Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

An AES System to Assist Teachers in Grading Language Proficiency and Domain Accuracy Using LSTM Networks

Aditya SAHANI[a*], Forum PATEL[a], Shivani MEHTA[a], Dr Rekha RAMESH[a] & Dr Ramkumar RAJENDRAN[b]
[a]Computer Engineering Department, University of Mumbai, India
[b]Education Technology Department, IIT Bombay, India
**aditya.sahani@sakec.ac.in*

Abstract: Automated Essay Scoring (AES) is a task of automatically grading the students' answers to subjective or essay type questions. AES is an area where assessing the answers rationally is very important. Assessing these subjective answers has always been a challenging process concerning reliability and effort. In such times, where the entire education system has shifted to being online, it becomes necessary to develop a system that assesses students based on their subjective answers. However, the existing AES system primarily focuses on assessing essays on a single dimension that is either grading domain accuracy or grading the language correctness of the answers. Moreover, there are few AES systems to grade student's responses in the computer science domain. To address these gaps, we propose an AES system to grade the subjective answers of students from the computer science domain. The proposed system grades the student's responses in two dimensions, namely domain accuracy, and language proficiency. In order to test the system, we collected data from 200 students and manually labeled them for domain accuracy and language proficiency. The system graded the student's responses automatically with domain accuracy of 89.47 percent and language proficiency of 84.79 percent.

Keywords: Parallel networks, domain accuracy, language proficiency, LSTM, Computer Science, Word Embedding.

1. Introduction

Assessment has been an integral part of the teaching and learning process. Essay type questions are always part of assessments (Dong, Zhang, & Yang, 2017) and such questions offer students an opportunity to demonstrate knowledge, skills, and abilities in a variety of ways such as writing skills and formulating arguments supported with reasoning and evidence (Valenti, Neri, & Cucchiarelli, 2003). However, because of the large number of student participation in assessments, manual evaluation and grading of answers to these essay type questions is a challenging task for the teachers. Manual evaluation by multiple teachers may also introduce inconsistent or erroneous grading because of mutual disagreements (Dasgupta, Naskar, Dey, & Saha, 2018).

To address this challenge automated essay scoring (AES) has been explored for over 50 years as a part of natural language processing. The existing works in AES were primarily developed to assess the essays on a single dimension such as either assessing domain knowledge or language proficiency. In order to provide detailed feedback to students and measure the outcome of all the objectives of subjective assessments, it is necessary to grade them on both domain accuracy and language proficiency. There exist few systems to detect both domain accuracy and language proficiency. However, they provide the final score as a combination/average of the domain and language accuracy. To address this gap, we propose a system that grades essays on mainly two parameters: domain accuracy and language proficiency.

The related systems with AES implement word embedding, transfer learning and feature engineering to score the answers (Hussein, Hassan, & Nassef, 2019) (Hussein, A., & Nassef, 2020). Many systems with respect to AES have been using LSTMs and CNNs to find meaning behind text and score it (Taghipour & Ng, 2016)(Hussein, A., & Nassef, 2020) (Cao, Jin, Wan, & Yu, 2020). A Siamese Bidirectional LSTM based Regression was designed for grading the answers of the Computer Science domain (Prabhudesai, Arya, Duong, 2019).

A few systems work in the CS domain (Ndukwe, Amadi, Nkomo, & Daniel, 2020) (Hussein, Hassan, & Nassef, 2019) and hence, there is limited availability of dataset from CS domain. Very few systems predict a score on English language proficiency (Ndukwe, Amadi, Nkomo, & Daniel, 2020) (Zhao, Zhang, Xiong, Botelho, & Heffernan, 2017). Moreover, most of the papers are improving the performance over the existing ASAP dataset.

Our proposed system uses a parallel Long Short Term Memory (LSTM) network, one for checking the domain accuracy of the answer and another for checking the language proficiency. We selected LSTM networks because they are widely used in recent related works and it utilizes the long-distance dependencies in the answers (Taghipour & Ng, 2016). From the existing AES systems, we found that there is no dataset in the CS domain that can be used for scoring. To address scarcity of dataset in CS domain, we created a dataset. The dataset is called the UM dataset and has answers Data Structures (DS). 200 students were given four essay-type questions as part of their assignments. Students' answers were graded by the instructors for domain and language accuracy and were given a score using a rubric designed by subject experts.

To train and test our system, a corpus of words was developed related to the Data structures domain using three standard textbooks from the university curriculum. The words in the corpus are then converted into word embedding using the Gensim Library, the embedding size is set to 300 dimensions The proposed AES system graded the student's response with a domain accuracy of 89.47 percent, language proficiency of 84.7 percent.

Since the UM dataset is small, to test the performance of our system on a large dataset we searched for an existing dataset in the DS course. We found a database from the University of North Texas (UT dataset). The dataset had short answers from the Data Structures domain. However, the UT dataset was associated with only domain scores and the rubrics used for grading were unavailable. Hence, we graded the UT dataset for language proficiency by experts using the rubrics that we developed. The proposed system graded the answers in UT dataset with a domain accuracy of 89.43 percent and language proficiency of 89.92 percent. The results indicate that our proposed dual LSTM network can perform well on both small and large datasets.

The rest of the paper is divided as follows. Section 2 discusses our proposed system. Section 3 describes the detailed implementation of our system and the results of the system are provided in section 4. Following this, section 5 throws light on the discussion and conclusion of the system.

2. Proposed Methodology

We use a parallel LSTM networks for automated grading of students' subjective answers to predict a separate individual score for domain accuracy as well as language proficiency for these answers. Figure 1 shows the block diagram of the proposed system. This trained word2vec model is stored and then fed to the LSTM model where the domain accuracy and language proficiency score is calculated.

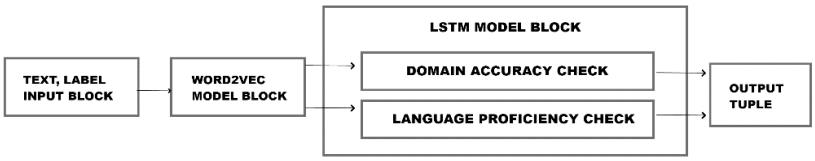

Figure 1. System Flow Diagram.

3. Implementation

The model takes in the answers of the data structures course graded on the domain accuracy and language proficiency as its input. The input fed is the word embeddings. The system uses a parallel LSTM network to score the answers. After successfully training and testing the model, the scores are displayed in the form of a tuple. The functions of each block are described in the following subsections.

The input block of the system takes in the answers of the data structures course that are graded by experts on both domain accuracy and language proficiency. The block performs the data cleaning by removing the nulls and performs exploratory data analysis on the subjective answers.. The dataset and pre-processing are explained in detail below.

We created our data set called the UM (University of Mumbai) dataset, consisting of descriptive answers for the data structures course offered in the second-year engineering curriculum. Students were given a set of four questions to solve in one hour as part of their assignment. The questions are shown in Table 1. 200 students participated in the assignment. These answers were collected online.

Two domain experts who taught the course three times graded the answers on domain accuracy. To have consistency in grading, a rubric was designed as shown in Table 3 and the inter-rater reliability achieved with Cohen kappa was 0.82. Table 2 talks about the language rubric that has been used to score the answers on language proficiency by an Expert.

Table 1. *Assignment questions for creating UM dataset*

Question no	Questions on the quiz
1	Consider a CPU scheduling task, where in each process has a process execution time and priority assigned to it. Processes are stored in the order of their priorities, that is the process having high CPU time
2	Given an arithmetic expression, find all possible outcomes of this expression. Different outcomes are evaluated by putting brackets at different places. We may assume that the numbers are single dimension.
3	Suppose we want to implement a navigation option in a web browser. Now we have two options for this particular purpose, a circular queue array based and doubly linked list. compare both with each other
4	What kind of data structure would you recommend to store the large amount of data in a computer system? Also how it will be better than other available options

Table 2. *Rubrics for accessing Language Proficiency*

Points	Level of Achievement			
	4-Expert	3-Accomplished	2-Capable	1-Beginner
Quality of writing	Piece was written in an extraordinary style and voice	Piece was written in an interesting style and voice	Piece had little style or voice	Piece had no style or voice
Sentence Structure	Sentences are coherent	Sentences are mostly coherent	Sentences are somewhat coherent	Sentences are not coherent
Understanding	Writing shows strong understanding	Writing shows a clear understanding	Writing shows adequate understanding	Writing shows little understanding
Spelling errors	Virtually no spelling errors	Few spelling errors	A number of spelling errors	So many spelling, errors
Punctuation and grammatical errors	Virtually no punctuation or grammatical errors	Few punctuation errors, minor grammatical errors	A number of punctuation or grammatical errors	So many punctuation and grammatical errors that it interferes with the meaning

We created a corpus that consisted of words from the domain from three standard books from the university curriculum. We use the gensim library to create the word2vec model. The model is trained with an embedding size of 300 and uses the skip-gram model (Agarap, 2019). After this trained model is stored, when we input the answers the model checks for new words and adds them to the word2vec model corpus.

Table 3. *Rubrics for accessing Domain Accuracy*

Points	Level of Achievement			
	4-Expert	3-Accomplished	2-Capable	1-Beginner
Completeness	All the concepts are covered	Few of the concepts are missing	Most of the concepts are missing	None of the concepts are covered
Correctness	Virtually no incorrect facts	Few incorrect facts	A number of incorrect facts	So many incorrect facts that it interferes with the meaning.
Knowledge Construction	Strong connections among the concepts and none of them are incorrect	Few connections are missing or incorrect	Most of the connections are missing or incorrect	No attempt made to connect the concepts
Understanding of topic	Writing shows strong understanding of topic	Writing shows a clear understanding of topic	Writing shows adequate understanding of topic	Writing shows little understanding of topic
Explanation with an example	Explained the topic with a correct example	Example is correct but explanation did not connect to the given topic	Attempted to explain the topic with a somewhat correct example	Did not give any example

The LSTM model consists of five layers. The recurrent dropout throughout the model is set to 40 percent and dropout is set to 50 percent. The first layer is the embedding layer, second LSTM layer consists of 300 neurons, input shape is set to 1,300 and the return sequence parameter is set to true. The third layer is an LSTM layer with 64 neurons, fourth layer is a dropout layer and last is a dense layer to take output.

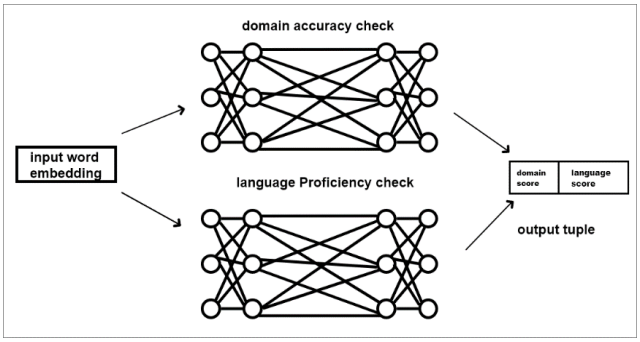

Figure 2. Parallel LSTM Network Model.

The activation function used is a rectified linear (ReLU) unit throughout the model (Agarap, 2019). We used mean square error, mean absolute error and cohens kappa metrics to evaluate the model. The input is optimized using the Adam optimizer (Kingma & Ba, 2017). The system uses a parallel LSTM network that work to score the answers in Figure 2.

The output block of the system consists of the results of the two parallel LSTM networks into a single tuple {Score 1, Score 2}. The first value in the tuple will be the score of domain accuracy and the second will be the language proficiency.

4. Results

The parallel LSTM model is first tested on the UM dataset. This dataset is fed to the model for both domain accuracy and language proficiency. The model is trained for 10 folds of cross-validation and

50 epochs each fold. Table 4 indicates the domain accuracy confusion matrix when the model is tested on the UM dataset and table 5 indicates the language proficiency. After successfully training the model, its performance is summarized as shown in Table 4.

After evaluation of the model, it produces a mean square error (mse) of 0.4 for domain accuracy and 0.2 for language proficiency. A quadratic weighted kappa (qwk) of 76.5 and 86.2 for domain accuracy and language proficiency, a precision score of 68.2 and 80, a recall score of 86.7 and 83, the overall accuracy of the model is 89.47 for domain accuracy and 84.7 for language proficiency.

Table 4. *The Confusion matrix for the language proficiency & domain accuracy of the dataset UM*

	Language Proficiency			
	Score 1	Score 2	Score 3	Score 4
Score 1	4	1	0	0
Score 2	1	19	7	0
Score 3	0	3	99	25
Score 4	0	0	10	140

	Domain Accuracy			
	Score 1	Score 2	Score 3	Score 4
Score 1	2	4	1	8
Score 2	0	37	2	1
Score 3	0	6	44	12
Score 4	0	3	9	308

The UM dataset is smaller in size, due to which we also tested it on the UT dataset to check the generalizability of our model. The UT dataset created by the University of North Texas consists of questions from the Data Structures course given as part of a course assignment (Mohler, Bunescu, & Mihalcea, 2011) (Mohler & Mihalcea, 2009). The dataset in all had 1400 answers rated on domain accuracy in the range of 1(Beginner) to 4(Advanced). The UT dataset does not have a language score assigned to the answers. Hence, we had an expert (doctorate in English literature) who graded the answers on language proficiency using a rubric shown in Table 2.The model runs for 10 folds of cross-validation with 50 epochs each fold. Table 5 indicates the domain accuracy confusion matrix when the model is tested on the UT dataset and table 8 indicates the language proficiency confusion matrix.

Table 5. *The Confusion matrix for the language proficiency & domain accuracy of the UT dataset*

	Language Proficiency			
	Score 1	Score 2	Score 3	Score 4
Score 1	29	8	0	0
Score 2	8	443	34	6
Score 3	0	30	424	35
Score 4	0	7	48	522

	Domain Accuracy			
	Score 1	Score 2	Score 3	Score 4
Score 1	117	19	6	30
Score 2	7	56	15	16
Score 3	3	11	93	45
Score 4	4	13	11	1110

After evaluation of the model, it produces a mean square error (mse) of 0.1 for domain accuracy and 0.3 for language proficiency. A quadratic weighted kappa (qwk) of 81 and 92 for domain accuracy and language proficiency, a precision score of 70.7 and 80, a recall score of 73.3 and 86.5, the overall accuracy of the model is 89.43 for domain accuracy and 89.92 language proficiency.

5. Discussion and Conclusion

We developed a system that uses parallel LSTM networks to grade students' subjective answers. Due to the limited availability of data in the CS domain, we created our own UM dataset by conducting an assessment consisting of essay type questions for the second year Engineering students from University of Mumbai in the data structures course. The answers are rated by experts for domain accuracy and language proficiency. The expert raters used two different rubrics to grade these answers and the grades ranged between 1(Beginner) to 4(Expert) as explained previously in section 4.1.1.

The system was also tested on the bigger dataset (UT) which we got from the University of North Texas (Mohler, Bunescu, & Mihalcea, 2011) (Mohler & Mihalcea, 2009) to check how the model would generalize for a large dataset. The results show that it performs well across both datasets for domain accuracy and language proficiency. We plan to share UM data publicly so that other researchers can use and build an AES system around it.

Currently, there is very limited availability of datasets in the CS domain. The UM dataset consists of only subtopics of the data structures course whereas the UT dataset has answers from all the topics of the course. The results of the UM dataset, when evaluated and analyzed, portray that the dataset might be unbalanced. The system performance can be improved if more data is available for training. The limitation of the system is that it can grade the answers only with integer values. Another limitation of the system is that it cannot provide feedback to the user on why the answer has been given a particular score.

In future, this system could be generalized to more domains to include other courses of CS Domain. We can also dive deeper into NLP and try including various text features that can be used with these word vectors to generate a more accurate system. The system could incorporate a self-explanatory feedback mechanism based on the rubrics and help the students in self-learning and improvement.

References

Agarap, A.0 F. (2019, February 07). Deep Learning using Rectified Linear Units (ReLU). https://arxiv.org/abs/1803.08375

Cao, Y., Jin, H., Wan, X., & Yu, Z. (2020). Domain-Adaptive Neural Automated Essay Scoring. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. doi:10.1145/3397271.3401037

Dasgupta, T., Naskar, A., Dey, L., & Saha, R. (2018). Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring. Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. doi:10.18653/v1/w18-3713

Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). doi:10.18653/v1/k17-1017

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. PeerJ Computer Science, 5. doi:10.7717/peerj-cs.208

Hussein, M. A., A., H., & Nassef, M. (2020). A Trait-based Deep Learning Automated Essay Scoring System with Adaptive Feedback. International Journal of Advanced Computer Science and Applications, 11(5). doi:10.14569/ijacsa.2020.0110538

Kingma, D. P., & Ba, J. (2017, January 30). Adam: A Method for Stochastic Optimization. Retrieved from https://arxiv.org/abs/1412.6980

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL 09. doi:10.3115/1609067.1609130

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. Retrieved from https://aclanthology.org/P11-1076/

Ndukwe, I. G., Amadi, C. E., Nkomo, L. M., & Daniel, B. K. (2020). Automatic Grading System Using Sentence-BERT Network. Lecture Notes in Computer Science Artificial Intelligence in Education, 224-227. doi:10.1007/978-3-030-52240-7_41

Prabhudesai, A., & Duong, T. N. (2019). Automatic Short Answer Grading using Siamese Bidirectional LSTM Based Regression. 2019 IEEE International Conference on Engineering, Technology and Education (TALE). doi:10.1109/tale48000.2019.9226026

Taghipour, K., & Ng, H. T. (2016). A Neural Approach to Automated Essay Scoring. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. doi:10.18653/v1/d16-1193

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. Journal of Information Technology Education: Research, 2, 319-330. doi:10.28945/331

Zhao, S., Zhang, Y., Xiong, X., Botelho, A., & Heffernan, N. (2017). A Memory-Augmented Neural Model for Automated Grading. Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale. doi:10.1145/3051457.3053982