UR-102

Auto Grader for Unstructured Open-Ended Questions using Long Short-Term Memory and Singular Value Decomposition

Author: Anna Furrow
Advisors: Dr. Mohammed Aledhari

Abstract

Grading an open-ended exam or paper takes a large amount of time, and to do that well across many classes is infeasible for professors. This leads to students taking exams that do not accurately detail their knowledge on a subject. Thus, the development of auto-grading systems can help mitigate this workload and time commitment. It is common that solely natural language processing techniques are used, the source material is learned and then compared to an answer key. While this is how professors score, only using natural language processing to derive a score has its limitations. This study aims to use natural language processing techniques, like word vectors and long short-term memory to understand the meaning of a passage. This study also aims to use recommender systems, namely collaborative filtering techniques, to determine an appropriate score for an exam. This system achieves an average Cohen's kappa score of 0.6482. The recommender system is also able to give a score similar, if not the same, as human graded responses. Through fine tuning and more training, the auto-grading system can be further explored for grading student responses.

Introduction

Examinations are an important aspect of education and educational institutions. To properly understand the knowledge a student has gained, an exam must be well-formulated. Those that are best for testing students are of open-ended format. Students will answer in a couple of sentences up to multiple paragraphs. However, on the professor's end this is very time consuming. As more young people pursue education, it is becoming increasingly important to develop written communication skills [2]. This creates a problem for professors - do they test students on their knowledge and help them develop necessary skills or do they create an exam that lightens their workload [2]?

Auto-grader systems are a natural language processing (NLP) problem that has been explored across many types of essays and writing pieces. Question scoring is divided into two main types: essay and short-answer [3]. Each type has specific metrics that are used for scoring. An essay is graded on the quality of the content written while a short answer is graded on accuracy [3]. For open-ended questions that fall in between an essay and a short-answer, this allows us to explore a combination of scoring for accuracy and writing quality. For this problem, the middle ground between a short-answer and an essay is explored. The answer to an open-ended question can range anywhere from a few sentences to a couple paragraphs.

Recommender systems are also an integral part of this problem because they are the way a score is given [1]. Recommender systems are typically used for business applications and recommend products to consumers. The systems will take in user preferences and give them a recommendation based on similarity [1]. These systems give similarity scores which can be used for essay scoring. Based on the similarity between the student response and the answer key, the recommender system can give a recommendation of the score that it believes the student response should receive

Research Question(s)

- Does the use of recommender systems enhance auto-grading systems?
- Is it possible for an auto-grader system's results to be comparable, if not better, than that of a professor?

Materials and Methods

- Python libraries: Tensorflow, Keras, Pandas, Surprise
- Dataset: Hewlett Foundation Essay Scoring datasets
  - 20,000+ essays written by high school students
  - Persuasive/Expository/Narrative over 8 sets

Results

Data Preprocessing

- Essay data is preprocessed by removing the aspects of language that do not contribute to meaning of the sentence. These are then converted to word embeddings (numerical representations of words).
- GloVe vectors are also used to help the model learn the representations of words

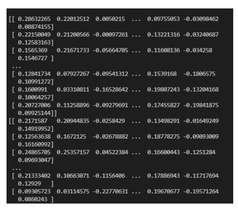

Figure 1: Word Embeddings after data preprocessing

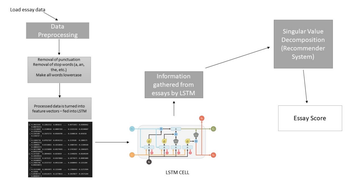

Figure 2: Flow of system architecture

Long Short-Term Memory (LSTM)

- LSTM is used to process the language and learn a base representation of the essays. The LSTM is trained using 5-fold cross validation.
- Evaluated using mean absolute error (MSE)
- Cohen's Kappa score gives an idea of how well the essays are written

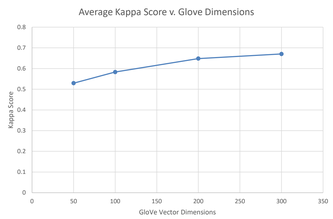

Figure 3: GloVe vector dimensions and how they affect the Kappa score while training the LSTM

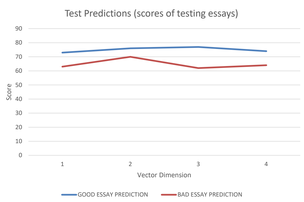

Figure 4: GloVe vector dimensions and how they affect score prediction using solely LSTM

Recommender System

- Singular value decomposition (SVD) takes what the LSTM has learned and then assigns a score to the essays.
- Root mean squared error and mean absolute error are used to evaluate the recommender system.

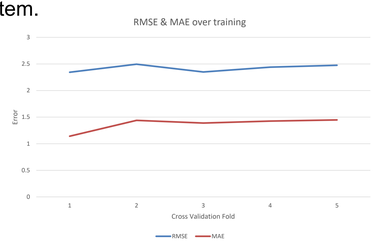

Figure 5: Change of RMSE and MAE over each fold of training SVD algorithm

Conclusions

The combination of LSTM and SVD show promise in creating a robust auto-grading system. LSTM is a state-of-the-art model used commonly for natural language processing tasks and was a natural choice for this problem. The main issue to focus on is improving the performance of recommender systems within the scope of this problem. The recommender system underperforms, the margin of error is quite wide when looking at the prediction in relation to the ground truth scores of the essay. In this current state of the solution, the recommender system does not enhance the auto-grader system, rather it hinders it and makes it much more complex.

This complexity and hinderance can also be attributed to the dataset used. While a variety of data can help create a robust and well-equipped machine learning model, in this case it is detrimental to the performance. To improve this system further, college/university level essays should be used.

Acknowledgments

Dr. Mohammed Aledhari – mentor and professor

Contact Information

afurrow@students.kennesaw.edu | anna.furrow@gmail.com

← Scan for my LinkedIn!

References

1] M. Deschenes, "Recommender systems to support learners' agency in a learning context: a systematic review," International Journal of Educational Technology in Higher Education, vol. 17, no. 1, 2020.
[2] B. L. Sevcikova, "Human versus automated essay scoring: A critical review," Arab World English Journal, vol. 9, no. 2, pp. 157– 174, 2018.
[3] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. M. Lee, "Investigating neural architectures for short answer scoring," Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 2017.

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
Department of Computer Science