The Impacts of Unanswerable Questions on the Robustness of Machine Reading Comprehension Models

Son Tran, Uyen Le, Phong Do (collaborator)
Department of Computer Science, Denison University
Research Advisor: Dr. Matt Kretchmar

Introduction

Task: In Extractive Question Answering, an AI machine takes as input a passage of text and a question about that text. The machine attempts to extract part of the text from the passage which best answers the question. (see example below) Sometimes the question is unanswerable in the given passage.

Adversarial Attack: An adversary inserts purposely misleading text to cause the machine to respond to the question incorrectly. (see red text below)

Models: We use three state-of-the-art deep learning neural network models (BERT, RoBERTa, and SpanBert) to implement and test our algorithm.

Question Types	Question	Passage	Answer
Answerable	What is the name of the water body that is found to the east?	To the east is the Colorado Desert and the **Colorado River** at the border with Arizona, and the Mojave Desert at the border with the state of Nevada. To the south is the Mexico-United States border. **Sea is the name of the water body that is found to the west.**	**Colorado River**
Unanswerable	What desert is to the south near Arizona?	**To the east is the Colorado Desert and the Colorado River at the border with Arizona**, and the Mojave Desert at the border with the state of Nevada. To the south is the Mexico-United States border. **The desert of Edmonton desert is to the north near Burbank.**	

Hypothesis

We hypothesize that we can increase the performance of machines against adversarial attacks by first training them on unanswerable questions. This additional training causes the machines to learn deeper representations of the passage semantics. To test this hypothesis, we compare the performance of machines trained on answerable questions only (v1 models) vs those that receive additional training on unanswerable questions (v2 models).

Adversarial Performance

		Answerable		
		Original	Attacked	Decrease
BERT	v1	88.4	63.8	24.6
	v2	78.4	55.2	23.2
RoBERTa	v1	91.5	70.5	21.0
	v2	84.8	58.0	26.8
SpanBERT	v1	91.5	68.6	22.9
	v2	85.8	58.9	26.8

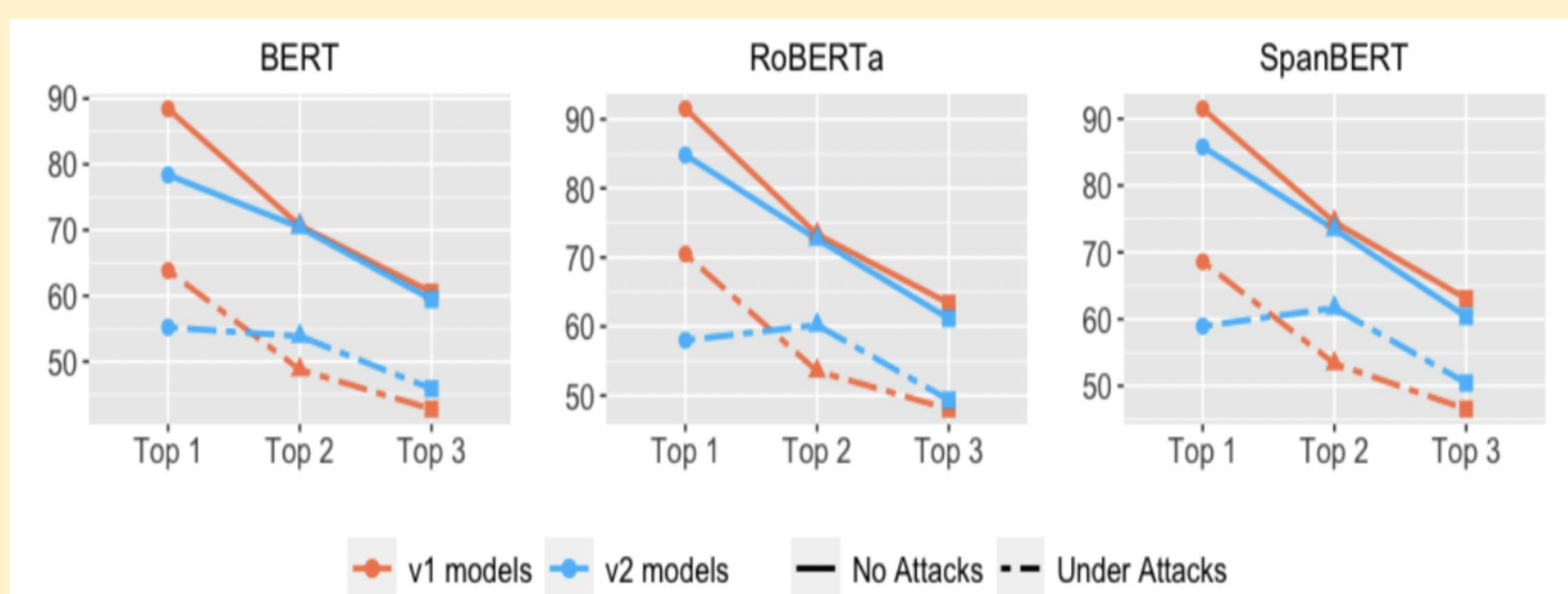
		I	C2I	C2U	C2C
BERT	v1	10.9	28.7	-	60.4
	v2	21.3	10.9	14.7	53.2
RoBERTa	v1	8.0	24.5	-	67.7
	v2	14.5	8.0	20.5	57.1
SpanBERT	v1	8.0	26.7	-	65.4
	v2	13.8	8.3	20.1	57.8

F1 scores measure the overlap between the prediction and ground truth answer.

Ground Truth: Denison University
- F1(Denison) = 0.5
- F1(Denison College) = 0.5
- F1(Denison in Granville) = 0.4

I: originally incorrect
C2I: correct to incorrect
C2U: correct to incorrectly unanswerable
C2C: correct to correct

Analysis of Top 3 Predictions



Force To Answer

Hypothesis: v2 models with additional training on unanswerable questions have the ability to perceive the attacks on answerable questions but fail to completely overcome them

Force To Answer technique:
Top 3 predictions by v2 model: ["no answer", "Colorado River", "Sea"]
The final answer will then be "**Colorado River**".

		Answerable		
		Original	Attacked	Decrease
BERT	v1	88.4	63.8	24.6
	v2	88.5	69.6	18.9
RoBERTa	v1	91.5	70.5	21.0
	v2	91.4	75.1	16.4
SpanBERT	v1	91.5	68.6	22.9
	v2	91.3	75.8	15.5

Conclusions

- In our first results shown in the top middle tables, v2 models do not appear to perform significantly better than v1 models. In fact, they often report that the question is now "unanswerable" even though a correct response appears in the passage.
- However, the correct responses of v2 models are often hidden as second-best answers (hidden robustness). This is shown in the three middle graphs where the performance of v2 models improves if we consider machine responses beyond the first one.
- By forcing v2 models to output a response to answerable questions, we leverage this hidden robustness to improve the performance of models to attacks on answerable questions. The table in the bottom shows approximately a 5 to 7 percent boost in accuracy of responses when attacked.

References

[1] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In EMNLP.

[2] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In EMNLP.

[3] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In ACL.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL.

[5] 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. ArXiv.

[6] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. In TACL.

Acknowledgements

This research was supported by Denison University through the generosity of
- The William G. and Mary Ellen Bowen Research Endowment

and
- The Laurie and David Hodgson Faculty Support Endowment