ESAS: Towards Practical and Explainable Short Answer Scoring (Student Abstract)

Palak Goenka,[*1] **Mehak Piplani,**[*2] **Ramit Sawhney,**[3] **Puneet Mathur,**[4] **Rajiv Ratn Shah**[2]
[1]Indian Institute of Technology, Roorkee, [2]MIDAS Lab, IIITD
[3]Netaji Subhas Institute of Technology, [4]University of Maryland, College Park
goenkapalak11@gmail.com, 201551072@iiitvadodara.ac.in, ramits.co@nsit.net.in,
puneetm@cs.umd.edu, rajivratn@iiitd.ac.in

Abstract

Motivated by the mandate to design and deploy a practical, real-world educational tool for grading, we extensively explore linguistic patterns for Short Answer Scoring (SAS) as well as authorship feedback. We approach the SAS task via a multipronged approach that employs linguistic context features for capturing domain-specific knowledge while emphasizing on domain agnostic grading and detailed feedback via an ensemble of explainable statistical models. Our methodology quantitatively supersedes multiple automatic short answer scoring systems.

Introduction

Assessment of the acquired knowledge is one of the most crucial aspects of the learning process. Automating this process may tremendously alleviate the quality of instruction, helping teachers to shift their focus from tedious evaluation tasks to imparting knowledge. Moreover, the study by (Prendergast and Topel 1993) highlights the influence of favoritism and the emotional mindset on the assessment procedure. Apart from these human biases, problems occur in online learning platforms due to the shortage of time, money, and the number of instructors in proportion to the number of students. So, to overcome all these obstacles in the path of providing fair education, ESAS, an automated short answer scoring practical and explainable domain agnostic tool has been developed to assist teachers.

The main highlight of ESAS is the extensive feedback provided along with suggestions to the students in a uniform unbaised fashion which sets it apart from the previous work (Kumar et al. 2019) (Riordan et al. 2017). ESAS utilizes Natural language Processing techniques (Ramachandran, Cheng, and Foltz 2015) and a few novel semantic overlap features designed to capture domain-specific knowledge and student's response similarity with sample responses. We also emphasize on building an end-to-end pipeline to handle all the question from the Automated Student Assessment Prize (ASAP) dataset[1] together instead of an individual model for all questions (Kumar et al. 2019) (Riordan et al. 2017).

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

Problem Formulation

Let $Q = \{q_1, q_2, ..., q_n\}$ represent the set of n question prompts with n = 10, and $S = \{s_1, s_2, ..., s_m\}$ be the set of students who answer these question prompts. The set $A = \{a_1, a_2, ..., a_m\}$ represents the set of m responses for a particular question given by corresponding m students such that a unique one-to-one mapping exists between Q, A and S. The sample response (gold standard) for a question prompt is taken to be the concatenation of rubric (if provided) and top three scored responses. Lastly, we define a set of grades $G = \{g_1, g_2, ...g_m\}$ assigned to each response where $g_i \in [0, 1, 2, 3]$. We treat the problem of grading these answer responses as a multi-class classification problem where we assign a grade g to a given question prompt and answer response pair (q_i, a_i).

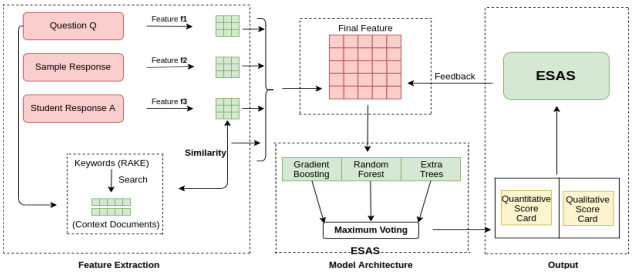

Figure 1: ESAS pipeline

Methodology

Feature Extraction

Mainly five categories of features are incorporated: Lexical (number of words/sentences, average sentence length, spell error), Readability, Text Cohesion, Syntactic (POS-tags count, depth of tree), and Semantic Overlap. The features proposed for the semantic overlap category are summarized below.

- **Context feature:** Context-specific external information sources like Wikipedia have been utilized via extracting a set of domain-specific words using RAKE[2] and Pager-

[2]https://pypi.org/project/rake-nltk/

Approach	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8	Set9	Set10	Avg	Overall
(Riordan et al. 2017)	0.795	0.718	0.684	0.700	0.830	0.790	0.648	0.554	0.777	0.735	0.723	NA
(Ramachandran et al. 2015)	0.86	0.78	0.66	0.70	0.84*	0.88	0.66	0.63	0.84	0.79	0.78	NA
(Kumar et al. 2019)	**0.872***	**0.824***	0.745	**0.743**	**0.845**	**0.858**	0.715	0.624	0.843	0.832	0.791	NA
ESAS	0.853*	0.788	**0.783**	0.676	0.719*	0.736	**0.846***	**0.880***	**0.845***	**0.853**	**0.80***	**0.853***

Table 1: Comparative study, where overall refers to combined model approach. * indicates statistically significant ($p < 0.05$) compared to (Riordan et al. 2017) using Wilcoxon's test.

ank using the Google API[3]. The context vector $\vec{C}(q,a)$ thus obtained is defined in Equation 1.

$$\begin{gathered}\vec{C}(q,a) = \{\, cosine(\gamma_{iq}, \gamma_a) : i \in [1,10]\} \\ \gamma_{iq} : Doc2Vec(d_i) : d_i \rightarrow D_q, \gamma_a : Doc2Vec(a) \\ D_q = \{\, Pagerank_i(keyphrase_q) : i \in [1,10]\} \\ keyphrase_q = Rake(q)\end{gathered} \tag{1}$$

- **POS-Tags Overlap:** To score a candidate response, a mapping of context has been established from the sample response with the help of POS (see Equation 2).

$$\begin{gathered}tag\_score(q,a) = \frac{words\_tag_{aq} \cap max\_count\_tag_q}{len(max\_count\_tag_q)} \\ words\_tag_{aq} = \{Word \iff POS(Word) = tag \\ \forall Word \in A_{iq}\}, tag \in [noun, verb, adj, adv] \\ max\_count\_tag_q = \{Max(Counter(words\_tag\_sr_q)\} \\ words\_tag\_sr_q = \{Word \iff POS(Word) = tag \\ \forall Word \in High\_score\_set_q\} \\ High\_score\_set_q = \{A_{iq}\} : s_{iq} = max(S_q), \\ q \in [1,10]\end{gathered} \tag{2}$$

- **Concept Overlap:** This feature determines the score to a candidate response based on its semantic relatedness with sample response using vector-based sentence representations like BERT, Word2Vec, Doc2Vec (see Equation 3).

$$\begin{gathered}Score(q,a) = \sum Cosine(\gamma_{iq}, \gamma_a) \\ i \in [0, len(sample\_response_q)] \\ \gamma_{iq} : Vec(h_i) : h_i \rightarrow sample\_response_q \\ \gamma_a : Vec(a)\end{gathered} \tag{3}$$

Model Architecture

The model architecture, **ESAS (Ensemble Short Answer Scoring)** is a max voting ensemble of the following statistical machine learning models: Random Forest, Gradient Boosting Classifier and Extra-Trees Classifier as demonstrated in Figure 1. The chief reasons for choosing these tree-based classifiers are feature interpretability, robustness due to ensembling of weak learners and reduction in variance due to cut-point randomization.

Feedback: Qualitative Analysis

We designed the following evaluation criterion: Lexical correctness, Readability score, Creativity, and Relevance for evaluating the qualitative score. For indicating **lexical correctness**, the feedback includes the count of spelling errors (via pyspellchecker[4]) and a score for deviation of spelling errors from the correct words. The **readability score** is indicated by incorporating the lengthy words (words having the number of syllables >2) in the response and their synonym replacement using NLTK Wordnet[5]. Lexical diversity and Connective incidence indicates **creativity** in a response indicated via count of connectives (tool used: Stanford CoreNLP[6]). Lastly, determination of **relevance**, based on application of domain knowledge, portrayed through the least and most contributing sentences calculated via context feature.

Results

The results are formulated using Quadratic Weighted Kappa (Brenner and Kliebsch 1996), an assessment metric to determine the agreement between the evaluations anticipated via ESAS and the human grader. Table 1 shows how ESAS outperforms the state of the art results by 7.8% by employing the combined model approach (training on all questions prompts together). Even though the features were not tuned individually ESAS also performed exceptionally well on five out of ten comprehension type questions, with a maximum of 41% improvement on the question *prompt-8* compared to the current best model by (Kumar et al. 2019).

Conclusion

In this paper, we have presented an approach for SAS that applies ensembling method over domain agnostic as well as context-specific features. Furthermore, we focused on building a feedback-oriented end-to-end deployable solution by training on cumulative dataset instead of singular question prompts, which beats the state-of-the-art results.

References

Brenner, H., and Kliebsch, U. 1996. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology (Cambridge, Mass.)* 7(2):199–202.

Kumar, Y.; Aggarwal, S.; Mahata, D.; Shah, R. R.; Kumaraguru, P.; and Zimmermann, R. 2019. Get it scored using autosas-an automated system for scoring short answers.

Prendergast, C., and Topel, R. 1993. Discretion and bias in performance evaluation. *European Economic Review* 37(2-3):355–365.

Ramachandran, L.; Cheng, J.; and Foltz, P. W. 2015. Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. In *BEA@NAACL-HLT*.

Riordan, B.; Horbach, A.; Cahill, A.; Zesch, T.; and Lee, C. M. 2017. Investigating neural architectures for short answer scoring. 159–168.

[3]https://pypi.org/project/google/

[4]https://pypi.org/project/pyspellchecker/

[5]http://www.nltk.org/howto/wordnet.html

[6]https://stanfordnlp.github.io/CoreNLP/