Adversarial Attack Strategies on Machine Reading Comprehension Models

香港中文大學
The Chinese University of Hong Kong

Romario Timothy Vaz, supervised by Dr Michael R Lyu and Shilin He

The Chinese University of Hong Kong

Motivations

Current Deep Learning models for Natural Language Processing are very powerful on a wide variety of tasks like Sentiment Analysis, Question Answering, Translation, and so on. However, many of these models are not robust, and can be easily tricked with adversarial examples. Despite the amazing advances in Deep Learning and NLP, it cannot be stated that these models are genuinely understanding text, due to some almost embarrassing errors that they make on seemingly simple tasks.

Objectives

We explore ways to trick the predictions made by models on reading comprehension tasks, using the Stanford Question Answering Dataset (SQuAD) and present the strategies that were successful.

Methods

The methods are tested on the R-NET and BERT models.

Paraphrasing context paragraphs:

- Automated and manual paraphrasing
- Both simple and complex paraphrase styles

Paraphrasing questions:

- Manual paraphrase to ensure that the syntax of the question is as different as possible from the syntax of the sentence that contains the answer.
- Replace question words (who, what, why, which, where, how) while retaining the original meaning, wherever it is possible to do so.

Adding irrelevant information to the paragraphs

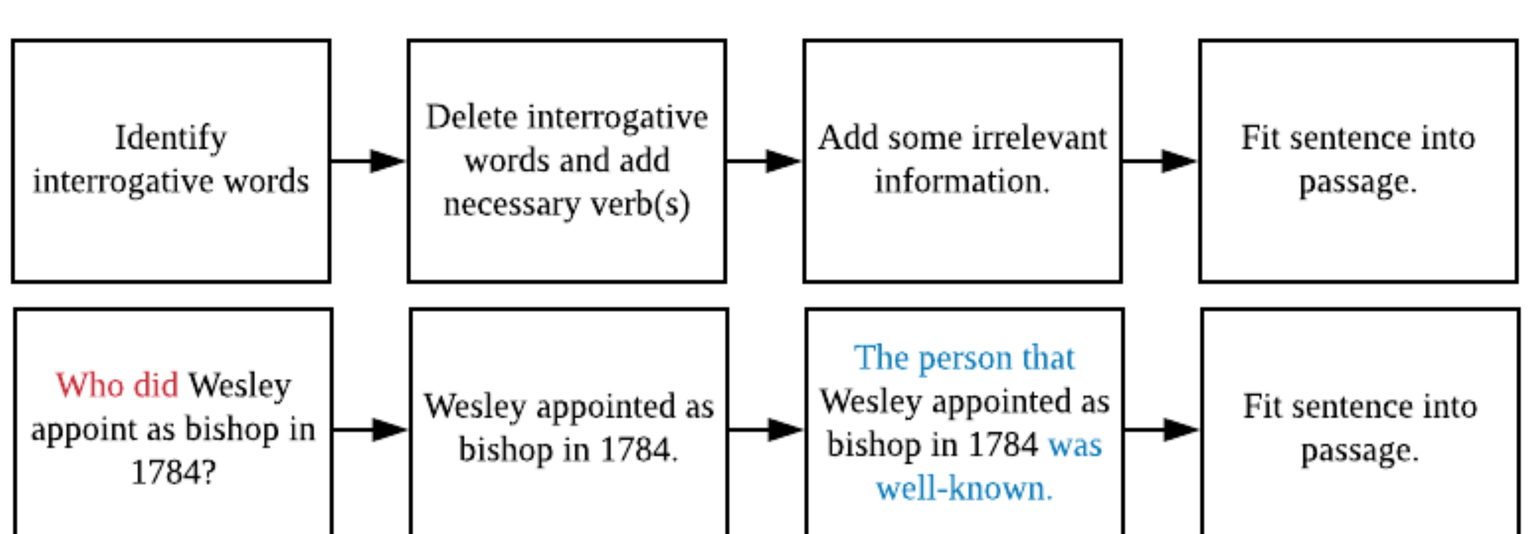
Question word substitution patterns:

- The question is transformed into a declarative sentence, and some irrelevant information is added in order to trick the model into thinking that this irrelevant information contains the answer.
- Done both manually and in an automated manner using simple if-else based rules and POS-tagging.

AddExtraneous:

- Manually transform a question from the paragraph into a statement and insert into the context paragraph using this method:
  1. Convert the question to a declarative sentence with an incorrect answer, such that the modified statement directly contradicts what has been stated in the paragraph.
  2. Qualify the sentence using some distractor clauses.
  3. Add a clause to the end of the sentence that nullifies the verb associated with the contradicting information (optional, only required if the sentence still contradicts the context paragraph).
  4. Add this final sentence to the paragraph such that it fits cohesively.

Figure 1: Steps to add irrelevant information to paragraphs using question word substitutions, with an example question.



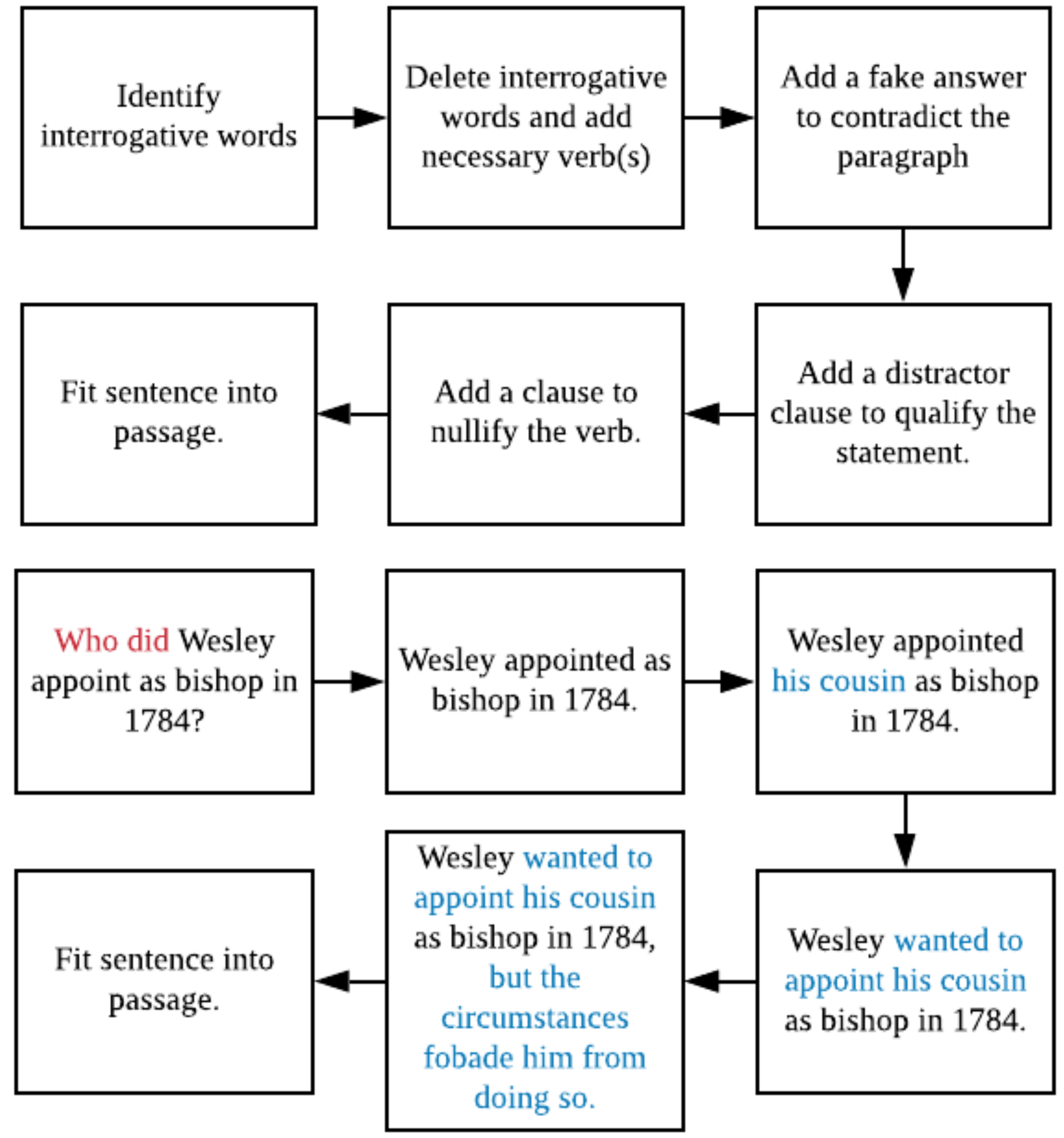


Figure 2: Steps to add irrelevant information to paragraphs using the AddExtraneous method, with an example

Figure 3: An example of tricking the BERT model using irrelevant information to the paragraphs using question word patterns

Paragraph: The place that Martin Luther went to school is important to discuss . In 1501, at the age of 19, he entered the University of Erfurt, which he later described as a beerhouse and whorehouse. He was made to wake at four every morning for what has been described as "a day of rote learning and often wearying spiritual exercises." He received his Master's degree in 1505.
Question: Where did Martin Luther go to school?
Original Prediction: University of Erfurt
Prediction under adversary: important

Figure 4: An example of AddExtraneous tricking the BERT model:

Paragraph: Martin Luther wanted to go to Harvard University, but the competition was too intense. In 1501, at the age of 19, he entered the University of Erfurt, which he later described as a beerhouse and whorehouse. He was made to wake at four every morning for what has been described as "a day of rote learning and often wearying spiritual exercises." He received his Master's degree in 1505.
Question: Where did Martin Luther go to school?
Original Prediction: University of Erfurt
Prediction under adversary: Harvard University

Results

Table 1:Model Performance when paraphrasing the context paragraphs

Model	Original Sample	Samples with paraphrased context
R-NET	77.23	75.05
BERT	85.53	84.38

Table 2: Model Performance when paraphrasing the questions

Model	Original EM Score	Adversarial EM Score	Original F1 Score	Adversarial F1 Score
R-NET	73.671	62.308	84.718	72.256
BERT	74.117	74.231	84.759	85.084

Table 3: Model Performance when adding irrelevant information using question word substitution patterns

Model	Original EM Score	Adversarial EM Score	Original F1 Score	Adversarial F1 Score
R-NET	70.60	61.30	78.55	67.71
BERT	80.90	68.96	88.28	76.78

Table 4: Model Performance when adding irrelevant information using the AddExtraneous method

Model	Original EM Score	Adversarial EM Score	Original F1 Score	Adversarial F1 Score
R-NET	68.213	36.167	79.626	43.929
BERT	73.592	45.750	87.321	52.018

Conclusions

- Word meanings are reasonably captured by the embedding layers.
- Pattern-matching nature of R-NET's simpler attention-based recurrent architecture is a likely reason for the success of question paraphrasing on the model.
- Overly stable nature of the networks is most likely responsible for the accuracy drop on examples where irrelevant information is added using question word patterns.
- The inability of the models to perceive subtle connotations within nullifying clauses results in the accuracy drop on the AddExtraneous adversarial examples (example in Figure).
- Current NLP models for Machine Comprehension are unable to differentiate between referring to a fact and stating the fact outright, which further explains their vulnerability to the AddExtraneous adversaries.

References

[1] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. CoRR, abs/1707.07328, 2017.
[2] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. CoRR, abs/1606.05250, 2016.
[3] Yicheng Wang and Mohit Bansal. Robust machine comprehension models via adversarial training. CoRR, abs/1804.06473, 2018.
[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[13] Natural Language Computing Group. R-net: Machine reading comprehension with self-matching networks. May 2017.
[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.