



Hochschule  
Bonn-Rhein-Sieg  
University of Applied Sciences



R&D Project Proposal

# Evaluating Active Learning for Short Answer Grading

*Mohandass Muthuraja, Jeeveswaran Kishaan*

Supervised by

Prof. Paul G. Ploeger

Deebul Nair

Month 20XX

## 1. Introduction

- An introduction to the general topic you are covering.
- Why is it important?

Short answer questions are characterized by the following aspects:

- the question must require a response that recalls external knowledge instead of requiring the answer to be recognized from within the question
- the question must require a response given in natural language
- the answer length should be roughly between one phrase and one paragraph
- the assessment of the responses should focus on the content instead of writing style
- the level of openness in open-ended versus close-ended responses should be restricted with an objective question design

Automatic short answer grading essentially deals with using computational methods to predict the grades for the answers, thus cutting down the effort and time of teachers / professors. Many automated approaches have been proposed in the past for grading short answer questions. These methods compare how similar the students' answers are to the one provided by the teacher or professor and assign a grade proportional to the magnitude of similarity. Different approaches of computing these similarity measures include handcrafted pattern matching, automatic pattern matching, lexical similarity (how particular words effect other words), semantic similarity (deals with the meaning of sentences), and entailments (whether one sentence leads to another without any contradiction). The abovementioned methods all belong to a learning paradigm called supervised learning where the right answer for every question is available. Though these approaches were able to produce decent results, they suffer from many shortcomings such as;

## Evaluating Active Learning for Short Answer Grading

- the failure to capture the different wordings/phrasing of the students while trying to answer the short answer questions. It is obvious that anticipating all different ways of answering his questions is practically impossible for the professor.
- lack of sufficient amount of labeled training data in the domain to learn the models. Reasons include privacy, availability, and the quality of the correct answers.
- accounting for small deviations in the answers which might affect the whole meaning of the sentence (for ex. in mathematical terms, though each and every word of the student's answer align with that of the professor, a small negation or inverse operation would change the whole meaning).
- finding a way to understand the underlying concept of various students' answers and bagging the similar ones (or the right and wrong ones separately) is also a very tedious task.

This work proposes a method where a system tries to learn this task of measuring the correctness of each answer with the help of a human in the loop. Active learning seems to be the best choice for this task as it actively queries the human expert to grade the answer where the training utility of grading that particular answer is high. The query should be effective such that the response from the human expert is informative as possible. By actively selecting the data samples to label, it reduces the need for a considerable amount of labeled training samples thereby alleviating the problem of insufficient labeled data which is prevalent in supervised learning approaches. Thus, it tries to reduce the human effort of going through all the answers while improving its understanding of the problem on an iterative basis.

The challenging task involved in following this approach is the selection of samples for which the labels could be queried from the human expert. Various active learning strategies are available to deal with the above mentioned

challenge like querying on grading the most uncertain answers, answers that could increase the efficiency of grading system, answers that could reduce the generalization error etc.,

### 1.1. Problem Statement

- What are you going to solve?
- How are you evaluating?
- One of the biggest challenge involved in supervised learning is the need for large labeled datasets. Generating a large data for short answer grading is also an obstacle. Many automatic short answer grading systems rely on supervised learning techniques. This, in turn, leads to high annotation cost of labeling the large training data. Also, the workload for labeling the data is more.
- As new data are generated on a regular basis, retraining the model for every new data also have a caveat of high computation cost. Developing a generic model to overcome the deficit of retraining the model for every new answer can solve this issue.
- Short answer grading typically focuses on the content rather than the style of writing. Short answers written by students have a lot of lexical diversity. Active learning can be used to handle the lexical diversity by involving human in training the model.

## 2. Related Work

- What have other people done?
- Why is it not sufficient?

### 2.1. Section 1

### 2.2. Section 2

## **3. Project Plan**

### **3.1. Work Packages**

The bare minimum will include the following packages:

WP1 Literature Search

WP2 Experiments

WP3 Project Report

Keep in mind that depending on your project, you will probably need to add work packages that are more suited to your projects.

### **3.2. Milestones**

M1 Literature search

M2 Experimental setup

M3 Experimental Analysis

M4 Report submission

### **3.3. Project Schedule**

Include a gantt chart here. It doesn't have to be detailed, but it should include the milestones you mentioned above. Make sure to include the writing of your report throughout the whole project, not just at the end.

## Evaluating Active Learning for Short Answer Grading

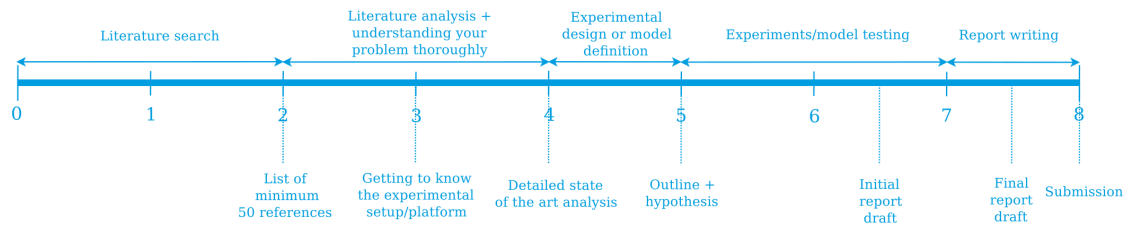


Figure 3.1:

### 3.4. Deliverables

#### 3.4.1. Minimum Viable

- Survey the existing approaches for interactive short answer grading
- Review existing datasets for automatic short answer grading.
- Compile new datasets from different in-house domain

#### 3.4.2. Expected

- Evaluate the quality of different features for interactive short answer grading
- Evaluate Active Learning strategies for Short-Answer grading
- Implement a working model of 100 clicks for 1000 grades.

#### 3.4.3. Desired

- Integrate the best model with a graphical user interface(GUI)