A Transformer for SAG: What Does it Grade?

Nico Willms
Hochschule für Technik
Stuttgart, Germany

Ulrike Padó
Hochschule für Technik
Stuttgart, Germany
ulrike.pado@hft-stuttgart.de

Abstract

Automatic short-answer grading aims to predict human grades for short free-text answers to test questions, in order to support or replace human grading. Despite active research, there is to date no wide-spread use of ASAG in real-world teaching. One reason is a lack of transparency of popular methods like Transformer-based deep neural networks, which means that students and teachers cannot know how much to trust automated grading. We probe one such model using the adversarial attack paradigm to better understand their reliance on syntactic and semantic information in the student answers, and their vulnerability to the (easily manipulated) answer length. We find that the model is, reassuringly, likely to reject answers with missing syntactic and semantic information, but that it picks up on the correlation between answer length and correctness in standard training. Thus, real-world applications have to safeguard against exploitation of answer length.

1 Introduction

Automated short-answer grading (ASAG) promises to support or replace human grading decisions for student-constructed answers to test questions and in this way avoid human error and save teachers' time and effort. In the context of formative testing for frequent feedback, online teaching and self-study, ASAG is especially attractive, since human grading effort is significant due to repeated testing or large groups, and the need for feedback can arise at any time of day or night in the case of self-study (Burrows et al., 2015).

ASAG models are not currently in wide-spread use in real-world teaching contexts (e.g., Lee and Shin (2020); Wilson et al. (2021) for the related task of essay scoring). Three requirements for their adoption are reliable performance on small-scale, real-word data, ease of development for non-experts and transparency of model decision making, both for teachers and students.

Transformer-based models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have been recently successfully explored for ASAG (Camus and Filighera, 2020; Bexte et al., 2022). Given the relatively small size of available training data for ASAG (in the low ten thousands of data points), the great advantage of these models is that they are freely available pre-trained on large data sets and require only relatively small data sets for fine-tuning to a specific task. Another advantage is that they require no manual feature engineering, as relevant patterns are derived by the complex neural networks from word distributions in the very large pre-training data sets. Transformer-based models therefore seem like good candidates to address the reliability on small data sets and ease of development criteria.

However, the grading decisions made by neural models are intransparent. This makes it hard for teachers to understand how to best use ASAG on specific data sets - are the predictions reliable enough to replace human grades, should they be manually revised, or are the available models unreliable altogether for their data? A related question is what the model predictions are based on - do they consider the content of the short answers, as intended, or do they also rely on extraneous signals, and can they be swayed by trivial manipulations of the input that would not convince a human grader? Since real-world grading applications have to gain the trust of teachers and students alike, these questions are highly relevant for practical application. This paper aims to further understand the functioning and limitations of a standard Transformer-based ASAG model.

Since ASAG is a semantic task (similar to the Natural Language Inference and Paraphrase Detection sub-tasks in the GLUE benchmark, which BERT does well on, see Devlin et al. (2019)), we

Nico Willms and Ulrike Pado. A Transformer for SAG: What Does it Grade?. *Proceedings of the 11th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2022)*. Linköping Electronic Conference Proceedings 190: 114–122.

hope to see sensitivity to the content of the input beyond keyword spotting. At the same time, trivial manipulations of the input should not affect the predicted grade.

One strategy for probing model behaviour and representations are adversarial attacks (Goodfellow et al., 2015), modifications of the input data that allow us to evaluate model behaviour in a controlled experiment. The strategy has been used before to establish relevant insights about neural ASAG: Ding et al. (2020) established that a recursive neural network was sensitive to combinations of content words (rather than just keywords, for example) for ASAG. Looking at the possibility of fooling the model, Filighera et al. (2020) were able to identify two-word trigger phrases that in some cases suffice to switch the predictions of a BERT-based model when added to student answers – while not altering the content of the student answer in a meaningful way.

We present several experiments to investigate a Transformer-based model's sensitivity to syntactic and semantic information in ASAG student answers, as well as a confounding (and potentially exploitable) length effect. Experiment 1 (Section 4) investigates the system's reaction to removal of syntactic information (namely, word order and function words). Experiment 2 (Section 5) explores the extent of the system's reliance on content words from different word classes and its robustness in case they are removed. Finally, Experiment 3 (Section 6) investigates the impact of input length.

We find that the model uses syntactic information (such as word order and function words for English), but its loss is not catastrophic for model performance. Removing nouns from the input data has the most tangible effect in Experiment 2, reducing the model's ability to identify a correct answer to 50% (when a human grader would likely be similarly affected). These results underscore that the model does rely on the meaning of the short answers to arrive at its grade prediction, as we had hoped.

However, we also find that the model is easily swayed by input length: Longer answers are much more likely to be graded *correct*. This pattern is visible in the training data and clearly picked up by the model. This result is alarming, since the length signal is easy to manipulate.

2 Related Work

Traditionally, extensive feature engineering on the lexical, syntactic and semantic level has been employed for ASAG (see Burrows et al. (2015) for an overview). More recently, neural network-based approaches have been tested, for example in work by Riordan et al. (2017) using an LSTM (Long Short-Term Memory, Hochreiter and Schmidhuber (1997)) or by Sung et al. (2019) or Camus and Filighera (2020) using the Transformer-based BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) models.

While Riordan et al. (2017) report that their LSTM-based model approaches the state of the art, the BERT-based approach of Sung et al. (2019) is the first to improve on the state of the art for a standard ASAG data set. The use of domain-specific data in pre-training and fine-tuning proves helpful, but makes performance brittle in unknown domains. Camus and Filighera (2020) demonstrate that fine-tuning on tasks related to ASAG (like Natural Language Inference, where systems decide whether a hypothesis follows from a premise) yields more robust improvements, even though the fine-tuning data has little connection to the topic domain of the test data.

To date, the state of the art on standard benchmark data sets is set by combinations of neural and traditional machine learning approaches: Saha et al. (2018) and Sahu and Bhowmick (2020) combine word and sentence embeddings with string-based similarity methods. Since these approaches inherit both the need for feature engineering and for extensive pre-training of the embeddings, they are harder to re-create for the application of ASAG methods in teaching practice. We therefore focus on the Transformer-based models in this paper due to their comparable ease of use.

BERT and related models have been investigated extensively with different strategies over the last years, finding that BERT learns syntactic (Hewitt and Manning, 2019; Tenney et al., 2019) and semantic representations (Ettinger, 2020) that are generally preserved through fine-tuning for semantic tasks like paraphrasing (Pérez-Mayos et al., 2021). However, Hessel and Schofield (2021) find that on the GLUE tasks (Wang et al., 2018), BERT is relatively insensitive to shuffling of the input sentences, which removes many syntactic clues in English. For ASAG, this behaviour is double-edged: On the one hand, ASAG focuses

on scoring answer content over answer form, so insensitivity to shuffled (or syntactically incorrect) input is an advantage. On the other hand, input words in truly random order would certainly be noticed by human graders and could indicate an attempt at manipulating the grade. Insensitivity to word order removes the system's ability to filter out such answers.

More problematic for the use of BERT-based models for ASAG are results from Ettinger (2020) that BERT is insensitive to negation in a word prediction task. For a task like ASAG, the removal or addition of negation to a student answer will likely immediately change the correct grade, so sensitivity to this information is vital.

Adversarial attacks specifically have been a fertile approach for studying neural networks in NLP in recent years (Zhang et al., 2020). Specifically for ASAG, Ding et al. (2020) found that attacks randomly generated from prompt specific words were more easily accepted by the system, more so if longer word sequences remained intact and most readily if the attack was generated by shuffling, since all lexical material is preserved. This is in line with the results by Hessel and Schofield (2021) and points to semantic association at the core of ASAG performance.

Filighera et al. (2020) identified a number of two-word trigger sequences that would frequently switch an ASAG grade from *incorrect* to *correct* when simply prepended to the student answers, increasing the misclassification rate of the attacked model by about 130-160%. This is a clear attack vector for grade manipulation, although it does not guarantee misclassification: Adding the triggers does not flip classification for any answer but only for ones that were already somewhat similar to the target answer.

3 Method

Data We work with two corpora that, together, constitute the standard English-language SemEval-2013[1] data set (Dzikovska et al., 2013), Beetle and SciEntsBank (SEB). The corpora contain student answers to science domain questions; Beetle (3.6k answers) was collected from interactions with a tutoring system, while SEB (4.5k answers) stems from a conventional test setting.

Evaluation Both corpora offer in-domain and out-of-domain test sets. For the in-domain test sets, additional *unseen answers* (UA) to questions from the training set are presented. In addition, there are also test sets containing completely new questions and their answers, called *unseen questions* (UQ). Finally, for SEB, there are also questions from an *unseen domain* (UD).

The task is to determine the human-annotated grade for a student answer by comparing it to a given correct reference answer. In the literature, Beetle is rarely used, since it provides several reference answers per question. Here, we append these reference answers into a single input.

We report Macro F_1 scores (for comparison to the literature state of the art) and Accuracy (for experimental evaluation) on the test sets, using the binary classification labels. In addition to overall Accuracy, we also break down the results into label-wise Accuracy. Across all data sets, the `incorrect` answers are the majority class (consistently at about 60% across all data subsets).

Model We aim to create a model close to the state of the art. Given the results in Camus and Filighera (2020), RoBERTa$_{base}$ as well as models pre-trained on the MRPC paraphrasing task (RoBERTa$_{MRPC}$) and the MNLI Natural Language Inference task (RoBERTa$_{MNLI}$) were separately fine-tuned on SEB and Beetle.

On a development set comprising 10% of the training data, we determined the optimal number of training epochs and compared the results for three versions of each RoBERTa model based on different random seeds. The models received a maximum of 256 tokens per input sentence. We used the Adam optimiser with an initial learning rate of 5e-5, and ϵ of 1e-8; batch size for training was 8. RoBERTa$_{MNLI}$ consistently outperformed the other model instances on the development set, so this model (with seed 100 and 6 epochs of training for SEB and seed 1 or 100 and 6 epochs of training for Beetle) was chosen.

Table 1 shows that we have succeeded in training a model that closely matches or numerically outperforms the state of the art for both corpora using macro F_1: We compare to Saha et al. (2018) on SEB.[2] and report the first results for 2-way Beetle since SemEval-2013[3].

[1] Available from https://www.cs.york.ac.uk/semeval-2013/task7/index.php%3Fid=data.html.

[2] Ghavidel et al. (2020) achieved a slightly higher F_1 score for UA at 79.7, but lower scores for UQ and UD.

[3] Results for the best model for each test set from the top-

	Beetle		SEB		
	UA	UQ	UA	UQ	UD
best SemEval-13	83.3	72.0	76.8	73.7	70.5
Saha et al. 2018	–	–	78.6	73.9	70.9
RoBERTa$_{MNLI}$	**89.7**	**78.1**	**82.2**	**74.1**	**72.1**

Table 1: Macro F_1 on the test sets for literature benchmarks and RoBERTa$_{MNLI}$.

Adversarial Attacks We modify the SEB and Beetle test data in different ways and compare model performance on the original and modified data using the difference in overall and label-specific model accuracy. This strategy allows us to both show the effect of the attack and to factor in imperfect model performance on the original data.

We will create attacks to evaluate the model's reliance on syntactic and semantic cues. In both cases, we will remove information from the student answers in the official test sets. For syntactic information, this means removing word order by shuffling and removing function words (as identified by the NLTK[4] tagger). For semantic information, we remove different content word classes (nouns, verbs, adjectives and adverbs). Since our strategy shortens the original student answers, we also closely look at the influence of answer length (by duplicating the original answers and by generating synthetic answers in different length bands).

If our attacks impair the model's ability to recognise the correct student answers, we expect a drop in overall prediction Accuracy, and more specifically, a strong decrease in prediction Accuracy for originally *correct* items (below, Acc$_{corr}$) and possibly an increase in Accuracy for originally *incorrect* items (Acc$_{incorr}$). If the model's ability to recognise *incorrect* answers suffers, overall prediction Accuracy will drop as well, but this time driven by lower Acc$_{incorr}$.

4 Experiment 1: Syntax

Our first experiment tests the impact of deleting syntactic information from the student answers. We try two strategies: Shuffling the input data (so word order information is lost), and deleting all tokens not belonging to the noun, verb, adjective and adverb classes: for example, pronouns, determiners, prepositions or conjunctions. In a third attack, we delete non-content tokens *and* shuffle. Sample attack items can be found in Table 2.

Table 3 shows the results: Shuffling and deleting non-content words both lead to lower Accuracy scores for both corpora (the table shows Δ Accuracy to the unaltered test data). The effect increases when we combine the attack strategies *and* the student answers are reduced to bags of content words.

Interestingly, the Beetle model is much more sensitive to the attack than the SEB model. Inspection of the data shows that Beetle contains many questions on opened and closed electrical circuits, where the direction of relations like *connected-to* is highly relevant and often signalled through syntactic means. Possibly, this is why model performance is hurt so much when syntactic signals are removed.

We look at the label-specific Accuracy results (see Table 6) to determine the cause of the observed drops in overall Accuracy. The results can be summarised as follows: As hypothesised, the drop in overall Accuracy is driven for both corpora and all test sets by a strong shift towards always predicting *incorrect*. For instance, looking at the most extreme attack of *shuffle+content only*, the label-specific Accuracy for *correct* instances drops by 50 percentage points for Beetle-UA while the label-specific Accuracy for *incorrect* rises by almost seven percentage points. The picture for Beetle-UQ is similar, and while the drops are generally less dramatic for SEB, the pattern is the same. Acc$_{corr}$ drops by about 13 points for SEB-UA and -UD (and by 33 points for SEB-UD), Acc$_{incorr}$ rises by 2-3 percentage points. This means that almost half of the bags of content words created by the attack are now so dissimilar to the reference answers that the models no longer recognise them as a correct answer.

In sum, the impact on the Accuracy of grading *correct* student answers is quite strong across all test sets, demonstrating that the RoBERTa models do use syntactic information in their decision-making. However, for the SEB model, the lack of syntactic cues is never catastrophic: The ma-

ranked Heilman and Madnani (2013) and Ott et al. (2013).

[4] https://www.nltk.org/

Syntax and Semantics Attacks		Length attacks	
Original	there is a damaged bulb	Original	there is a damaged bulb
Syntax: Shuffle	bulb there damaged is a	Rand. Short	was path in is or is closed has incorrect
Syntax: delete Non-Content	is damaged bulb	Rand. Avg.	a affect terminal terminal by bulb [...] (34 words)
Semantics: delete Nouns	there is a damaged	Rand. Long	a and c path state difference bulb [...] (93 words)
Semantics: delete Verbs	there a bulb	Duplicate	there is a damaged bulb there is a damaged bulb

Table 2: Adversarial attack items for syntax, semantics and length (Rand: randomly generated) attacks (Beetle).

	Beetle		SEB		
	UA	UQ	UA	UQ	UD
Test data	89.7	78.1	82.2	74.1	72.1
Shuffle	-9.3	-7.2	-3.7	-2.2	**-1.8**
Content only	-9.5	-6.4	**-4.4**	-0.3	+0.5
Both	**-16.1**	**-11.2**	-3.7	**-7.3**	**-1.5**

Table 3: Exp.1: Removing syntactic information: Shuffling and removing non-content words from the SEB and Beetle Unseen Answer (UA), Unseen Question (UQ) and Unseen Domain (UD) test sets, overall Δ Accuracy (lowest result in boldface).

	Beetle		SEB		
	UA	UQ	UA	UQ	UD
Test data	89.7	78.1	82.2	74.1	72.1
No Adj	-7.7	-4.8	-7.6	-1.7	-2.1
No Adv	-5.2	-2.2	-1.5	-3.0	-0.2
No Nouns	**-10.2**	**-14.5**	**-8.3**	**-3.8**	-2.3
No Verbs	-4.3	-0.7	-4.8	+0.4	**-3.5**

Table 4: Exp.2: Removing various content word classes from the SEB and Beetle Unseen Answer (UA), Unseen Question (UQ) and Unseen Domain (UD) test sets, overall Δ Accuracy (lowest result in boldface).

jority of correct student answers is still recognised based on shuffled content words only. For ASAG this means that a relevant combination of content words still has a chance of being recognised as a correct answer even if it is not syntactically correct. This potentially helps non-native speakers and is in line with the focus on content in ASAG. Looking at the attack items in Table 2, it is likely that human graders would be able to interpret some of these answers and grade them as *correct*, as well - we did not investigate this point further, however.

5 Experiment 2: Semantics

In Exp. 1, we probed the influence of syntactic information by excluding all non-content words. In Exp. 2, we ask about the relative importance of the different classes of content words instead. We create four different sets of attack items by selectively removing all nouns (or verbs, adjectives or adverbs) as identified by the NLTK tagger. We hypothesise that nouns and verbs furnish the most crucial information for correct grading, so removing them from the test set answers should affect grading Accuracy most. Negation expressed by "not" will be removed with the adverbs, so grading may suffer in this case, as well (since the meaning of the student answers will be substantially changed by the deletion).

We find a clear impact of removing content words (see Table 4), with the greatest effect from deleting nouns (while the SEB-UD model does worst without verbs). For three out of five test sets (Beetle-UQ and SEB-UA and -UD), the performance drop from removing nouns is larger than when syntactic information was removed. This performance drop is again caused by a tendency of the models to label the attack items as *incorrect*, which is visible in Table 6 across all data sets and for all content-word classes. This is plausible, as the student answers become very hard to interpret for humans, as well (cf. the sample item "there is a damaged" in Table 2).

Removing adverbs, and thereby negation expressed by "not", at first glance seems to be less damaging than removing adjectives and much less so than removing nouns and verbs. However, note that not all student answers contain adjectives and adverbs in the first place, so fewer changes are made to the test data. The fact that we still see a noticeable effect speaks to the semantic importance of these word classes in the student data. As for the syntactic attack items, it would be interesting to see whether humans and the models accept and reject the same attack items to gauge the importance of the word classes to human interpretation versus machine grading.

We also see that model performance strongly

	Beetle		SEB		
	UA	UQ	UA	UQ	UD
Test data	89.7	78.1	82.2	74.1	72.1
Repeat 1x	-9.1	-6.8	-5.2	-4.1	-4.8
Repeat 2x	**-14.3**	**-10.8**	**-9.6**	**-14.1**	**-8.5**
Rand. Short	–	97.5	–	–	95.5
Rand. Avg.	–	89.9	–	–	83.0
Rand. Long	–	**43.0**	–	–	**33.0**

Table 5: Exp.3: Testing the influence of answer length: Repeating answers from the SEB and Beetle test sets and randomly generated input sequences in three length bands; overall Δ Accuracy to test data (absolute Accuracy for randomly generated input).

deteriorates in the UA setting for both corpora (and for Beetle-UQ). We hypothesise that performance in the UA setting, where the model sees new answers to questions encountered in training, depends on keyword spotting more than in the UQ and UD settings. This is consistent with the well-established deterioration of performance on the unaltered test sets when moving away from the UA setting.

The model's remaining robustness towards removal of content words (after all, about 50% of correct answers are still recognised by the SEB RoBERTa model even if nouns are removed) may be rooted in RoBERTa's masked pre-training task which specifically teaches the model to reconstruct missing input.

Again, this result is reassuring in the context of ASAG: The model uses information from all groups of content words and is more likely to reject as *incorrect* inputs with some missing content words.

6 Experiment 3: Input Length

When we remove content words, we also shorten the input. At the same time, answer length is correlated with grade in the training data: *correct* Beetle answers have a median length of 54 characters (min: 3, max: 367), while *incorrect* answers are only 41 characters long in the median (min: 0, max: 256). For SEB, the numbers are 60 characters (min: 4, max: 532) for correct and 51 (min: 2, max: 413) for incorrect answers. Therefore it is relevant to ask whether the models pick up on this correlation.

We use two strategies to probe sensitivity to length while keeping the meaning of the utterances constant: One is to repeat the student answer, thus doubling or tripling the input in length without making a change to its meaning. The other is to randomly generate synthetic test items of different lengths (but without discernible meaning). More specifically, we build an attack set with synthetic length-controlled items generated randomly from the vocabulary of the Beetle-UQ and SEB-UD test sets (which are most different from the training data). We generate 200 attack items for each of three length classes: Short attack items are in the range between the minimum and median length of all relevant answers, average-length items are in the range of the first to third quartile and the length of long items is between the median and maximum lengths for the test sets. All of these attack items should be rejected as *incorrect* by the model since they are nonsensical (see Table 2 for sample items).

The results are shown in Table 5. Repeating each student answer once (doubling the answer length) or twice (tripling the answer length) clearly reduces model Accuracy. However, the result pattern at label level is inverted to the first two experiments (see Table 6). Now, Acc_{corr} increases for double- and triple-length answers, while Acc_{incorr} drops by more than 20 percentage points for all corpora. The model now *accepts* answers more easily the longer they are, although their content has not changed.

We turn to the length-controlled synthetic items to gauge the effect of submitting short items (which we could not probe in the replication attack without modifying answer meaning). The synthetic items show that the shortest inputs are in fact labelled *incorrect* even more frequently than the average length ones, so short items are somewhat at a disadvantage (the table shows absolute overall Accuracy). Long items are again labelled *correct* with very high probability (leading to low Accuracy, since all synthetic items are *incorrect*), and this effect is much stronger than the disadvantage for short items. This is a concerning finding for ASAG, since item length can easily be influenced by test-takers independent of their understanding of the task.

7 Word Deletion Attacks and Length

Given the results from Exp. 3, we need to reconsider our strategies and results in Exp. 1 and 2, where our attacks rely on deleting words from

	Beetle				SEB					
	UA		UQ		UA		UQ		UD	
	corr	incorr	corr	incorr	corr	incorr	corr	incorr	corr	incorr
Test data	88.6	90.5	62.5	89.5	75.1	87.6	86.6	74.7	73.4	74.5
Shuffle	-30.1	+5.5	-23.3	+4.4	-4.7	-2.9	+0.2	-3.2	-1.0	-3.0
Content only	-31.8	+5.3	-25.	+2.0	-10.7	+0.3	-19.3	+3.8	-7.8	+3.1
Both	**-50.6**	+6.8	-37.5	+7.8	-12.6	+2.3	-32.8	+1.2	-13.0	+3.5
No Adjs	-22.7	+2.2	-16.0	+3.1	-19.7	+2.6	-31.1	+9.8	-15.8	+4.5
No Advs	-14.8	+1.7	-8.1	+2.0	-1.3	-1.6	-22.1	+1.0	-6.1	+0.7
No Verbs	-13.0	+1.5	-7.3	+4.0	-11.6	-1.0	-20.5	+5.6	-21.8	+6.5
No Nouns	-30.6	+2.2	**-39.2**	+3.3	**-22.7**	+2.6	**-36.4**	+9.6	**-27.4**	+12.5
Repeat 1x	+2.3	-16.7	+11.9	-20.4	+7.9	-15.0	-4.5	-13.1	+5.4	-15.6
Repeat 2x	+6.3	**-28.1**	+16.2	**-30.6**	+13.8	**-27.0**	+1.6	**-23.0**	+10.6	**-25.8**

Table 6: Exp.1-3: Label-wise Δ Accuracy for different types of answer manipulation on the SEB and Beetle Unseen Answer (UA), Unseen Question (UQ) and Unseen Domain (UD) test sets (lowest result per column in boldface).

the student answers, thereby shortening them. Indeed, we found for both experiments that the models showed a tendency to reject the modified student answers, which could now also be explained by their shorter length. Recall, however, that the results for deleting non-content words in Exp. 1 were backed up by the shuffle attack, which preserves length.

In order to gauge the effect of length reduction in the word deletion attacks, we re-ran the experiments after replacing each non-content word rather than deleting it – e.g., nouns by "thing", verbs by "do", and non-content words by the particle "to" or, alternatively, any deleted word by "—" The attack items kept their length in this way.

Across all data sets and attacks, we found that replacing content words with valid lexical items generally further reduces model performance. Replacing words distorts the sentences even more strongly than just deleting them, because no guessing or filling in the blanks is possible (which is the task RoBERTa was trained to do during pre-training). There is very little difference between deleting words and replacing them by "—" placeholders, except that the extremely low performance for Beetle-QA in Exp. 2 is mitigated to something closer to the SEB performance. We therefore conclude that any length effect confounded with the deletion attacks is minor. This is supported by our observation that short answers are somewhat more likely to be graded `incorrect`, while long answers are much more likely `correct` (so the effect is smaller for shorter answers). Therefore, we believe that the results from Exp. 1 and 2 are not due to the length effects of the word deletion strategy but indeed to the loss of syntactic or semantic information from the student answers.

8 Conclusions

Across our three experiments, we have observed the performance of the RoBERTa$_{MNLI}$ model on the SAG task using the SEB and Beetle corpora.

A first, striking insight across all three experiments is that the size of the impact of our attacks differs strongly between corpora, while the general patterns stay the same. Removing syntactic information causes the models to label previously *correct* student answers as *incorrect*, but the model fine-tuned on SEB is much more forgiving and ready to retain the *correct* label than the Beetle model. The same is true for removing semantic information. This result shows how much of SAG model performance depends on the fine-tuning and test data and how misleading it can be to generalise insights from one data set to another.

Second, we saw clear evidence of RoBERTa's sensitivity to syntactic information in Exp. 1 – removing structural and word order clues causes the model to no longer accept originally correct student answers in many cases. This is plausible, since the student answers also become harder to interpret for humans. Model performance is not completely impaired, however, so slightly imperfect syntax will likely not preclude a *correct* grade.

Removing semantic information (even when ut-

terance length is preserved) in Exp. 2 is similar. When nouns are removed, only about 50% of all *correct* student answers are still recognised (for both Beetle and SEB) – understandably so, as the meaning of the answers is strongly distorted also to humans. Removing all other classes of content words showed similar effects; normalising the results by the number of affected student answers per manipulation in order to more accurately weight the influence of the word classes remains for future work.

Confirming that the RoBERTa models are sensitive to the syntax and semantics of student answers is reassuring in the context of ASAG. However, the strong length effect shown in Exp. 3 is very concerning for a SAG model, since it is clearly independent of content and could be used to gain an unfair advantage. Any serious use of the models as they stand should therefore install safeguards, for example a human review of all unusually long answers. In the long run, adversarial training (Madry et al., 2017) could be employed to mitigate the length effect.

While we carry out our experiments on one specific model (RoBERTa), the effects we find are likely to generalise to other Transformer-based ASAG models because they appear to stem from the training data and training regime. Further, the effect of insensitivity to word order (Hessel and Schofield, 2021) has been observed for another semantic task in previous work and the importance of semantic information (in terms of the choice of question-relevant lexical material) is also observed in (Ding et al., 2020).

In both Exp. 1 and 2, we were as yet unable to answer the question whether the attack items that were still accepted as *correct* by the models would also be acceptable to human graders or whether they are completely spurious. Comparing human and machine grades for these attack items is another interesting avenue for future work.

Acknowledgments

The authors acknowledge support by the state of Baden-Württemberg through the bwHPC high-performance computing infrastructure. N.W. was supported by AraCom IT Services AG for the duration of his Master's thesis.

References

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2022. Similarity-based content scoring - how to make S-BERT keep up with BERT. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 118–123.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25:60–117.

Leon Camus and Anna Filighera. 2020. Investigating transformers for automatic short answer grading. In *Artificial Intelligence in Education AIED*, Lecture Notes in Computer Science, pages 43–48.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuning Ding, Brian Riordan, Andrea Horbach, Aoife Cahill, and Torsten Zesch. 2020. Don't take "nswvtnvakgxpm" for an answer –the surprising vulnerability of automatic content scoring systems to adversarial input. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 882–892, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2020. Fooling automatic short answer grading systems. In *Artificial Intelligence in Education AIED*, Lecture Notes in Computer Science, pages 177–190.

Hadi Abdi Ghavidel, Amal Zouaq, and Michel C. Desmarais. 2020. Using bert and xlnet for the automatic short answer grading task. In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU)*, pages 58–67.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*, San Diego, CA, USA.

Michael Heilman and Nitin Madnani. 2013. ETS: Domain adaptation and stacking for short answer scoring. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 275–279, Atlanta, Georgia, USA. Association for Computational Linguistics.

Jack Hessel and Alexandra Schofield. 2021. How effective is BERT without word ordering? implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers)*, pages 204–211. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 8(9):1735–1780.

Boh Young Lee and Sang Keun Shin. 2020. Doable and practical: A validation study of classroom diagnostic tests. *Journal of Asia TEFL*, 17(2):349–362.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. In *Proceedings of ICLR*.

Niels Ott, Ramon Ziai, Michael Hahn, and Detmar Meurers. 2013. CoMeT: Integrating different levels of linguistic modeling for meaning assessment. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 608–616, Atlanta, Georgia, USA. Association for Computational Linguistics.

Laura Pérez-Mayos, Roberto Carlini, Miguel Ballesteros, and Leo Wanner. 2021. On the evolution of syntactic information encoded by BERT's contextualized representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2243–2258, Online. Association for Computational Linguistics.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168, Copenhagen, Denmark. Association for Computational Linguistics.

Swarnadeep Saha, Tejas I. Dhamecha, Smit Marvaniya, Renuka Sindhgatta, and Bikram Sengupta. 2018. Sentence level or token level features for automatic short answer grading?: Use both. In *Artificial Intelligence in Education*, pages 503–517.

Archana Sahu and Plaban Kumar Bhowmick. 2020. Feature engineering and ensemble-based approach for improving automatic short-answer grading performance. *IEEE Transactions on Learning Technologies*, 13(1):77–90.

Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. 2019. Improving short answer grading using transformer-based pre-training. In *Artificial Intelligence in Education AIED, Proceedings*, Lecture Notes in Computer Science, pages 469–481.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of the International Conference on Learning Representations*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Joshua Wilson, Yue Huang, Corey Palermo, Gaysha Beard, and Charles A. MacArthur. 2021. Automated feedback and automated scoring in the elementary grades: Usage, attitudes, and associations with writing outcomes in a districtwide implementation of mi write. *International Journal of Artificial Intelligence in Education*, 31(2):234–276.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3).