Get IT Scored using AutoSAS - An Automated System for Scoring Short Answers

Yaman Kumar
Adobe
ykumar@adobe.com

Swati Aggarwal
NSUT-Delhi
swati@nsit.ac.in

Debanjan Mahata
Bloomberg
dmahata@bloomberg.net

Rajiv Ratn Shah
IIIT-Delhi
rajivratn@iiitd.ac.in

Ponnurangam Kumaraguru
IIIT-Delhi
pk@iiitd.ac.in

Roger Zimmermann
National University of Singapore
rogerz@comp.nus.edu.sg

Abstract

In the era of MOOCs, online exams are taken by millions of candidates, where scoring short answers is an integral part. It becomes intractable to evaluate them by human graders. Thus, a generic automated system capable of grading these responses should be designed and deployed. In this paper, we present a fast, scalable, and accurate approach towards automated *Short Answer Scoring* (SAS). We propose and explain the design and development of a system for SAS, namely AutoSAS. Given a question along with its graded samples, AutoSAS can learn to grade that prompt successfully. This paper further lays down the features such as *lexical diversity*, *Word2Vec*, *prompt*, and *content overlap* that play a pivotal role in building our proposed model. We also present a methodology for indicating the factors responsible for scoring an answer. The trained model is evaluated on an extensively used public dataset, namely *Automated Student Assessment Prize Short Answer Scoring* (ASAP-SAS). AutoSAS shows state-of-the-art performance and achieves better results by over 8% in some of the question prompts as measured by Quadratic Weighted Kappa (QWK), showing performance comparable to humans.

Introduction

Essays and other types of writing practices have been extensively used for evaluation purposes. Graduate Record Examination (GRE), Scholastic Aptitude Test (SAT), Senior School Examinations such as Zhongkao in China and All India Senior School Certificate Examination (AISSCE) in India are just some of the many examples. The stakes for getting high grades in the essays and hence in these exams are tremendous for *pupils*, *teachers* and *schools* alike. The essays and short answers written by the students in the exams determine their future colleges and hence have a career wide impact.

Under the *No Child Left Behind Regulations*, U.S. States have been asked to use uniform and regulated test scores for evaluation of teachers for determining their salaries and tenures (Higgins 2014). This underlines the importance of getting good scores in these writing practices. A school's reputation is often determined by the SAT scores of its graduating students, which in turn is impacted by how well they have been taught to write their essays and short answers (Dale and Krueger 2002).

Motivation

Automated Essay Scoring (AES) and *Short Answer Scoring* (SAS) systems such as the one presented in this work (AutoSAS), provides economic advantages to testing companies and state-wide corporations. These systems reduce the economic and time burden of getting each response checked by human graders close to zero, bringing down the cost and effort significantly.

It has been noted that about 30% of a teacher's time is spent in evaluating students that subsequently translates to close to 4.02 Billion US Dollars per year coming from the taxpayers (Mason and Grove-Stephensen 2002). To eliminate this, it is necessary to design an automated system that a teacher can trust, and can use to mark essays and short text responses. In addition, one often hears about biases in marking students based on region, religion, and ethnicity. AES systems can possibly aid in uprooting any such biases from the education system.

Currently, AES systems have been successfully deployed by Educational Testing Service[1], where GRE essays are graded by a human grader as well as an AES system (Burstein et al. 1998). A second human grader is required only if there is a non-negligible difference between the two grades. AES systems form a major use case for Massive Open Online Classes (MOOCs) where economies of scale are required. As MOOCs advent towards offering courses in subjects such as literature and humanities, a range of assessment techniques such as AES will come in handy.

Systems similar to AutoSAS can be deployed in other agencies where it can not only be used for reducing the economic cost related to grading, but also in providing a scalable system for uniform grading in a time-bound manner. Students can also benefit from use of these systems if they can verify and check their work before actually submitting it for final reviews.

In this work, we primarily focus on utilizing natural language processing (NLP) for the task of Short Answer Scoring (SAS), which involves automated scoring of short answers provided for a given prompt that presents questions

[1] https://www.ets.org/

from a fixed set of subjects. This problem is typically formulated as a supervised learning problem where samples are graded on an ordinal scale (say, 1-10). AES as a NLP problem has been studied extensively (Balfour 2013; Xi 2010; Valenti, Neri, and Cucchiarelli 2003; Yang et al. 2002), with a variety of methods being deployed to perform the task of grading. We attempt to solve it as a supervised regression problem and develop a system, namely Automated Short Answer Scoring (AutoSAS) on top of a popular publicly available dataset.

The quality of text from the perspective of scoring short answers is dependent on many factors, some of them being *content*, *grammar*, *vocabulary*, *flow*, *coherence of ideas* and *relevance to the topic*. We propose a novel model utilizing many of these elements to grade short answers. A thorough overview of all the features used in building AutoSAS is presented. Using these features, AutoSAS performs better than the current state-of-the-art models for grading students' short answers. For some of the prompts, the improvements are more than 8%.

AutoSAS provides a listing of all the features that are important for the grading of a particular response in a ranked manner. It also presents a listing of the features which contributed to the score of a particular candidate. Through this raw feedback, the students can assess their weak areas. This may serve as an invaluable feedback for the students. Using the feedback, they can improve their writing before submitting the final document. This can also be used as an alternative as well as to augment the feedback that teachers provide while they grade student responses.

Contributions

Towards the objective of scoring short answers we make the following contributions in the work presented in this paper:

1. We present a supervised model for automatically scoring short answers that shows state-of-the-art performance with improvements of more than 8% in certain sets of question prompts as measured by Quadratic Weighted Kappa (QWK).
2. We augment existing features used by previous works with new set of features along with their ablation study.
3. AutoSAS as a system can not only score short answers but can also possibly find its usage in providing feedback to its users about their response, giving a detailed overview of what went behind its decision making process.

Next, we briefly present some previous research and systems that are relevant to the scope of our work in this paper.

Related Work

Different authors and organizations have ventured into building AES systems extending the Project Essay Grade (PEG) (Page 1994), which is one of the first systems developed. AES is ordinarily considered as a regression or a classification problem. In regression based analysis, essay score is considered to be a dependent variable and depends on values of features of an essay. These features are then used to learn a regression equation which is further used for grading essays.

In classification based approaches, essays are segregated into different classes according to their scores. These classes then form the basis for segregating and scoring the future essays. Different techniques and models have been developed on various datasets. *E-Rater*, a system designed by Educational Testing Services (ETS) (Attali and Burstein 2006), utilizes stepwise regression analysis on diverse linguistic features. It is used in popular exams like GRE and Test of English as a Foreign Language (TOEFL).

Apart from standard features such as *TF-IDF*, *word frequency analysis*, many have attempted to use more diverse and heterogeneous set of features for this task such as *lexical chain* (Somasundaran, Burstein, and Chodorow 2014), *students' demographic information*, *reading comprehension*, *vocabulary knowledge*, *writing apprehension*, among others for scoring essays (Crossley et al. 2015).

There have been many previous attempts to automate the scoring of short answers as well as essays. The methods utilized have ranged from regression to classification based supervised learning. In spite of the breadth, the previous works considered a very restrictive set of features (Chen et al. 2010) which were often hand-picked and were restrictive to the domain they were applied to (Ramachandran, Cheng, and Foltz 2015; Bachman et al. 2002; Riordan et al. 2017), compromising the reliability and accuracy of predictions substantially.

We propose a supervised model, namely AutoSAS, which is developed to grade short answers. This model shows a significant improvement over the current state of the art technologies in the accuracy of predictions and scalability over disparate domains. In summary, this paper presents a simple to use, fast and reliable approach to grade short answers that can be easily used in a classroom setting.

Task Overview

In this section, we present the details of the dataset used, a comparison of AES and SAS, and the reasons behind SAS being more difficult in nature. Then, we detail the features used by AutoSAS for grading short responses.

Topic	Question
Science	Replicate an experiment based on the details of another experiment
Arts	Similarity between Pandas, Koalas and differences wrt. Pythons
Biology	Describe protein synthesis
English	Describe a character Mr. Leonard

Table 1: Sample questions from the dataset.

Dataset

We conduct the experiments on a public dataset released by Automated Student Assessment Prize (ASAP) competition hosted in Kaggle[2], and was sponsored by Hewlett Foundation. This is the largest publicly available dataset, consisting of student responses for a total of 10 different questions and more than 16000 responses. This dataset has

[2] https://www.kaggle.com/c/asap-sas/

also been popularly used by researchers who have reported works similar to us (Ramachandran, Cheng, and Foltz 2015; Riordan et al. 2017).

The responses were written by high school students and then manually graded and double scored (on a scale of 0-3) by the ASAP graders. The questions covered a range of topics from Science to Language and Arts. A brief overview of the questionnaire is presented in Table 1. The questions belonged to a variety of topics, with their response length ranging from 1 word to 300 words, with an average of 50 words. Information provided in the responses ranged from the question itself (verbatim in some cases) to the author's preformed knowledge. Due to the realistic nature (non-lab environment) and diversity of the dataset, it is ideal for our analysis.

Automatic Essay vs Short Answer Scoring:

Although AES and SAS as NLP tasks have a lot in common, yet SAS is significantly different from AES in following ways:

- **Length of Response** : Short answers are typically shorter in length than essays. This means, essentially, that the author is constrained to present his ideas in a shorter response, and has scope to present fewer ideas. This is a challenge for an automated system, as there are lesser number of tokens (words, phrases *etc.*) that are related to the domain about which the writer is writing (Ramachandran, Cheng, and Foltz 2015).
- **Genre** : Essays emphasize on *narration* and *imagination*, which is not possible in short answers since they are required to be precise and to the point. The model short answers should cover all the major points respecting the space provided. For a concrete example, an intrigued reader can access the marking scheme of the questions asked in the datasets from the data set description.

Next, we describe the various groups of features that were extracted from the responses and forms the basis of our trained model and further analysis. Wherever applicable we point to the previous studies that has used any of them.

Features

- **Word2Vec and Doc2Vec based Features:** Word2Vec (Mikolov et al. 2013) and Doc2Vec (Le and Mikolov 2014) are useful techniques that capture semantic relationships between words and documents from the different contexts in which they occur. We used pre-trained Word2Vec and Doc2Vec models trained on Google News corpus and Wikipedia dump, respectively. These corpuses are used so as to model the generic nature of the question prompts. The questions in the dataset ranges from Science to Arts and Literature. Sentences from such a variety of topics can only be covered by a corpus such as that of news reported from different domains and encyclopedic entries on a vast range of topics. Word2Vec although provides high quality word vectors but averaging them makes them lose their order information. Thus Doc2Vec was also used over the full short answer.
- **Part of Speech (POS) Tagging:** POS tags based n-grams capture context very well. Each response word was tagged with its corresponding part-of-speech (eg., Verb, Noun, Preposition). To take care of the context information, POS-tagged bi-grams, tri-grams and tetra-grams were extracted from the prompt. To avoid the trivial n-grams, based on the training set, a list of significant bi-grams, tri-grams and tetra-grams was constructed. For constructing such a list, we used the responses that were graded high by the human graders. For a particular response, the n-grams which are present in this set were considered and all others were ignored. The final selection of the n-gram set was based on considering only those that have greater than a particular incidence count. The threshold was determined by running the model on validation data, keeping other features constant.
- **Weighted Keywords:** Certain prompts demand a set of domain-specific keywords to be present in them. For getting a list of keywords, we took following steps. Firstly, a set of domain-specific words were identified from the set of answers to questions related to a specific subject. Then using Google API, the top 20 articles related to that word were extracted. Each page was subsequently scraped and the term frequencies of keywords were stored. Then, tf-idf values were calculated for each word, thus getting a list of keywords based on their tf-idf importance.
- **Prompt Overlap:** Short answers derive some of their information from the question itself. Therefore, an overlap between the prompt and the response serves as an important grading metric for any grader. As an example, in reading comprehension based prompts, the answers derive their context and content from the comprehension and the question itself. This content can be in the form of taking information of subject, verb, argument, time, from the question itself and then using this information to answer it. For example, one of the questions given in the dataset, "Based on Rose's conversations with two other characters, describe her." This question requires an answer that includes context from the text of the question. Thus an overlap between the answer and a question is expected and necessary.
- **Lexical Overlap:** In many of the prompts, even after taking prompt overlap, many words that can be found in the questions such as reading comprehensions were not captured. These words can only be extracted from the comprehension itself, neither Google keywords, nor a simple overlap with question prompt would yield something significant. Thus, AutoSAS takes into consideration different types of lexical overlaps between the sentences present in the short answers: *Noun Overlap*, *Argument overlap*, and *Content Overlap*.

  Noun Overlap is a measure of the frequency of overlap of nouns across two sentences. *Argument Overlap* is measured by the overlapping intersection of arguments among sentences. We extract argument from a sentence using a list of hand-crafted heuristics. *Content overlap* measures the amount of overlap of content words across sentences. Wherever applicable, this set was further extended to in-

clude the words' synonyms using WordNet[3]. All the overlaps are calculated w.r.t the reading comprehension text. For example, in case of noun overlaps if there are 5 nouns in the comprehension text and the answer has 3 nouns the overlap is calculated to be $3/5$. These scores are further normalized across individual prompt sets.

Previous studies show that lexical overlap significantly aids in text analysis. (Rashotte and Torgesen 1985; Ferris 1994; Douglas 1981). For instance, response no. 19953 in the dataset states - Paul finds out that Mr. Leonard was a track star but he could not read. 'No school wanted a runner who couldn't read'. Thus the *Nouns*- Mr. Leonard, school runner, Paul match with what is given in the comprehension. The overlap between the *Argument*, "No school wanted a runner...." matches with the prompt. This is one of the reasons of this response's high grades (it scored a 2).

- **Word Frequency, Difficulty and Diversity:** Word frequency and diversity indicate a student's command over language. Although, many earlier works (Nation and Heatley 1996), use this technique but they look at the frequency of top k words only. We find this approach non-holistic. This is so since the top words (frequency wise) can be more quickly accessed by a writer and thus are consequently easier to decode by a reader (Perfetti 1985; Rayner 1989), whereas, as indicated by many studies (Frase et al. 1998; Reppen 1995; Reid 1990) writers using lesser frequent words are, generally, more proficient than others. Thus with this in mind, Webster Dictionary[4] was divided into 20 different levels of words with varying difficulties (Breland et al. 1994). Then each word used by the candidate was mapped to a difficulty level using this dictionary. The frequencies of words for each of the difficulty levels were noted as features of that response.

  In addition, the number of unique words that appear in each response and *Type Token Ratio* (TTR) (Templin 1957) serve as another set of features. TTR is a value which ranges from 0 to 1 and is indicative of the lexical diversity of a prose. The writer with a larger vocabulary is generally more proficient and hence is better graded than a writer using limited vocabulary (Engber 1995; Reppen 1995).

- **Statistics of Sentence and Word Length:** In general, sentence length indicate the complexity of the sentence, with longer sentences requiring greater use of working memory and hence being more difficult to understand. Thus word and sentence length can be used to indicate the sophistication of a writer (Hiebert 2011). Several research projects have shown that higher-rated essays, in general, contain more words (Carlson et al. 1985; Ferris 1994; Reid 1990) and generally use longer words (Frase et al. 1998; Reppen 1995). Thus sentence length, word length, average sentence and word length were noted down for each response and used as features.

- **Logical Operators based Features:** Logical operators are directly related to the number, density and abstractness of ideas of a student. These then translate to the quality of arguments in a text (Fayol 1997). Some other studies have also used them, for example, Coh-Metrix (Crossley and McNamara 2012). AutoSAS uses *and, or, not, if-else, if-then, unless, whether, although, but* and their various other combinations as logical operators for grading purposes. As an example, response with Id 231, states, "*If* I used different amounts of water *when* washing the samples, one may not be as thoroughly washed as another *which* could mess up the results." The logical operators combination *if*, *when* and *which* indicate the logical complexity of the sentence.

- **Temporal Features:** Temporal features such as tense and aspect words help the grader in forming a timeline of events thus enhancing the validity and pithiness of the arguments of a student. While tense helps us in formation of a sequence, aspect represents the dynamics of the events with respect to time (Klein 2013). It has been argued (McCarthy et al. 2007) that repeating tenses and aspects in the text create more cohesion in the arguments presented in a response, hence improving its quality. Some other studies (Crossley and McNamara 2012) have also used aspect and tense based features to extract temporal features. Thus various tense and aspect words were identified in a response and then, were associated with events.

  For instance, response with Id 8078, responded, "I *believe* in this article 'invasive' *means* hidden/unchecked or *referred* to pythons MaccInnes *uses* this word when he *states* ," I *think* that invasive *is passing* judgement" he *used* this because he in happy that pythons *are going* birth...." Due to wrong usage of tenses, among other reasons, it scored a 1.

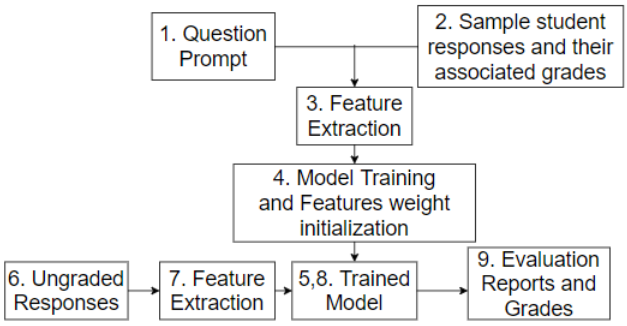

Figure 1: Pipeline for scoring short answers using AutoSAS.

Experiments

In this section, we provide a detailed description of the supervised training process of our model. The complete pipeline of the proposed model, AutoSAS is shown in Figure 1. For training the model, it is required that the users (*i.e.* teachers, schools, interviewers, *etc.*) grade some of the responses, and provide them as an input to AutoSAS along

[3]https://wordnet.princeton.edu/

[4]https://www.https://www.merriam-webster.com/

with their respective question prompts. The question prompt is required for extracting features, as explained in the previous section. Different sets of question prompts will have different number of features. The sample responses given to train the model will help to finalize the weights of these features, which will then be used for grading of the ungraded responses. Once the grading has been done successfully, it also produces feedback for a given response.

Augmentation of Training Set using Jumbled Content

For our experiments, we also augmented the training set using jumbled content. Two types of jumbled content were included along with the normal responses in order to train AutoSAS for a particular prompt. This was done in order to avoid grading those responses highly, that were written well, but had content irrelevant to the question asked.

We included 10 highly rated responses from prompts other than the one for which the model is being trained for and gave them the lowest grade possible. This made sure to penalize the irrelevant responses. Another source for jumbled content was from the answers of the same prompt on which AutoSAS was being trained. After an initial training for that particular prompt, some of the these otherwise highly graded responses were taken, jumbled up and then included with the training samples assigning them lowest possible grades. This was done to avoid grading those responses highly, which included a soup of gibberish keywords related to the question while not having context, connecting information (Perelman 2014).

Training the Model

Firstly, all the responses for a particular prompt are prepared. The responses are obtained either from the dataset corresponding to a particular prompt or are obtained via dataset augmentation. Then, each response is checked for spelling mistakes as it is mentioned in most of the grading rubrics grammar, penmanship and spellings were important for the clarity of responses but were not important for scoring. Words involving scientific names such as chemical compounds, proper nouns and other tokens which are not found in the dictionary but are employed in the responses are taken care of appropriately during spell correction. The set of features as described previously were subsequently extracted from the responses and the questions. Responses from different prompts were saved separately and subsequently loaded in different dataframes.

In order to train AutoSAS for a particular question, all the features and grades of that question are loaded in a dataframe, which is then used for *regression analysis*. For splitting the total data into train, validation and test data, a ratio of 70:10:20 was used. It is to be noted that none of the jumbled responses were included in the test data. The test set solely consisted of the original data. Other type of data such as jumbled responses were included solely for training purposes. Testing and training data was stratified so as to get an equalized distribution of samples across all grades.

A *Random Forest* model was trained on all the features. Random forest model has been used chiefly because of two reasons. Firstly, it performs well over the multitude of features extracted from the responses (as presented in the results). Secondly, in order to know the importance of each feature set used in the analysis, which is not possible with popularly used neural network based approaches (as presented in Ablation study). An example of feedback produced by AutoSAS is given in Figure 2. With multi layer neural networks (RNN, CNN), one can feed the network a listing of all the features but cannot easily expect the network to tell the performance of each feature for a particular response (Leray and Gallinari 1999; Montavon, Samek, and Müller 2017).

In the case of *Random Forest* model, we get the performance of all the features on a particular response by making use of the package TreeInterpreter[5]. It gives a list of contribution of each feature in getting the grade for a particular response. For a task of such nature as grading is, students expect the teacher to give atleast a crude indication of the reason behind the scores assigned to them. This feedback might be useful for a student's improvement. This makes *interpretability* of the model an important task. Though we do not explore an exhaustive mechanism for comments and review part of the grading process, however, a crude indication of the scores assigned is presented.

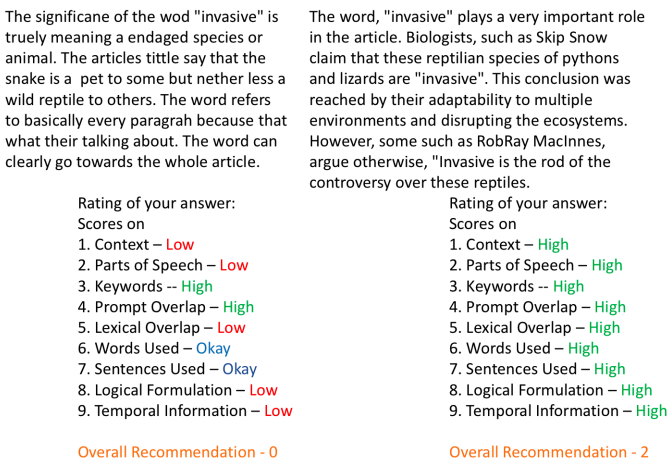

Figure 2: An example of feedback produced by AutoSAS.

As a part of our future work, we would like to explore the proposition of giving a more exhaustive explanation for the grades assigned, similar to some other recent work (Chen et al. 2018) dedicated to this goal. With a *Random Forest* model, AutoSAS presents a listing of all the features computed for a particular response along with the importance of each such feature in its grading process. Thus, the author of the response has the opportunity to know what are his/her weak areas, and can focus on them. The results were computed using the process mentioned, and are presented in the next section along with a comparison of the current state-of-the-art models.

[5] https://github.com/andosa/treeinterpreter

Approach	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Mean
AutoSAS	**0.872**	**0.824**	**0.745**	**0.743**	**0.845**	0.858	**0.725**	0.624	**0.843**	**0.832**	**0.791**
Ramachandran et al.	0.86	0.78	0.66	0.70	**0.84**	**0.88**	0.66	**0.63**	**0.84**	0.79	0.78
Riordan et al.	0.795	0.718	0.684	0.700	0.830	0.790	0.648	0.554	0.777	0.735	0.723

Table 2: Comparison of performance of models on the dataset, ASAP-SAS. The data presented is QWK scores for each of the ten prompts in the dataset.

Results

Evaluation Metrics

We use Quadratic Weighted Kappa (QWK) (Brenner and Kliebsch 1996) as the evaluation metric for finding the agreement between the grades predicted by AutoSAS and the human graders. This metric was chosen since this was used in the official competition of ASAP-SAS. Other works (Chen and He 2013; Ramachandran, Cheng, and Foltz 2015) also used it to evaluate their results. It calculates the level of agreement between the two raters. It also takes into account the *by chance* probability of assigning the same grade to a sample by both the raters.

Quadratic Weighted Kappa is calculated as follows. Firstly, the weight matrix W is constructed according to Equation 1. Here i is the reference rating of human rater, j is the rating assigned by the model and N is the total number of possible ratings.

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{1}$$

After this, *QWK* is calculated as:

$$\kappa = 1 - \frac{\Sigma_{i,j} W_{i,j} O_{i,j}}{\Sigma_{i,j} W_{i,j} E_{i,j}} \tag{2}$$

Here matrix O contains the observed scores such that rating i is given by human grader and j is given by the model. $W_{i,j}$ contains the weights as derived in Equation 1 and E contains the expected scores obtained by multiplying the histogram vectors of the two scores *i.e.* the ones by human graders and the other by the proposed model AutoSAS. Subscripts in Matrix $O_{i,j}$ correspond to the number of essays that score i from the first rater and j from the second one.

Experimental Setup

We conducted the experiments on a machine with Intel(R) Core(TM) i5-3210M CPU @2.50GHz with a 10.0 GB of RAM and Operating System as 64-bit Windows 10. The Internet speed is close to 8 Mbps. On this system, it took slightly lesser than 25 minutes (average time) to extract features, train and test the model for a particular set of question prompt. This indicates that AutoSAS could be used by common students, schools, teachers and MOOCs alike without much overhaul of their existing systems. Such scoring can even happen on a teacher's or a student's personal computer and not just on a powerful laboratory computer.

Discussion

The results for the evaluation of AutoSAS and those of the models used by (Ramachandran, Cheng, and Foltz 2015) and (Riordan et al. 2017) are presented in Table 2. Ramachandran *et al.* used word order graph in order to capture the order of the tokens and lexico-semantic matching technique for identifying the degree of relatedness across tokens and phrases. They replaced the manually coded patterns used in the best performing model in Kaggle with the automatically generated patterns produced by their method, and used them as features for training a Random Forest model. Riordan *et al.* used neural networks with n-grams and word embeddings as features. The performance of their systems are directly reported from their papers. As shown, AutoSAS outperforms (Riordan et al. 2017) on each of the prompts. AutoSAS also outperforms model given by (Ramachandran, Cheng, and Foltz 2015) on 6 out of 10 sets. In the remaining 4 of the short answer sets, it performs equally well in 2 of them, in one of the sets it performs slightly worse and in set 6 it lags behind the other models.

AutoSAS performs exceptionally well on set 3, performing 8.8% better than the current best model of (Riordan et al. 2017). The question for set 3 asks students to explain the similarities between Pandas in China and Koalas in Australia and how they are different from Pythons. This question demands some specific details from the students as its answer. These details can be found in the comprehension question prompt given to them. Thus majority of the answer can be derived from the question itself, but many responses go beyond the details mentioned in the question. They include some details which cannot be derived just from the question but require some prior knowledge. This is what most hand-tailored approaches and features such as those used by (Ramachandran, Cheng, and Foltz 2015) fail to grasp. Prior knowledge of the subject can be fed into the training models only by introducing it to the texts that contain those specific concepts and facts. Only then it can effectively grasp what the student have written, and in the process, modeling what the teacher would have done when faced with a similar scenario. AutoSAS does this task by acquiring information that is outside the purview of the prompt using the features such as *Weighted Keywords* and *Word2Vec/Doc2Vec* embeddings.

Next, we show the feature importance and ablation study. It is worthy to note that neither of the works (Ramachandran, Cheng, and Foltz 2015; Riordan et al. 2017) with which we compare our performance have conducted such a study. Thus an important aspect of the grading process is absent from the present state-of-the-art systems as reported to the research community.

Ablation Study

Table presents the findings of various feature groups. It lists the rankings of the feature groups as well as the fall in accu-

racy observed after removing the said features. For getting the fall in accuracy value for a particular group, the features extracted from that group were removed while keeping the other groups intact. The smaller set of features are fed to the Random Forest model and the results computed are presented.

Word2Vec and *Doc2Vec* were the most important features. This might be due to the fact that the dataset is generic in nature and is represented well by these embeddings. Features that are based on in-domain information are also highly valued. The examples of such features are *prompt information*, *weighted keywords*, *lemmatized response* and *lexical overlap*.

The additional features such as *word frequency*, *difficulty*, *statistics of word* and *sentence length* do not figure highly neither in the rankings nor are the accuracy values being affected significantly. But, in any case, they do prove to be useful for predicting the scores.

Although, it was mentioned that the graders did not consider word frequency, penmanship, for grading a particular response, but as indicated, these biases do show up in the gradings either knowingly or unknowingly. With AutoSAS, it is a virtue of the system that one can turn off these features if one does not want to take into consideration these specific details.

Rank	Feature Group	Fall in Accuracy
1	Word2Vec, Doc2Vec	23.54%
2	Prompt Overlap	20.85%
3	Weighted Keywords	16.93%
4	POS Tags	12.36%
5	Lexical Overlap	8.45%
6	Logical Operators	6.40%
7	Temporal features	4.2%
8	Stats of Sentence and Word Length	2.11%
9	Word Freq, Difficulty	1.02%

Table 3: Importance of various features used in AutoSAS.

Conclusion and Future Work

In this work, we proposed a supervised regression model and explored different linguistic features for grading short answers, and a system encompassing it named AutoSAS, which can be easily used and deployed in various educational and professional testing settings. Experiments on the publicly available dataset ASAP-SAS showed that AutoSAS outperforms the current state-of-the-art algorithms and approaches. According to (Powers et al. 2000), the agreement between machine learning models and expert human graders range between 0.7 to 0.8, and AutoSAS achieved a mean QWK score of 0.79.

We also showed how AutoSAS can be useful for assessing the decision making process for assignment of a score and provide valuable feedback to the users about the characteristics of a response. Unlike the existing state-of-the-art systems, we perform an ablation study and discuss about the most important features that contribute towards the performance of our trained model. One of the major aspects where AutoSAS still lacks is review comments. We would like to work on it in the future and also try out hybrid methods that takes into account the Random Forest model along with a deep neural network architecture in order to improve our current system.

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China under Grant no. 61472266 and by the National University of Singapore (Suzhou) Research Institute, 377 Lin Quan Street, Suzhou Industrial Park, Jiang Su, People's Republic of China, 215123.

MIDAS lab gratefully acknowledges the support of NVIDIA Corporation with the donation of a Titan Xp GPU used for this research.

References

Attali, Y., and Burstein, J. 2006. Automated essay scoring with e-rater [r] v. 2. *Journal of Technology, Learning, and Assessment* 4(3).

Bachman, L. F.; Carr, N.; Kamei, G.; Kim, M.; Pan, M. J.; Salvador, C.; and Sawaki, Y. 2002. A reliable approach to automatic assessment of short answer free responses. In *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, 1–4. Association for Computational Linguistics.

Balfour, S. P. 2013. Assessing writing in moocs: Automated essay scoring and calibrated peer review (tm). *Research & Practice in Assessment* 8.

Breland, H. M.; Jones, R. J.; Jenkins, L.; Paynter, M.; Pollack, J.; and Fong, Y. F. 1994. The college board vocabulary study. *ETS Research Report Series* 1994(1).

Brenner, H., and Kliebsch, U. 1996. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology* 199–202.

Burstein, J.; Braden-Harder, L.; Chodorow, M.; Hua, S.; Kaplan, B.; Kukich, K.; Lu, C.; Nolan, J.; Rock, D.; and Wolff, S. 1998. Computer analysis of essay content for automated score prediction: A prototype automated scoring system for gmat analytical writing assessment essays. *ETS Research Report Series* 1998(1).

Carlson, S. B.; Bridgeman, B.; Camp, R.; and Waanders, J. 1985. Relationship of admission test scores to writing performance of native and nonnative speakers of english. *ETS Research Report Series* 1985(1).

Chen, H., and He, B. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1741–1752.

Chen, Y.-Y.; Liu, C.-L.; Lee, C.-H.; Chang, T.-H.; et al. 2010. An unsupervised automated essay scoring system. *IEEE Intelligent systems* 25(5):61–67.

Chen, J.; Song, L.; Wainwright, M. J.; and Jordan, M. I. 2018. Learning to explain: An information-theoretic perspective on model interpretation. *arXiv preprint arXiv:1802.07814*.

Crossley, S. A., and McNamara, D. S. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35(2):115–135.

Crossley, S.; Allen, L. K.; Snow, E. L.; and McNamara, D. S. 2015. Pssst... textual features... there is more to automatic essay scoring than just you! In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 203–207. ACM.

Dale, S. B., and Krueger, A. B. 2002. Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics* 117(4):1491–1527.

Douglas, D. 1981. An exploratory study of bilingual reading proficiency. *Learning to read in different languages* 33–102.

Engber, C. A. 1995. The relationship of lexical proficiency to the quality of esl compositions. *Journal of second language writing* 4(2):139–155.

Fayol, M. 1997. On acquiring and using punctuation: A study of written french. *Processing interclausal relationships. Studies in the production and comprehension of text* 157–178.

Ferris, D. R. 1994. Lexical and syntactic features of esl writing by students at different levels of l2 proficiency. *Tesol Quarterly* 28(2):414–420.

Frase, L. T.; Faletti, J.; Ginther, A.; and Grant, L. 1998. Computer analysis of the toefl test of written english. *ETS Research Report Series* 1998(2).

Hiebert, E. H. 2011. Beyond single readability measures: Using multiple sources of information in establishing text complexity. *Journal of Education* 191(2):33–42.

Higgins, J. 2014. Loss of waiver will mean districts labeled failing. *The Seattle Times (Seattle, WA)*.

Klein, W. 2013. *Time in language*. Routledge.

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, 1188–1196.

Leray, P., and Gallinari, P. 1999. Feature selection with neural networks. *Behaviormetrika* 26(1):145–166.

Mason, O., and Grove-Stephensen, I. 2002. Automated free text marking with paperless school.

McCarthy, P. M.; Lehenbauer, B. M.; Hall, C.; Duran, N. D.; Fujiwara, Y.; and McNamara, D. S. 2007. A coh-metrix analysis of discourse variation in the texts of japanese, american, and british scientists. *Foreign Languages for Specific Purposes* 6:46–77.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Montavon, G.; Samek, W.; and Müller, K.-R. 2017. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*.

Nation, I., and Heatley, A. 1996. Vocabprofile, word and range: Programs for processing text. *LALS, Victoria University of Wellington*.

Page, E. B. 1994. Computer grading of student prose, using modern concepts and software. *The Journal of experimental education* 62(2):127–142.

Perelman, L. 2014. When "the state of the art" is counting words. *Assessing Writing* 21:104–111.

Perfetti, C. A. 1985. *Reading ability.* Oxford University Press.

Powers, D. E.; Burstein, J. C.; Chodorow, M.; Fowles, M. E.; and Kukich, K. 2000. Comparing the validity of automated and human essay scoring. *ETS Research Report Series* 2000(2).

Ramachandran, L.; Cheng, J.; and Foltz, P. 2015. Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 97–106.

Rashotte, C. A., and Torgesen, J. K. 1985. Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly* 180–188.

Rayner, K. P. 1989. A.(1989). the psychology of reading englewood cliffs.

Reid, J. 1990. Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. *Second language writing: Research insights for the classroom* 191–210.

Reppen, R. 1995. Variation in elementary student language: A multi-dimensional perspective.

Riordan, B.; Horbach, A.; Cahill, A.; Zesch, T.; and Lee, C. M. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 159–168.

Somasundaran, S.; Burstein, J.; and Chodorow, M. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 950–961.

Templin, M. C. 1957. Certain language skills in children; their development and interrelationships.

Valenti, S.; Neri, F.; and Cucchiarelli, A. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research* 2:319–330.

Xi, X. 2010. Automated scoring and feedback systems: Where are we and where are we heading?

Yang, Y.; Buckendahl, C. W.; Juszkiewicz, P. J.; and Bhola, D. S. 2002. A review of strategies for validating computer-automated scoring. *Applied Measurement in Education* 15(4):391–412.