Using Active Learning for Assisted Short Answer Grading

Jeeveswaran Kishaan kishaan.kishaan@smail.inf.h-brs.de
Mohandass Muthuraja mohandass.muthuraja@smail.inf.h-brs.de
Deebul Nair deebul.nair@h-brs.de
Paul G. Plöger paul.ploeger@h-brs.de
Autonomous Systems Group
Hochschule Bonn-Rhein-Sieg
Sankt Augustin, Germany

Abstract

Efficient and comprehensive assessment of students knowledge is an imperative task in any learning process. Short answer grading is one of the most successful methods in assessing the knowledge of students. Many supervised learning and deep learning approaches have been used to automate the task of short answer grading in the past. We investigate why assistive grading with active learning would be the next logical step in this task as there is no absolute ground truth answer for any question and the task is very subjective in nature. We present a fast and easy method to harness the power of active learning and natural language processing in assisting the task of grading short answer questions. A web-based GUI is designed and implemented to incorporate an interactive short answer grading system. The experiments show that active learning saves the time and effort of graders in assessment and reaches the performance of supervised learning with less amount of graded answers for training.

Keywords: Active Learning, Automatic Short Answer Grading

1. Introduction

Assessing the knowledge of students is one of the most important phases of the learning process Mohler et al. (2011). Different forms of assessments that exist today include multiple choice questions, fill-in-the-blanks, essay questions, and short answer questions. Prior works have shown that multiple choice questions and fill-in-the-blanks fail to capture the vital aspects of the acquired knowledge such as reasoning and self-explanation Wang et al. (2008). In contrast, questions which require the students to construct responses in natural language have been found to be more effective in assessing their grasp on the subject matter Roy et al. (2016). Essay questions and short answer questions belong to this category. This work is more concerned about short answer questions where students construct answers in natural language.

Assessing the students' responses in time and giving quick feedback's enables the students to realize the mistakes and learn from them. Limited availability of teachers, online learning platforms, and individual or group study sessions done outside classrooms necessitated quick and efficient assessment of free text responses Mohler et al. (2011). Computer assisted assessment / automatic grading evolved as a solution to this problem and a lot

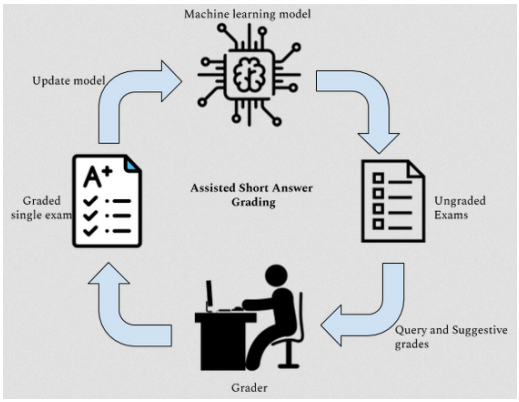


Figure 1: Workflow of active learning in assisted short answer grading. Image adapted from Burrows et al. (2015)

of research has been done on automating the grading short answer responses Leacock and Chodorow (2003); Pulman and Sukkarieh (2005); Mohler and Mihalcea (2009).

Automatic short answer grading essentially deals with using computational methods to compute the grades for students' answers. A typical work flow of an automatic short answer grading would include learning a machine learning model to compute the grade based on the features extracted from the students' answers. Though these approaches were able to produce decent results, they suffer from many shortcomings such as; lack of sufficient amount of labeled training data in the domain to learn the models, failure to capture the different wordings/phrasing of the students, and being a passive learner where the models learn the rules once and apply them on new input answers (thus being less robust to new data). Figure 1 illustrates the work flow proposed in this paper which implements active learning for the task of short answer grading. A generic scoring model tries to learn this task of measuring the correctness of each answer continuously with a human in the loop. Active learning seems to be the best choice for this task as it actively queries the human for grades of the samples it is most uncertain of. By actively selecting the data samples to label, it reduces a considerable amount of labeled training samples, thus, alleviating the problem of insufficient labeled data which is prevalent in supervised learning approaches.

We evaluated various active learning strategies as a potential solution to overcome the deficits of automated short answer grading in three different datasets. The best active learning query strategy is selected for the final comparison with the supervised learning approach and finally we conducted the experiment to measure the ease in grading by measuring the number of clicks. A generic interactive model for assessments in different domains is implemented based on the results of experiments performed using a web-based Graphical User Interface (GUI).

The main contributions of this paper are;

- Application of active learning to improve the grading experience.
- Comparison of active learning query strategies useful for short answer grading.
- New dataset for short answer grading.

2. Related Work

Previous works in automating the task of short answer grading mostly comprised of either matching predefined templates over student answers or measuring the similarity of student answers to that of the teacher's answer and compute the grade based on it. The history of works on this task could be categorized into four approaches according to Burrows et al. (2015).

- Works done by Callear et al. (2001) and Leacock and Chodorow (2003) belong to a paradigm called concept mapping where the student answers are segmented into different concepts and the score is given based on the presence of concepts in the reference answer.
- Bachman et al. (2002); Pulman and Sukkarieh (2005); Ramachandran et al. (2015); Meurers et al. (2011) followed a different approach called information extraction where the systems extract a pattern from the student and the reference answers.
- Mohler et al. (2011); Sultan et al. (2016); Zesch et al. (2015) applied machine learning based approaches by to training a model based on the features extracted using natural language processing techniques.
- Kumar et al. (2017); Liang et al. (2018) have applied deep learning approaches to improve the accuracy in the task of automatic short answer and essay scoring.

All the above methods have the goal of completely automating the grading procedure and no one has achieved a 100% accuracy. A solution is to use assisted grading, in which we solve the problems of data collection and deploying a machine learning model. Active learning is a well established branch in machine learning for doing such assisted learning tasks.

3. Active Learning

Active learning belongs to a special case of semi-supervised learning algorithm where the learner is allowed to query the user to get the labels for data points which will help the learner to perform better. When the learner is allowed to choose the data points for learning, the performance of the learner is better with less labeled data. Supervised learning algorithm performs well when the model is trained with a large number of labeled instances. But the labeled instances are expensive, time-consuming, and difficult to obtain. The main aim of the active learning is to improve the accuracy by labeling less number of instances which is determined by the active learning query strategies.

3.1 Active Learning Scenarios

There are three main problem scenario that can occur when the learner queries the user. They are 1) membership query synthesis, 2) stream-based selective sampling and 3) pool-based sampling. All the experiments are done based on pool based sampling as it suits our use case. The pool-based labeling setting consists of two pools of data which include a large set of unlabeled data and a small set of labeled ones. The instances are queried based on the usefulness to evaluate the other instances in the unlabeled pool.

3.2 Query strategy frameworks

Passive learning approach has a large amount of labeled data which are sampled from an underlying distribution is used to train a model. Then the trained model is used for prediction. But in active learning, the learner query the most informative instance or the best instance which is the main difference between active and passive learning. There are various querying strategies to select the most informative query. Out of these uncertainty sampling based methods performed better for our usecase.

3.2.1 Uncertainty Sampling

In uncertainty sampling, the active learner query the data for which the prediction probability is very uncertain to classify into a particular class. The two types of uncertainty sampling are discussed below.

Least confident uncertainty sampling queries the instance for which the learner is not confident about its prediction. The instance that has to be queried using the least confident uncertainty sampling is found using the following equation given by Lewis and Catlett (1994).

$$x^*_{LC} = \underset{x}{\operatorname{argmax}}\, 1 - P_\theta(\hat{y}|x) \tag{1}$$

Here, $P_\theta(\hat{y}|x)$ is the prediction probability of an instance belonging to a most likely class. This helps in finding the least confident instance for which the prediction probability for a particular class is high.

Margin uncertainty sampling chooses instance considering the prediction probability of both the first and second likely class for which the instance belongs. This is in contrast to the least confident uncertainty sampling where it considers only the class for which the instance has maximum prediction probability. Scheffer et al. (2001) compute the margin uncertainty of the instances using,

$$x^*_M = \underset{x}{\operatorname{argmin}}\, P_\theta(\hat{y_1}|x) - P_\theta(\hat{y_2}|x) \tag{2}$$

where, $P_\theta(\hat{y_1}|x)$ and $P_\theta(\hat{y_2}|x)$ were the prediction probability of the first and second likely class of a particular instance respectively. The instance for which the margin uncertainty is minimum is queried.

4. Experimental Evaluation

We conducted various experiments to select the appropriate querying strategy for active learning, to compare the effectiveness of active learning with respect to supervised learning and finally we demonstrate with a GUI how active learning can improve the grading experience.

4.1 Datasets, Features and Classifiers

We used 3 short answer grading datasets from 3 different domains to evaluate the different active learning strategies. The dataset used are Mohler'11 Dataset Mohler et al.

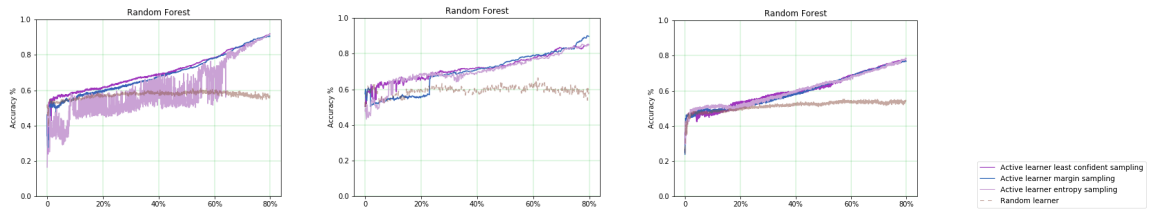
(2011),SemEval-2013 Task 7 dataset Nielsen et al. (2008) and finally a in-house ASAG dataset.

The machine learning model used in this project is implemented with the help of the scikit-learn library Pedregosa et al. (2011). The experiments were done with random forest classifier as it is known for dealing with multiple features which could be correlated, and reduce variance.

4.2 Experiment 1: Comparison of Different Query Strategies

Various active learning query strategies with different machine learning models were evaluated based on accuracy on different datasets. Figure 2 shows the performance of difference query strategies on different datasets. Committee based query strategies were not included in the graph due to their inferior performance when compared to the uncertainty based counterparts.

The experimental results on different datasets and using different query strategies show that there is no single active learning setting that works well for all datasets. Least confident based uncertainty sampling worked well in random forest classifier when features from Sultan et al. (2016) were used.



(a) Mohler11 dataset (b) AGAG dataset (c) SemEval dataset (d)

Figure 2: Performance of different query strategies on different datasets. The x-axis represents the percentage of labelled dataset used for training

4.3 Experiment 2: Efficiency of active learning based grading system

This experiment is conducted to measure the efficiency of using an interactive short answer grading system to grade the assignments with active learning in the background 2. We conducted the experiments in 2 parts, one without active learning support and other with active learning support. The number of clicks required to grade all the answers in both the settings were recorded. The experiments were done on all three datasets. All the students' answers are displayed question-wise in the active learning-free version. In the version which uses active learning, the model is trained on the grades given by the grader for 25% of the answers and then the answers are displayed question-wise along with the predicted score for every answer. The grader has the choice of changing the grade if he thinks that the suggested score by the model is wrong. Random forest classifier with 100 trees and margin based uncertainty sampling are used in the learning process. The number of clicks made in both the learning-free version and learning assisted versions have been tabulated in Table 1.

Datasets	Clicks with active learning	Clicks without active learning
Neural network	338	680
SemEval 2013	3527	5104
Mohler'11	1386	2352

Table 1: Number of clicks required to grade the answers with and without active learning

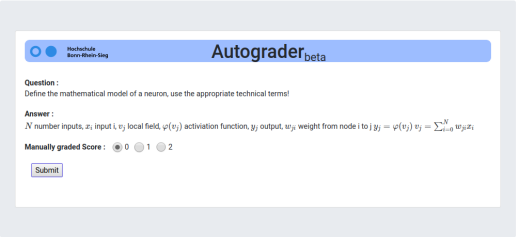


Table 2: Querying stage of active learning in GUI

As can be seen from the table 1, the grading process assisted by active learning massively reduces the effort and time of the grader. In addition, the grader could spend more time in constructing useful feedback's for the students.

5. Limitations and Future Work

Getting the features proposed by Sultan et al. (2016) is time-consuming. This hinders instant deployment of the grading system on a new dataset with the GUI. More time-efficient feature generators would save time. The current workflow requires reference answers for every questions to compute the similarity score and word alignment score. The quality of the reference answer has a very high impact on the results of the grading system.Workflows which are agnostic to model answers should be developed. The features work mostly in a syntactic nature where the words are considered individually rather than on a semantic basis where the whole meaning of a sentence is taken into account. Current deep learning methods (for ex. work by Kumar et al. (2017)) seems to work at a more semantic level which if incorporated can further reduce the time required for grading.

6. Conclusions

A workflow for assisted grading using active learning methods is proposed in this work. A detailed evaluation of different active learning strategies on three different datasets is done. Selected active learning strategy achieved similar performance as supervised learning with less training samples, thus reducing the effort of graders. A web-based graphical user interface (GUI) which can be used for the task of grading the answers with the help of active learning is developed and made publicly available [1]. The ASAG dataset along with the features extracted is made publicly available [2].

Based on our experiments and usability of the system, assisted grading of short answer questions is found to be helping in reducing efforts to grade. Thus, it could be seen as a potential solution towards deploying current machine learning and deep learning methods for short answer grading. We used basic feature extraction and machine learning models, which can be replaced by other state of the art approaches for better performance in future.

1. https://github.com/DigiKlausur/AssistedShortAnswerGrading
2. https://github.com/DigiKlausur/ASAG-Dataset

References

Lyle F Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Michael J Pan, Chris Salvador, and Yasuyo Sawaki. A reliable approach to automatic assessment of short answer free responses. *Proceedings of the 19th international conference on Computational linguistics* -, 2:1–4, 2002. doi: 10.3115/1071884.1071907.

Steven Burrows, Iryna Gurevych, and Benno Stein. *The eras and trends of automatic short answer grading*, volume 25. 2015. ISBN 4059301400268. doi: 10.1007/s40593-014-0026-8.

David Callear, Jenny Jerrams-Smith, and Victor Soh. Caa of short non-mcq answers. 2001.

Sachin Kumar, Soumen Chakrabarti, and Shourya Roy. Earth mover's distance pooling over siamese lstms for automatic short answer grading. In *IJCAI*, pages 2046–2052, 2017.

Claudia Leacock and Martin Chodorow. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405, 2003. ISSN 00104817. doi: 10.1023/A:1025779619903.

David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning Proceedings 1994*, pages 148–156. Elsevier, 1994.

Guoxi Liang, Byung-Won On, Dongwon Jeong, Hyun-Chul Kim, and Gyu Choi. Automated essay scoring: A siamese bidirectional lstm neural network architecture. *Symmetry*, 10 (12):682, 2018.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. Evaluating answers to reading comprehension questions in context: Results for german and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9. Association for Computational Linguistics, 2011.

M. Mohler, R. Bunescu, and R. Mihalcea. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. pages 752–762, 2011.

Michael Mohler and Rada Mihalcea. 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 (EACL '09)*, (April):567–575, 2009. doi: 10.3115/1609067.1609130.

Rodney D Nielsen, Wayne H Ward, James H Martin, and Martha Palmer. Annotating students' understanding of science concepts. In *LREC*, 2008.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Stephen G. Pulman and Jana Z. Sukkarieh. Automatic short answer marking. *EdAppsNLP 05 Proceedings of the second workshop on Building Educational Applications Using NLP*, (June):9–16, 2005. doi: 10.3115/1609829.1609831.

Lakshmi Ramachandran, Jian Cheng, and Peter Foltz. Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 97–106, 2015. doi: 10.3115/v1/W15-0612.

Shourya Roy, Sandipan Dandapat, Ajay Nagesh, and Narahari Y. Wisdom of Students: A Consistent Automatic Short Answer Grading Technique. *Proceedings of the 13th International Conference on Natural Language Processing*, pages 178–187, 2016.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. Fast and Easy Short Answer Grading with High Accuracy. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1070–1075, 2016. doi: 10.18653/v1/N16-1123.

Hao Chuan Wang, Chun Yen Chang, and Tsai Yen Li. Assessing creative problem-solving with automated text grading. *Computers and Education*, 51(4):1450–1466, 2008. ISSN 03601315. doi: 10.1016/j.compedu.2008.01.006.

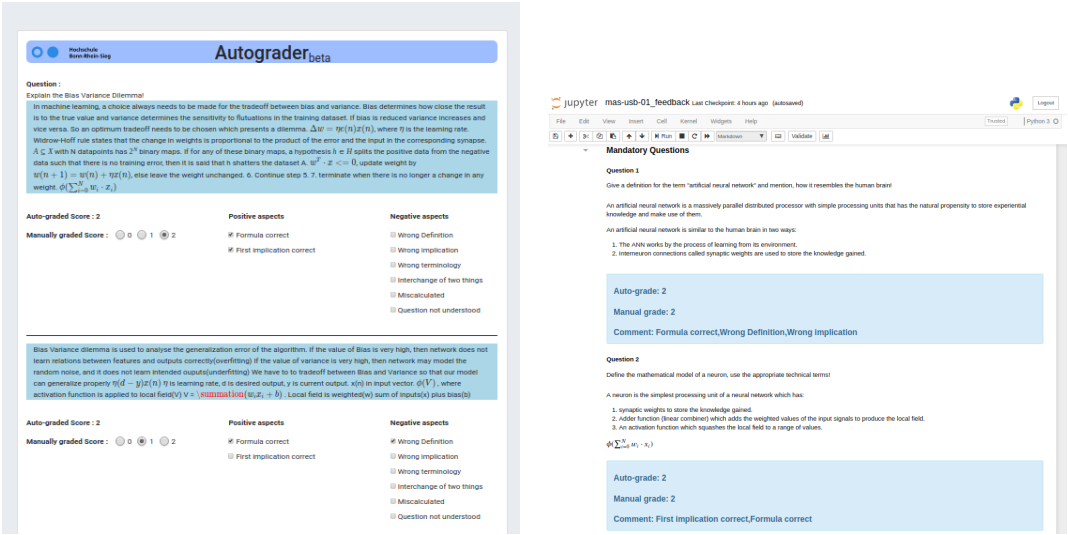
Torsten Zesch, Michael Heilman, and Aoife Cahill. Reducing annotation efforts in supervised short answer scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–132, 2015.

Appendix A. AI Assisted Grading System - GUI

The best active learning strategies (query strategy, and seed selection), features extracted from the answers, and machine learning model which were determined from the experimental results were incorporated into a Graphical User Interface (GUI) to be used by teachers or professors while grading. Initially, the system queries the grades for certain percentage of the total answers from the human grader, and then displays the notebooks in 'question-wise' view. The purpose of question-wise view is to make the task of assessing the answers easier for the humans.

This system reads all the notebook files of the students from a specified directory, converts them into a Pandas dataframe, extracts the features from the answers needed for the machine learning algorithm, learns the model for the task of grading, gets the grades confirmed by a human and stores the grades along with human expert's feedbacks which is saved in another directory as Jupyter Notebook files. Tasks which include getting the notebook files, processing them through the machine learning algorithm and saving the grades with feedbacks occur in the backend while the display of the questions, answers, grades and feedbacks is taken care of by the front end.

Once notebooks are converted into the Pandas dataframe format along with their features, the whole dataset is fed into the active learning algorithm. Figure 2 shows the stage where the queried answers appear in the GUI interface for the professor to grade. Certain



(a) Question-wise view of answers after auto-grading stage

(b) Final format of the notebook with grades and feedbacks

Figure 3: Different views of the AI assisted grading system

percentage of the answers are queried from the grader in this manner and the model is learned based on these grades. Once the grader is done grading the initial set of answers queried through active learning, the autograded scores are displayed along with the answers question by question.

Figure 3a shows this stage where all the students' answers to the first question for the grader to check. The grader could simply leave the grades as it is if he/she thinks that they are right or has the option to change the grades in the manual grade section. In addition, the professor can add further feedbacks by ticking the relevant opinions given under every question. All the answers are saved back into Jupyter notebook format and Figure 3b shows the final format of the notebook with the grades and feedbacks.