Anna Filighera, Tim Steuer, Christoph Rensing: *Fooling it - Student Attacks on Automatic Short Answer Grading.* In: Carlos Alario-Hoyos, María Jesús Rodríguez-Triana, Maren Scheffel, Inmaculada Arnedillo-Sánchez, Sebastian Maximilian Dennerlein (Eds.): Addressing Global Challenges and Quality Education, p. 347-352 Springer International Publishing, September 2020. ISBN 978-3-030-57717-9.
https://doi.org/10.1007/978-3-030-57717-9_25 .

Fooling it - Student Attacks on Automatic Short Answer Grading

Anna Filighera (✉)[1][0000−0001−5519−9959], Tim Steuer[1][0000−0002−3141−712X], and Christoph Rensing[1]

Technical University of Darmstadt, Multimedia Communications Lab, Germany
{anna.filighera,tim.steuer,christoph.rensing}@kom.tu-darmstadt.de

Abstract. Modern machine learning approaches have been shown to be vulnerable to adversarial attacks in many fields. This is a critical weakness, especially for models that are expected to function in an adversarial environment, such as automatic grading models in exams. However, as most of these attacks are either limited in their success rate, their applicability in diverse scenarios or require mathematical expertise of the attacker, the question arises to which extent students themselves are even capable of fooling state-of-the-art grading models. This work aims to investigate this question for the short answer question format. For this purpose, we tasked students of an educational technologies university course with probing the state-of-the-art automatic short answer grading model for weaknesses. Of the fourteen active participants, only one reported the model to be sufficiently free of deficits. The following weaknesses were identified by the students: a disregard for negation, no plagiarism detection, correct answers not being predicted as such and oversensitivity to small linguistic changes in answers, triggers, and keywords.

Keywords: Automatic Short Answer Grading · Adversarial Attacks · Automatic Assessment

1 Introduction and Related Work

Recently, a neural automatic short answer grading (ASAG) approach was published that reportedly outperforms human graders [6]. And while the authors show this to be true in the sense that the model's judgement correlates higher with the gold standard than the individual human graders', there are arguably more aspects that have to be investigated before we can claim to outperform humans on this task. One of these aspects is the ability to adequately deal with cheating attempts. Students may exploit systematic weaknesses in grading models. While methods to automatically identify such weaknesses exist, the question stands *to what extent students are capable of fooling state-of-the-art automatic short answer grading systems.* To investigate this question we employed a reproduction [3] of the state-of-the-art ASAG approach proposed by Sung et al. [6] in a university course about educational technologies and tasked students with finding answers the system was not able to judge correctly. They were also asked

to comment on identified weaknesses. We make the resulting dataset publicly available [1].

Manually constructing adversarial examples is not a new concept. Ettinger et al. [2] let linguistic and NLP researchers generate adversarial examples for model evaluation. Our work differs from theirs in key aspects. Firstly, the attackers in our scenario are university students with limited prior experience in NLP. Secondly, the students are not presented with a fixed set of predictions but instead can query the model as they see fit. Wallace et al. [7] propose a framework that assists humans in generating adversarial examples and evaluate it with trivia enthusiasts on a Quizzbowl question answering task. The framework provides the attacker with information about the inner workings of the victim model, such as the evidence text snippets utilized by the model. While these insights provide good clues to the attackers as to which parts of their questions they have to modify to fool the model, we do not expect students to have such information.

2 Methodology and Results

In practice, we expect grading models and datasets to be kept secret. This is an easy first line of defense against adversarial attacks as it forces attackers to limit themselves to black-box attacks or estimate needed information about the model's inner workings, e.g. gradients. However, we assume that students will have continuous or at least recurrent access to feedback assigned by the model. This could be the case if students have graded exercises before taking part in an exam or if students may retake exams. Alternatively, the students may collect information on the grading model over years as they do on exam questions now.

The short answer questions used in this study stem from two distinct datasets, the SciEntsBank dataset of the SemEval-2013 challenge [1] and the computer science questions proposed by Mohler et al. [5]. We selected 3 questions (see Table 1) from the SciEntsBank training set that were short and did not reference any external material, such as figures. All the computer science questions were included. However, students were tasked to mainly focus on the SciEntsBank questions, as the grading model has seen these during training. This is also the scenario we expect to see in real-world applications where models are trained to grade specific questions. According to the SciEntsBank's labeling scheme, we consider answers to either be *correct, incorrect* or *contradictory*. The *incorrect* class incorporates all false answers that are not direct opposites to the *correct* answers, such as partially correct, irrelevant or non-domain answers.

The attackers were students enrolled in Technical University of Darmstadt taking an educational technologies course. Of the 24 course participants, 13 were male and 11 female. On average, the students were 25 years old at the time of this study. 96% spoke English either proficiently or fluently. Of the 24 course participants, 14 answered the questions connected to the task. Nevertheless, the others may have submitted answers without commenting as the answer submission was anonymous. The students were tasked to find example answers the

[1] https://github.com/PumpkinPieTroelf/ASAG-Adversarial-Dataset.git

grading model misclassified. For this purpose, we built a web-interface where students could submit answers and receive the predicted classes. The students had no further insights into the grading model besides knowing that it was BERT-based. They were also asked to provide the - in their opinion - correct class for each answer. Finally, in the context of a voluntary graded exercise for bonus points in the end-of-term exam, they should report the five most interesting examples they had found and give a comment on how the model performed in general. They had one week to query the model and answer the graded questions during which they received no feedback beyond the model's predictions.

Table 1. Example misclassified student answers to questions from the training data.

Question:	When a seed germinates, why does the root grow first?
Reference Answer:	The root grows first so the root can take up water for the plant.
Labeled *correct*:	Because the seed needs to stay away from water.
Labeled *correct*:	The plant needs no water
Labeled *correct*:	plant water.
Labeled *correct*:	Because Chewbacca eats plant water.
Labeled *correct*:	The seed contains much water, so the root pumps it into the ground.
Labeled *incorrect*:	Because the seed needs liquid.
Question:	How do you define a controlled experiment?
Reference Answer:	An experiment is controlled if only one variable is changed at a time.
Labeled *incorrect*:	An experiment with only one person.
Labeled *correct*:	None exists An experiment with only one person.
Labeled *correct*:	A controlled experiment is one in which nothing is held constant except for one variable.

Results: We first analyze the success of the students' attacks quantitatively. Then, we present some of the interesting examples the students found underlined with the student statements regarding the quality of the model's predictions. The students' statements were processed using the summarising content analysis according to Mayring [4] to identify commonly identified weaknesses. Students submitted 620 answers in total. However, we excluded a remote code execution attempt from further analysis. We expected students to focus on linguistic attacks and were thus surprised to see this style of attack.

The other 619 student answers consisted of 262 *correct*, 328 *incorrect* and 29 *contradictory* answers. The model's confusion matrix and resulting metrics can be seen in Table 2. As we can see, the model seems biased towards the *incorrect* class. This is also an observation we made on the original SciEntsBank data. The model predicts almost half of the *correct* answers correctly, while not recognizing a single *contradictory* answer. However, this is unsurprising considering

the low representation of this class in the training data. The students managed to fool the model to classify 13.4% of their *incorrect* answers as *correct*. As students typically stopped submitting *incorrect* answers once they found a reliable strategy, this number should be viewed as a lower bound of their capabilities.

Table 2. Model's confusion matrix and classwise metrics on the student answers.

		Predicted Class			**Classwise Metrics**		
		correct	incorrect	contra.	Precision	Recall	F1-score
True Class	correct	130	129	3	0.71	0.50	0.58
	incorrect	44	273	11	0.65	0.83	0.73
	contra.	10	19	0	0.00	0.00	0.00

Table 3. Weakness categories identified during the content analysis of the student comments. The second column denotes the number of occurrences, the last examples of comment snippets for each category. Please note that the snippets were translated as most students opted to answer the questions in German.

Category	#	Example Snippet
Sufficient	1	In my opinion the model works very well
Plagiarism	1	Answers copied from Wikipedia were marked as correct. While this is true, a teacher would probably have realized that the answer is "copied".
Inversion	3	In addition, the examples suggest that sentences that are worded similar to the correct answer, such as Something that emits heat. for question 1 are still classified as correct although the difference means that they mean the opposite.
Brittleness	4	The model is not very robust, because small changes to the inputs falsify the result
Keywords	6	The model rates answers as correct if they contain the correct words. Whether these words are in the correct order, that is whether they fulfill the right roles in the sentence, or whether the syntax of the sentence is even approximately correct, is not checked.
False Negative	8	I have not been able to get an answer correctly rated as correct except in question 2, although the answers were valid and I tried them out in different versions.

If we now have a closer look at the answers (Table 1) and comments (Table 3) the students provided, interesting patterns emerge. Especially the answers to the plant question indicate an oversensitivity to specific keywords contained in the reference answer. This is also a weakness 6 of the 14 students consciously identified and reported. Another vulnerability 3 of the students mentioned was

the model's disregard for negation or inversion of answers. This is illustrated by the last example answer where the inverted reference answer is still labeled as *correct* despite its contradictory meaning. The bias towards labeling answers as *incorrect* we have discussed before also reflects in this data. Eight students criticized the model's tendency to refuse actually correct answers. Even exact definition quotes from Wikipedia were rejected in some cases. Furthermore, four students declared the model to be brittle in the sense that small, often nonsensical, changes in the answer led to vastly different predictions. An example of this behaviour is depicted in Table 1 where prepending "None exists" to a correctly labeled *incorrect* answer leads the model to predict it as *correct*.

3 Conclusion and Future Work

In conclusion, we have seen that students were able to identify and exploit systematic weaknesses of the state-of-the-art grading model, even for questions the model was trained for and despite only having the model's predictions to inform their perturbations. However, a larger study with a heterogeneous body of students is needed to ensure the generalizability of these findings. Nevertheless, this study illustrates the need for further research and refinement of automatic short answer graders before employment in high-stake scenarios is advisable.

References

1. Dzikovska, M.O., Nielsen, R.D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., Dang, H.T.: Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. Tech. rep., NORTH TEXAS STATE UNIV DENTON (2013)
2. Ettinger, A., Rao, S., Daumé III, H., Bender, E.M.: Towards linguistically generalizable nlp systems: A workshop and shared task. arXiv preprint arXiv:1711.01505 (2017)
3. Filighera, A., Steuer, T., Rensing, C.: Fooling automatic short answer grading systems. In: International Conference on Artificial Intelligence in Education. Springer (2020, in press)
4. Mayring, P.: Qualitative inhaltsanalyse. In: Handbuch qualitative Forschung in der Psychologie, pp. 601–613. Springer (2010)
5. Mohler, M., Bunescu, R., Mihalcea, R.: Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 752–762. Association for Computational Linguistics (2011)
6. Sung, C., Dhamecha, T.I., Mukhi, N.: Improving short answer grading using transformer-based pre-training. In: International Conference on Artificial Intelligence in Education. pp. 469–481. Springer (2019)
7. Wallace, E., Rodriguez, P., Feng, S., Yamada, I., Boyd-Graber, J.: Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. Transactions of the Association for Computational Linguistics **7**, 387–401 (2019)