate the quality of content in an essay is the foremost important task. Along with it an another important set of features for evaluating any piece of writing provide is the number of words in various syntactic classes like nouns, adverbs, verbs, adjectives etc. should also be evaluated . To get the counts of words in each POS (part-of-speech) class, NLTK library is used. This library provides us the POS tag for each word in an essay, and thus, we can extract the number of nouns, adverbs, adjectives and verbs separately. An important parameter while grading an essay is the spelling mistakes. Therefore, number of spelling mistakes in an essay has also been included as a feature for the model. To obtain the results for this, we use the spell checker provider named enchant.

By using Bayes theorem in essay scoring, expands the classification to a three or four point categorical or nominal scale (for example extensive, essential, partial, unsatisfactory) and introduce a large dataset. The contents in Bayesian essay scoring are mainly features of essay such as (specific words, phrases) and other characteristics of essay like the order in which certain noun-verb pair appears or the order of the concepts explained. Two models from the information science field were taken for creating student essays to test Bayesian essay scoring. The models chosen were improved using 462 essays and two score points. The improved system was tested with 80 new, pre-scored essays with 40 essays in each score group. Used variables were comprised of the two models; use of words, phrases and arguments; two ways for trimming; stemming; and usage of stop words. Although the text classification literature proposes urgent improvement on thousands of cases per group, even with the amount of inadequate data used in this study; accuracy over 80% was achieved.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

It is perhaps the most marked feature of this model is it tries to understand the semantics and information content of an essay. To get this feature working, we first figured out the best essay from each set (highest scored essay) then, we pulled out nouns from that essay. These nouns were served as keywords for the particular domain. Then, we fire these words into WordNet and take out their other equivalent. In this way, for each set, we got a bunch of different words, most relevant to its particular domain. Then, we count the number of domain words in the essay provided.

Automated essay grading is a very vital machine learning application. It has been studied quite a number of times, using different techniques like latent semantic analysis etc. This current approach tries to model the language features like language fluency, grammatical correctness, domain information content of the essays, and put an effort to fit the best polynomial in the feature space using linear regression with polynomial basis functions. The results which may be obtained will be quite encouraging and legitimate. We might achieve average absolute error that is significantly less than the standard deviation of the human scores. Across all domains used, the proposed approach appears to work very well. The future scope of the given problem can extend in various fields. One such area is to search and model good semantic and syntactic features. For this, various semantic parsers etc can be used. Other area of focus can be to come up with a better approach than linear regression with polynomial basis functions like neural networks.

The task of AES is usually treated as a supervised learning problem, typical models of which can be divided into three categories: classification, regression and preference ranking. In the classification scenario, scores are divided into several categories, each score or score range is regarded as one class and the ordinary classification models are employed such as Naive Bayes (NB) and SVMs [4]. In the regression scenario, each score is treated as continuous values for the essay and regression models are considered, like linear regression, Bayesian linear ridge regression [5]. In the preference ranking scenario, AES task is considered as a ranking problem in which pair-wise ranking and list-wise ranking are employed. The former considers the ranking between each pair of essays, while the latter considers the absolute ranking of each essay in the whole set.

B. Deep RNN / CNN Based Approach

Fei Dong et.al. [6] Neural network models have recently been applied to the task of automatic essay scoring, giving promising results. Existing work used recurrent neural networks and convolutional neural networks to model input essays, giving grades based on a single vector representation of the essay. On the other hand, the relative advantages of RNNs and CNNs have not been compared. In addition, different parts of the essay can contribute differently for scoring, which is not captured by existing models. We address these issues by building a hierarchical sentence-document model to represent essays, using the attention mechanism to automatically decide the relative weights of words and sentences. Results show that the model outperforms the previous state of-the-art methods, demonstrating the effectiveness of the attention mechanism.

The sequence of word embeddings obtained from the embedding layer is then passed into a long short-term memory (LSTM) network. The LSTM model is parameterized by output input and forgets gates, controlling the information flow within the recursive operation. For the sake of brevity, we omit the technical details of LSTM which can be found in many related works. At every time step t, LSTM outputs a hidden vector ht that reflects the semantic representation of the essay at position t. To select the final representation of the essay, a temporal mean pool is applied to all LSTM outputs. The main work-flow of the proposed approach is as follows. Firstly, a set of essays rated by professional human raters are gathered for the training. A list-wise learning to rank algorithm learns a ranking model or function using this set of human rated essays represented by vectors of the pre-defined features. Then the learned ranking model or function outputs a model score for each essay, including both rated and unrated essays, from which a global ordering of essays is constructed. Finally, the model score is mapped to a predefined scale of valid ratings, such as an integer from 1 to 6 in a 6-point scaleflow.

Current learning to rank algorithms fall into three categories, that is, the point wise, pair-wise, list wise approaches. Point wise approach takes individual documents as training examples for learning a scoring function. In fact, both multiple linear regression and support vector regression, which have been widely used in automated essay scoring, can be seen as point wise approaches. Pair-wise approaches process a pair of documents each time and usually model ranking as a pair-wise classification problem. Thus, the loss function is always a classification loss.

C. String Kernel and Word Embedding Approach

Formally, an AES model is trained to minimize the difference between its automatically output scores and human given scores on a set of training data. Long short-term memory units are the modified recurrent units which are proposed to

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

handle the problem of vanishing gradients effectively [7]. LSTMs use gates to control information flow, preserving or forgetting information for each cell units. In order to control information flow when processing a vector sequence, an input gate, a forget gate and an output gate are employed to decide the passing of information at each time step.

In String kernels, the Kernel functions capture the intuitive notion of similarity between objects in a specific domain. For example, in text mining, string kernels can be used to measure the pair wise similarity between text samples, simply based on character n-grams. Various string kernel functions have been proposed to date. One of the most recent string kernels is the histogram intersection string kernel [8].

Bag-of-super-word-embeddings. Word embeddings are long known in the NLP community (Bengio et al., 2003; Collobert and Weston, 2008), but they have recently become more popular due to the word2vec (Mikolov et al., 2013) framework that enables the building of efficient vector representations from words. On top of the word embeddings, Butnaru and Ionescu (2017) developed an approach termed bag-of-super-word-embeddings (BOSWE) by adapting an efficient computer vision technique, the bag-of-visual-words model), for natural language processing tasks. The adaptation consists of replacing the image descriptors useful for recognizing object patterns in images with word embeddings useful for recognizing semantic patterns in text documents.

The centroid of each cluster is interpreted as a super word embedding or super word vector that embodies all the semantically related word vectors in a small region of the embedding space. Every embedded word in the collection of documents is then assigned to the nearest cluster centroid (the nearest super word vector). Put together, the Super word vectors to generate a vocabulary (codebook) that can further be used to describe each document as a bag-of-super-word-embeddings. To obtain the BOSWE representation for a document, compute the occurrence count of each super word embedding in the respective document. After building the representation, we employ a kernel method to train the BOSWE model for specific task.

D. Discussion on datasets for Automated Essay Scoring

The Automated Student Assessment Prize (ASAP) dataset as evaluation data for our task, which contains 8 prompts of different genres as listed in Table 2. The essay scores are scaled into the range from 0 to 1. The settings of data preparation follow [9]. We use quadratic weighted kappa (QWK) as the metric. For domain adaptation cross-domain experiments are done.

Arxiv Academic Paper Dataset: As there is no existing dataset that can be used directly, we create a dataset by collecting data on academic papers in the field of artificial Intelligence from arxiv website. The dataset consists of 19,218 academic papers. The information of each source paper consists of the venue which marks whether the paper is accepted, and the source LATEX file. We divide the dataset into training, validation, and test parts

IV. FUTURE RESEARCH DIRECTIONS

There are few suggestions for building better Automatic Essay Scoring systems. The effectiveness of using domain adaptation is that only a small number of target domain essays are used. We have shown that domain adaptation can achieve better results compared to using just the small number of target domain data or just using a large amount from a different domain. As such this research will help to

Table I: Overview of Compared Approaches and Datasets Used

Type of Approach	**Proposed Approaches**		
	Model	**Highlight**	**Dataset**
LSTM	V. V.Ramalingam et.al. [2]	Bayes Theorem	Automated Student Assessment Prize (ASAP)
RNN and CNN	Pengcheng Yang et.al. [4]	Modularized Hierarchical CNN	Arxiv Academic Paper Dataset
String kernel and Word embedding	Fei Dong et.al.[7]	Convolutional Neural network	Automated Student Assessment Prize (ASAP)
	Yue Zhanget.al.[8]	Long Short-Term Memory	Automated Student Assessment Prize (ASAP)

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

The AES systems can be deployed in two different manners, namely prompt-specific and generic. A prompt-specific rating model is built for a specific prompt and designed to be the best rating model for the particular prompt. For different prompts, the features used, their weights, and scoring criteria, may be different. It usually requires several hundreds of graded essays for training, which is time-consuming and usually the same for all prompt and there for has validity-related advantages.

Reduce the amount of annotation work needed to be done by human graders to introduce a new prompt.

CONCLUSION

In this paper, the task of Automatic Academic Paper Rating (AAPR), which aims to automatically determine whether to accept academic papers. We propose a novel modularized hierarchical CNN for this task to make use of the structure of a source paper. Experimental results show that the proposed model outperforms various baselines by a large margin. In addition, the conclusion and abstract parts have the most influence on whether the source paper can be accepted or not. The abstract and conclusion gives better result for the acceptance of paper.

REFERENCES

[1]. H. Chen and B. He, "Automated essay scoring by maximizing human machine agreement," in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1741–1752, 2013.
[2]. G. K. Chung and H. F. O'Neil Jr, "Methodological approaches to online scoring of essays.," 1997.
[3]. M. Mahana, M. Johns, and A. Apte, "Automated essay grading using machine learning," Mach. Learn. Session, Stanford University, 2012.
[4]. L. S. Larkey, "Automatic essay grading using text categorization techniques," in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 90–95, ACM, 1998.
[5]. T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," Discourse processes, vol. 25, no. 2-3, pp. 259–284, 1998.
[6]. F. Dong, Y. Zhang, and J. Yang, "Attention-based recurrent convolutional neural network for automatic essay scoring," in Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 153–162, 2017.
[7]. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[8]. M. Cozma, A. M. Butnaru, and R. T. Ionescu, "Automated essay scoring with string kernels and word embeddings," arXiv preprint arXiv:1804.07954, 2018.
[9]. P. Phandi, K. M. A. Chai, and H. T. Ng, "Flexible domain adaptation for automated essay scoring using correlated linear regression," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 431–439, 2015.