The effect of prompted self-revision on student performance in the context of open-ended problems using Randomized Control Trials

by

Meghana Kasal Vinayakumar

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

by

May 2020

APPROVED:

Professor Neil T. Heffernan, Thesis Advisor

Professor George T. Heineman, Thesis Reader

Professor Craig E. Wills, Head Of Department

Abstract

Assessments improve student learning. More than 50 years ago, Benjamin Bloom showed how to conduct this process in practical and highly effective ways when he described the practice of mastery learning (Bloom, 1968, 1971). Open-ended problems in assignments, as opposed to more closed-ended problems where there are a small set of known correct responses, offer an opportunity for students to demonstrate their understanding by articulating their underlying thought processes. In such problems, students are required to explain in a sentence or two, how to solve a particular problem or how they arrived at a solution. Open-ended responses stimulate a thought process in a student and allow teachers to better evaluate the student's deeper understanding of a topic beyond what can be observed in other problem types. Due to the open-ended nature of student responses to these problems, however, it is sometimes difficult for teachers to devote time to assessing student work, which causes students to apply lower effort or disengage from such problems if it is believed that a teacher is unlikely to attend to it. In order to promote better student engagement with these open-ended questions and to motivate them to apply more effort in answering these questions, I have built an infrastructure to conduct RCTs(Randomized Control Trials) with open-ended problems within ASSISTments, an online assessment tool; I have built an infrastructure that caters to machine learning models for the automated assessment of the student work. I am using this infrastructure to design an RCT that will evaluate the effect of prompted self-revision on the quality of the student responses.

Acknowledgements

I would like to thank my advisor, Professor Neil T. Heffernan for the guidance I received from him during the last two years. I would like to thank my mentor, Anthony Botelho for his guidance over the Thesis work. His patience and attention to details has been instrumental in making me a better researcher.I would like to thank John Erickson for his support while working on the fairness aspect of the model in the thesis. I also want to thank my thesis reader, Professor George T.Heineman, for giving useful feedback to make this thesis a success.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

An observation made in assessments portrays the decline in usage of open-ended questions over time. Responses to the open-ended questions provide the potential to assess how well a student understands a concept as opposed to close-ended questions. We would like to find ways to get students to attend more and give more effort in writing/improving their own textual responses. I am trying to build a tool that tries to get students more engaged in the response to open-ended questions and also try to understand how to motivate students to provide a better response.

1.1 Goal of the Thesis

The goal of this work is to develop infrastructure and explore methods that can be applied to study self-regulated learning in the context of open response problems. As such, this thesis intends to study how such prompts for students to revise their work affects their performance by conducting a randomized control trial. Towards this goal there are four main contributions of this thesis:

- I present the design and implementation of software infrastructure that is able

to leverage machine learning models and run studies, to test the effectiveness of learning interventions based on open response problems.

- I use this infrastructure to design and build an RCT to study the effectiveness of allowing students to revise their work.
- I conduct a set of pilot exploratory analyses to assess the risk of bias and fairness within existing machine learning models designed to assess student open-ended work.
- I explore the application of a statistical method known as REBAR to more accurately estimate treatment effects in the context of the proposed RCT.

1.2 Research Questions

This thesis intends to achieve these goals by addressing three research questions:

1. **What is the effect of prompted revision on student performance on open-ended problems? Are there benefits to prompting only when it is estimated that a student is likely to receive a low score?**

   When students are provided an opportunity to revise their responses to open-ended questions, would it motivate them to review their answers and would reviewing improvise the quality of the response? These questions are targeted to be answered by running a randomized control trial with "revision" intervention in the open-ended questions. This will be presented in section 3.2

2. **What is the potential risk of bias in Machine Learning models that assess these open responses?**

NLP models might incorporate bias within them. Prior research has found that pre-trained versions of word2vec and GloVe contain biases in the learned associations of words (Bolukbasi et al. 2016; Swinger et al. 2019) [2][3]. The usage of such models in assessing open responses might result in incorrect analysis. This will be presented in section 3.3

3. **Does the use of machine learning models developed to assess student open responses help in better estimating treatment effects?**

   Randomized Control Trial is often run on a small set of data. Using a statistical technique called REBAR it is previously shown that with historical data, which is often high dimensional and longitudinal, we can improve causal estimates from A/B tests. With Big data, it is shown that it can play a significant role in reducing the standard errors of treatment effect estimates[4]. Using this technique I aim to provide estimates of treatment effects that are more accurate. This will be presented in section 3.4

1.3 Background

By grading assignments, teachers assess student's gain in knowledge. But are these teachers consistent in their grading? Systems like ASSISTments allow teachers to manually grade responses to open-ended problems. These systems do not have a prescribed rubric to assess students. Teachers assess students based on how well they are able to articulate and demonstrate their knowledge of assigned content. In prior development work, it is shown that a pilot study was conducted to study teachers grading policy suggested that there exists a huge variation in how teachers approach grading these open-ended questions. The internal consistency of teachers grading their students were also low[5]. These developments do not address the

fairness of the model. About Machine learning Ming said(Founder and CEO, Socos Labs), "When you're doing this in advertising, who cares if you get it wrong, When you're doing it in diagnostics or in education or in hiring, you potentially just ruined someone's life. You have a real moral obligation to understand why your system is making the recommendations it is making."[1]. An automatic student assessment tool, should not consider the gender of the student when making its predictions. We use word embedding models in predicting grades of the students, I hereby try to analyze how safe are these in running our experiment.

In prior work, it is shown that feedback demonstrated higher performance in the student than the process goal without feedback condition in both experiments, although the differences were not statistically significant. These findings are noteworthy given that many children have difficulty assessing their writing skills (Hillerich, 1985)[10].

Self-efficacy also correlates with the Development of Academic Self-Efficacy 14 indexes of self-regulation, especially the use of effective learning strategies. Self-efficacy, self-regulation, and cognitive strategy use are positively inter correlated and predict achievement (Pintrich & De Groot, 1990)[7]. Keeping this in mind, we try to introduce indirect feedback, where a student is prompted to revise based on student's predicted grade in our system.

We try to evaluate the feedback by running a randomized control trial. Randomized Control Trials is an experimental way to learn how much better a treatment system is doing in comparison to that of a control system. Candidates for the experiment are randomly chosen to be in control and treatment conditions. The Control system is the current system and the Treatment system is an improvised system or a system that intends to bring a new feature. Randomized A/B tests hold a lot of promise for the study of student learning within intelligent tutors[4]. It is more likely

that our R.C.T is going to have less amount of data. With the help of a method discussed in a statistical experiment REBAR, I plan to use this historical data which is often high-dimensional and longitudinal, to improve causal estimates from R.C.T. Users with similar co-variate and outcome data as the participants in the A/B test, but who were not randomized are called "remnant". Data from the remnant may be quite useful—in particular, the extra sample size could improve the statistical precision, i.e. reduce the standard errors of experimental effect estimates. Vast amounts of log data, collected prior to the experiment, in conjunction with powerful machine-learning methods, could help sharpen causal estimates considerably[4].

Chapter 2

Experimental Overview

We would be needing an infrastructure to address one of the research questions "Running Randomized Control Trial to analyze the Revision intervention" mentioned above. RCT infrastructure development is discussed in chapter 3. Chapter 4 describes how to use RCT infrastructure built in chapter 3 to fit with that of the ASSISTments platform ,where we intend to run this experiment. Chapter 5 describes the experiment conducted to address the fairness of the model i.e., research question 2 .Chapter 6 describes the experiment conducted to better estimate the treatment effects i.e., research question 3.

Chapter 3

Infrastructure to Support RCTs on Open Response Problems

I built an infrastructure that can be generalized to answer such type of questions. This infrastructure is developed to replace existing open-response problems delivered through the ASSISTments online learning platform with open-response problems in which there is an embedded RCT; in this way, the infrastructure will be able to generalize to support virtually any study intended to run on open-ended problems. Using this infrastructure, described in this section, I propose to conduct an RCT to study the effects of prompted revision on student performance.

3.1 Requirements and goals of the system

1. Functional requirements:

    - Given an open-ended problem, a student should be able to see the question of the open-ended problem
    - Given an open-ended problem, a student should be able to write a solution

and submit

- Our system should be able to assign students into different conditions based on experimental requirements
- Our system should be able to log the actions of a student
- Our system should be able to run independently of the hosting systems

2. Experiment Specific requirements:
   - Provide a revision intervention for open-ended responses
   - Ability to connect with ASSISTments infrastructure

3.2 System APIs

We use REST APIs to expose the functionality. I am listing a few of the significant APIs for the system.

- API to get problem body:

  getProblemBody(database, table, extPrID)

  Parameters

  database: Name of the database

  Table: Name of the table

  extPrID: Problem ID passed by the hosting system

- API to get the condition to which system has assigned:

  etUserCondition(database, table,"userID",userID)

  Parameters

  Database: Name of the database

Table: Name of the table to fetch data from

"userID": Column of the table

userID: Value of the userID

- API to create a student database for the experiment:

  createUser(extUserId)

  Parameters

  extUserId: external user id of the student passed to the infrastructure by the hosting system

- API to assign students to a condition for a given problem:

  createUserCondition(userID, prMappingID, userCondition)

  Parameters

  userID: userID generated in our system for a given extUserId

  prMappingID: ID of the mapping from problem ID created in our system to that of the problem ID from the hosting system.

  userCondition: Integer value ranging the number of experimental states in our system

- API to create action types:

  createActionType(database,table,actionName,actionDesc=None)

  Parameters

  Database: Name of the database

  Table: Name of the table

  actionName: Name of the action to be created

actionDesc: Description of the action

- API to log all the actions:

  logAction(actionType, userPrConditionID, actionRes, time=now())

  Parameters

  actionType: action type can be problem load, solution submit, revise.

  userPrConditionID: unique identifier mapping userId, prID, userConditionID

  actionRes: generic JSON object which stores the result of the action

  Time: timestamp of the action performed

3.3 Database Design

We are using a SQL database: postgreSQL

3.3.1 ER Diagram

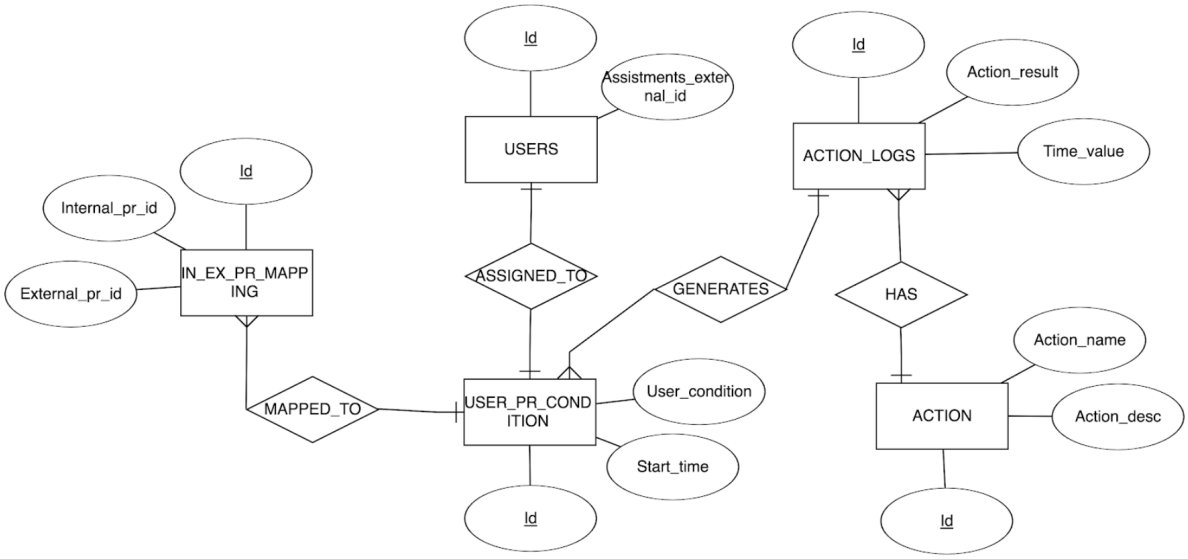

Figure 3.1: ER Diagram of RCT architecture

3.4 Codebase stack

IFrames can be used to enable our system to run independently of other systems and also to be able to be hosted in other systems. IFrame for the experiment is created using Jinja2 and rendered by Flask, a micro web framework written in Python. Communication with the database is done with Python. Javascript functions to connect our system with a hosted system.

3.5 Functionality of the architecture

Based on the problem information passed by the hosted system, students will see an open-ended problem, with a TinyMCE view to provide his/her solution. When a student enters the experiment for the first time, our system randomly assigns the user to a condition. Based on this condition, an intervention is provided by the system. Our system logs particular actions of a student like time at which the problem was loaded, what was the answer of the student, time at which his response was submitted and so on that would facilitate our analysis.

3.6 Connecting Randomized Control Trial Infrastructure to the host system-ASSISTments

For the current experiment, I am running RCT infrastructure in ASSISTments by establishing a secure connection with the ASSISTments system. ASSISTments will be notified of loading RCT infrastructure. From then on for every 10000th ms, a message with an action called "heartbeat" is sent to ASSISTments indicating that the RCT infrastructure is alive and is still doing some processing. Once the experiment is completed, a message with "complete" action is sent to ASSISTments and the connections will be closed.

Chapter 4

RCT on Student Revision

To address the second research question ie., to find the effect of prompted revision on student's performance, we devise the following process

4.1 Functionality

In our current experiment, students are randomly assigned to one of the three conditions: control condition, all-revise condition, conditional-revise condition. The control condition is our current system in which a student gets to work on an open-ended problem and submit his/her solution. In our experimental condition, students get a "revise" intervention. In this intervention, students would be given an opportunity to revise his answer before submitting. In all-revise condition, students will be asked to revise every time. In conditional-revise condition, a student will get an intervention only when his grades predicted are lower than a threshold.

In our experiment, once the answer is submitted, it's grade is predicted by a Machine Learning model, which would act as a threshold. For the current experiment, the threshold is set to 2.0 on a grading scale of 0.0 to 4.0. Although the interventions do not require us to predict the grades of the responses of the students

in control and all-revision conditions, it would cater for future analysis.

4.2 Frontend view of the RCT

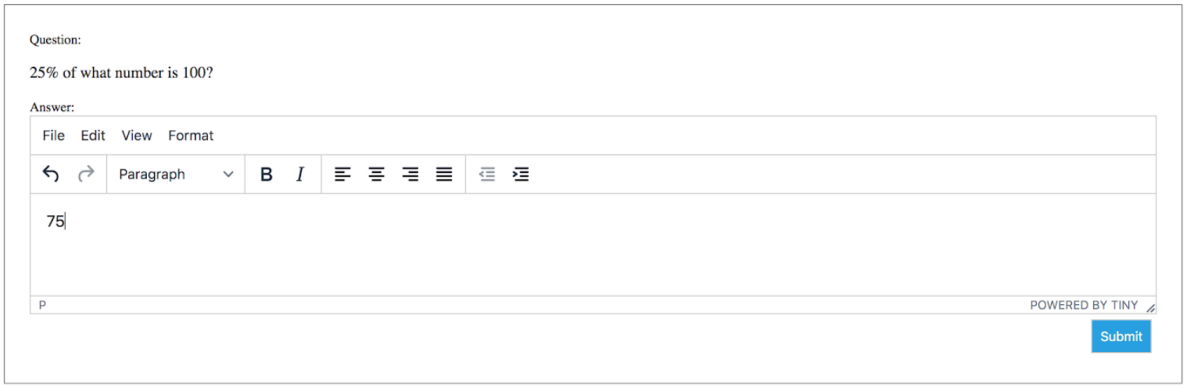

Figure 4.1: Open-Ended Problem

As soon as students start an assignment, they will see an open-ended question along with space to provide a response to this question [Fig-4.1]. Then they can submit their solution once they feel that they are all set.

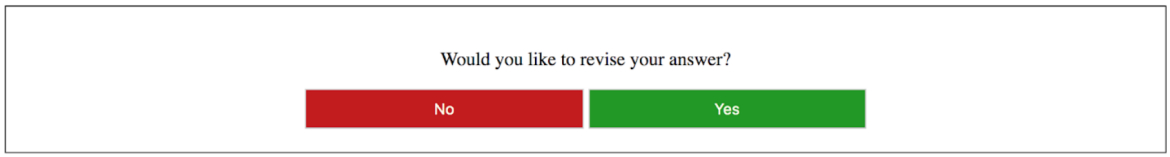

Figure 4.2: Intervention for students in all-revise condition and threshold-based conditional-revise condition of the experiment

If a student is in an "all-revise" condition or has scored less than the threshold in the "conditional-revise" condition, they are prompted to revise their answers once [Fig-4.2]. Then a student can choose to revise or not to. If students choose to revise, they will get the same view as in Fig-4.1 to improve their answers and then they can submit their final solution.

4.3 Discussion

With RCT in place, I propose to conduct the following two analysis was planned to address: Comparing the grade of a student's original response to the revised response, do we see a significant difference in their final grade? In the case of where we're only giving the revision suggestion when we think it is helpful, we'd want to do an additional analysis that looks at how much time was spent on the problem. Do we see significantly different amounts of time being spent on these problems, particularly between the two revision based conditions?

We have teachers working with a project DRIVER-SEAT which is funded by the National Science Foundation (NSF)[9]. We have these teachers help run an orchestrated study, where we're going to have them grade both the original and the revised responses which will become our source labels. As an additional exploratory analysis, we can see how well our models did compare to the grade that was provided by the teacher.

Chapter 5

Exploring Fairness in Open Response Grading Models

RCT relies heavily on the application of machine learning grading models; as such, it is important to study the fairness and other potential sources of bias that could influence the results of the RCT. There have been some developments in trying to build N.L.P models that automatically assess student open responses. It is shown that N.L.P models bring consistency in terms of grading[5]. We are using NLP models to predict the grades of the submitted responses of a student. The question we have to think is how fair is this model? I am conducting an experiment to answer research question 2 mentioned in the introduction of this paper.

5.1 Intuition

Consider a hypothetical slice of the embedding space, where each word mapped is not identifiable of gender as in [fig-5.1]. What if some words, consider words with red dots, are close to being differential?[fig-5.2]It is possible that the group of these

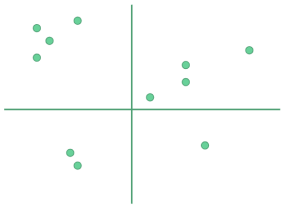

Figure 5.1: Hypothetical Slice

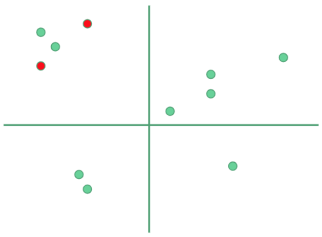

Figure 5.2: Hypothetical Slice with nearby words

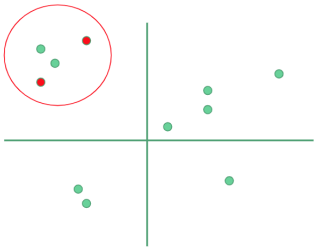

Figure 5.3: Hypothetical Slice with nearby words in a group

words may identify gender where the individual words could not.[fig-5.3]

Can a popular word embedding model like a GloVe, learn the gender of a student, just by the words they use, when trained with data that is free of gender i.e., is there a potential involvement of internally discovered gender variable in the prediction of the model?

$$Y = b_0 + b_1 W + b_2 X + G$$

where

W is the "words used by the student" ,

X is a matrix of some other covariates that we think would matter,

G is the gender parameter that GloVe learned from just the words of the student.

With this experiment, I am trying to see if a group of words can identify gender. Specifically, I am trying to explore if GloVe can learn gender from the association of words within the cluster of words.

5.2 Experiment

5.2.1 Null Hypothesis

No differences between the types of words used by gender as categorized by the geometry of glove

5.2.2 Dataset

Dataset is a collection of responses to open-ended mathematical questions of open educational curricula (EngageNY, Illustrative, and UtahMath). It consists of 4,639,258 words, 1,297,270 responses, 60,156 students, 3,703 classes.

5.2.3 Preprocessing

Open ended responses were tokenized with Stanford nltk tokenizer. Male and Female usage of these words were collected.

Word	Male usage	Female usage
acute	695	630
ad	170	178
add	5687	5410
added	2864	2957
...	...	...

Figure 5.4: Preprocessed data

Further with the help of my friends, Xangluin and Rui, filtered words from the corpus that have a significant difference in word usage by gender and also those words that are used fewer than 30 times in total. Preprocessed data is shown in the fig-5.4

5.2.4 Procedure

Words in the preprocessed data-set are mapped with 100-dimensional pre-trained word vectors of GloVe. Further, these words are clustered based on word vectors using k-means clustering.

In order to find the optimum number of clusters, I used the Elbow method.

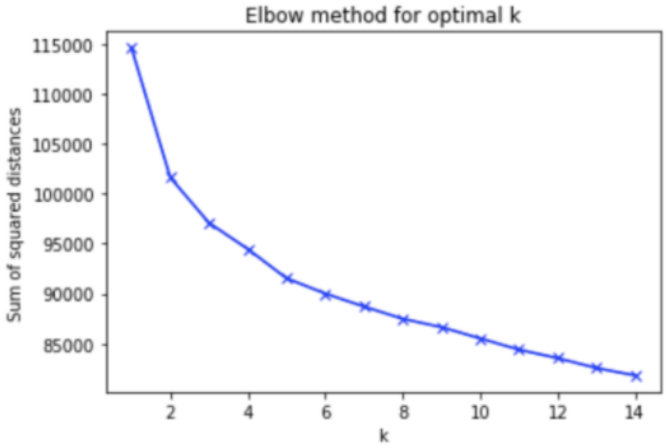

Figure 5.5: Elbow method to find an optimal number of clusters

In Fig-5.5 'k' specifies the number of clusters for which the kmeans model was trained on.

From Fig-5.5, although we see a slight elbow at k=5, it is not evident to point out on the exact elbow point.

Optimal cluster identification:

To decide on the optimal number of clusters with choices being k=5, k=10, k=15, I ran an experiment: Words closer to the center of the cluster will strongly represent features of the cluster and so we can interpret what every cluster represents. Hence for every cluster in cluster size k=5 , I extracted top 5 closest words to the center and interpreted the cluster representation. I repeated this experiment for k=10 and k=15.After examining words of all the choices for the optimal cluster sizes, we can

see from that cluster size 15 has words that are more interpretable. 5 closest words to the center are shown in tabel 5.1 and their interpretaion can be seen in table 5.2. It is clear that from this point onwards we can consider cluster size 15 for continuing with our main experiment.

The table 5.1 shows the 5 cluster-centric words under the column "Word", the cluster number to which each word belongs to is seen under the column "Cluster Group"

Word	Cluster Group
1,115	0
1,125	0
.	.
.	.
0-10	3
0-9	3
0-4	3
1-15	3
1-10	3
.	.
.	.
0.05	8
0.04	8
0.03	8
0.02	8
0.01	8
.	.
.	.
!!!!	14

Table 5.1: Words in Cluster

Cluster group	interpretation
0	mostly greater than 1000
1	Mathematical Words
2	includes mathematical operators
3	subtraction
4	special characters
5	decimals less than 0.1
6	-
7	units
8	decimals less than 0.1
9	decimals greater than 0.1
10	fractions
11	-
12	-
13	abbreviations
14	-

Table 5.2: Cluster Interpretation

Cluster	M	M-Proportion	F	F-Proportion	Difference	P-value
0	92869.0	50.662542	90440.0	49.337458	1.325084	0.0001
1	215215.0	49.759428	217296.0	50.240572	0.481144	0.0016
3	3761.0	52.127512	3454.0	47.872488	4.255024	0.0003
4	223831.0	49.748624	226093.0	50.251376	0.502752	0.0007
5	4048.0	46.342301	4687.0	53.657699	7.315398	0.0001
9	29545.0	50.511181	28947.0	49.488819	1.022362	0.0134
12	87301.0	49.336536	89649.0	50.663464	1.326928	0.0001
13	13343.0	48.666886	14074.0	51.333114	2.666228	0.0001

Table 5.3: Significant clusters obtained on conducting Proportional "N-1" Chi-squared results

Table 5.3 shows the significance level of the clusters experimented. Cluster column specifies the cluster number, M and F in the above column stands for the genders Male and Female respectively

5.2.5 Result

Proportional "N-1" Chi-squared test was performed manually on each of the clusters. 7 of the 15 clusters showed a significant difference.The significance was further confirmed by the Benjamini Hochberg method. Clusters representing 'subtraction' and 'decimals' showed higher usage differences of 4.24% and 7.32% respectively which are highlighted in table 5.3.

5.2.6 Experiment to explore the differential usage

To check the differential usage reported in clusters,representing 'subtraction' and 'decimals less than 0.1', which was reported previously from running the GloVe

model, I conducted "N-1 Chi-squared test" on these clusters without using the GloVe model. To run this experiment, consider just the words belonging to the clusters where the differential usage was high i.e., clusters representing subtraction and decimals. Male and female usage and the sample proportion were calculated and then a proportional "N-1" Chi-squared test was performed on each of the groups.

Cluster	M	M-Proportion	F	F-Proportion	Difference	P-value
Decimals	43263	49.986	43281	50.01	0.02%	0.9512
Subtraction	4515	50.78	4376	49.21	1.56%	0.1405

Table 5.4: Significant clusters obtained on conducting Proportional "N-1" Chi-squared results

Table 5.4 shows the significance level of the clusters experimented. M and F in the above column stands for the genders Male and Female respectively. For both the clusters, we see that the $p < 0.05$. Therefore Although GloVe model depicted usage difference between these clusters, there is no significant difference in actual usage of subtraction and decimal words between male and female when depicted manually without using the model.

5.2.7 Conclusion

There are potentially some words that are predictive of students' gender, but that does not really mean that there is bias in the models. This is the first step toward a much more extensive line of research studying fairness within these models

5.2.8 Discussion

This leaves us further to explore the ways to identify if the model is biased. Just because the model is potentially able to identify the genders by the words used, does that really mean that there is bias contained within such a model?

Chapter 6

Using REBAR to Better Estimate Treatment Effects

It is expected that the data collected from running the RCT will be less. To better estimate the treatment effects with small data I use a statistical method called 'REBAR'[4]. To see if a treatment works or not, we will plot a graph with an estimate for an experiment against effect size. The co-efficient of the Linear model gives the magnitude of the effect. To minimize the size of error bounds, I use REBAR. To increase the confidence of the effect size we can add more power to the experiment i.e., to increase the sample size. There are two ways to increase the sample size. One way is to increase the number of students in the experiment. This choice wouldn't be cost-effective in terms of time. We probably might need to span this experiment over multiple years, Just to run with 1000 students. The second way is to make use of a REBAR method. ASSISTments has a large pool of data of open-ended responses of all the students who have ever used this platform. Using REBAR we utilize this large pool of data to better estimate the outcome of the experiment. Thus add the predictive power to our experiment by artificially

increasing the sample size. Build a model on a large pool of data to predict the outcome variable and apply it to a small pool of data collected from running RCT to create a new outcome variable that is less noisy. The variance predicted outside of the experiment with the larger dataset can be used to remove variance form the outcome variable of the RCT dataset. Thus by extracting explainable variance leaves out the cleaner estimate of the RCT experiment.

To explore the potential benefits of REBAR, I present a simulation analysis that 1) adapts the REBAR method to utilize the automated grading models for open response problems, and 2) provides an example case and subsequent infrastructure that is intended to be included for pre-registration of the proposed RCT. I am using the models which do a good job of grading student open responses in mathematics developed by John et.al.

6.1 Methodology

To simulate the study, I have picked a problem for which scores seen across the samples are uniformly distributed over the grade range. 20% of the data is held out as an RCT experimental data. An Ensemble model of LSTM, Random Forest[5] is trained on 80% of data acts as a remnant data.

To simulate the various conditions in the RCT experiment, consider the smaller dataset which was held out as experimental data above. Then I divided this experimental dataset into two parts: One part representing the control condition and the other representing the all-revise condition. Assuming that there isn't much of the increase in the beneficial factor between all-revise and conditional-revise condition, I am running the simulation study with only control and all-revise condition. However, it is possible to run the simulation experiment with all three conditions with

small changes.

Students are randomly assigned to control and all-revise conditions. Let A_0 be the actual grade, A_0hat be the predicted grade, A_1 be the grade after revising their answer, and A_1hat be the predicted grade for the revised answer. For the students in the control condition, A_1 and A_1hat will be the same as that of the A_0 and A_0hat. It is assumed that when the actual experiment with the all-revise condition is executed, there is a small increase in grade value from that of the grade obtained on the original answer. To simulate this small increase in the grade value in our simulation study, we pick a value from uniform Gaussian Distribution with mean of 0.5 and standard deviation of 1. This value is added to the grades A_0 and A_0hat to get A_1 and A_1hat for the students in all-revise condition respectively i.e.,

$A_1 = A_0$ + Gaussian(0.5,1)

$A_1hat = A_0hat$ + Gaussian(0.5,1)

The new y = A_1 - A_0 and yhat = A_1hat - A_0hat. Residual, R is given as R = y - yhat.

A linear regression model is trained with R = intercept + C, where intercept is 1, C = [0,1] specifies the condition assigned for a student.

To compare the affect of the experiment, We will first fit linear regression model with outcome variable y , which tells us how the effect would be if students are in this experiment without revision and then we fit linear regression model separately with R and then we compare both the models.

6.2 Results

Problem number 1524987 is chosen to run the experiment. Upon training the linear regression model with the outcome variable y results from Table-6.1 Line 1 shows

Model-outcome	Variable	coef	p-value	[0.025	0.975]
y	x1	0.5435	0.020	0.088	0.999
R	x1	0.2311	0.410	-0.328	0.791

Table 6.1: Model summary of statistical significance

that the treatment effect was found to be 0.54 suggesting that if a student were to be in treatment in this model, he/she would likely have 0.54 better scores as a result of revising. This is a statistically reliable result with P = 0.02 and the confidence interval suggests that the model is 95% sure that the interval of the true co-efficient is somewhere between 0.088 and 0.999.

With the help of the REBAR trick, I plan to make the confidence bound smaller. The model summary is shown in Table-6.1 Line 2 says that the model has trouble finding the true coefficient. We see that the confidence bounds are very large, suggesting that it has a null effect. The result is not statistically reliable with P = 0.41 and with 0 falling in the range of confidence bounds -0.328 and 0.791.

6.3 Discussion

For this simulation experiment, The REBAR method did not lead to an improvement in the confidence of the treatment effect over the traditional methods. The model that we had did not represent the data very well. It added noise to our estimate as opposed to that of subtracting the noise. We can improve this by making a better model. The current baseline model has auc=0.6923, rmse=1.823, kappa = 0.1409. If we were to apply REBAR to the actual result of the study, we would like to use a better model that does better than the baseline.

Chapter 7

Summary

I have built an infrastructure to conduct RCTs(Randomized Control Trials) with open-ended problems within ASSISTments, an online assessment tool; I have built an infrastructure that caters to machine learning models for the automated assessment of the student work. I am using this infrastructure to design an RCT that will evaluate the effect of prompted self-revision on the quality of the student responses.

I studied the fairness of Machine Learning models and other potential sources of bias that could influence the results of the RCT. We found that there are potentially some words that are predictive of students' gender, but that does not really mean that there is bias in the models. This is the first step toward a much more extensive line of research studying fairness within these models

I explored REBAR method to better estimate the treatment affects by a simulation experiment.The REBAR method did not lead to an improvement in the confidence of the treatment effect over the traditional methods. However, It serves as a baseline model to beat if REBAR were to be applied in the actual experiment.

Chapter 8

References

[1] Education Technology's Machine Learning Problem—and Responsibility - By Tina Nazerian: https://www.edsurge.com/news/2018-07-11-education-technology-s-machine-learning-problem-and-responsibility

[2] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016)Ṁan is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems (pp. 4349-4357).

[3] Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., & Kalai, A. T. (2019, January). What are the biases in my word embedding?. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 305-311). ACM.

[4] Adam C Sales, Anthony Botelho, Thanaporn Patikorn, Neil T. Heffernan. (2018).Using Big Data to Sharpen Design-Based Inference in A/B Tests

[5] John Erickson, Anthony Botelho, Neil.T.Heffernan. The Automated Grading of Student Open Responses in Mathematics

[6] Schunk, D. H., & Swartz, C. W. (1993). Goals and progress feedback: Effects on self-efficacy and writing achievement. Contemporary Educational Psychology, 18(3), 337-354.

[7] Dale H. Schunk, Frank Pajares, The Development of Academic Self-Efficacy

[8] Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. Journal of Educational Psychology, 82, 33-40.

[9] For the current DRIVER-SEAT study, the project is funded by the National Science Foundation (NSF) under Grant Number 1822830.

[10] HILLERICH, R. L. (1985). Teaching children to write, K-8: A complete guide to developing writing skills. Englewood Cliffs, NJ: Prentice-Hall.