BERT & Family Eat Word Salad: Experiments with Text Understanding

Ashim Gupta, Giorgi Kvernadze, Vivek Srikumar

THE UNIVERSITY OF UTAH

AAAI 2021

1. Key Points

This paper is about ***analyzing BERT-like models using ill-formed, unnatural examples.***

We propose ***nine destructive transformations*** for this analysis.

The punchline: Models cannot differentiate natural text from unnatural text, let alone understanding natural text.

Hypothesis: Happens because models learn spurious correlations instead of understanding complexities of a natural language.

Actionable Insight: Simple mitigation strategies work. Best ones make use of unnatural examples at training.

2. *Failed* Natural Language Inference

Premise: In reviewing this history, it's important to make some crucial distinctions.

Hypothesis: Making certain distinctions is imperative in looking back on the past. (ENTAILMENT with 99% Probability)

Alphabetically Sorted Hypothesis: back certain distinctions imperative in is looking making on past the . (ENTAILMENT with 97% Probability)

- Sorting the hypothesis makes it meaningless to humans.
- But model's prediction remains same with high confidence.
- Trend consistently observed in Multi-NLI evaluation

3. Destructive Transformations

`Large changes in input aimed at destroying label determining information from input`

Prediction should not remain same after making large changes to the input.

Types of Destructive Transformations

1. **Lexical Overlap Based Transformations**
   To diagnose sensitivity to word order of input.
2. **Gradient Based Transformations**
   To study effect of removing, replacing, repeating words in input.
3. **Statistical Correlation Based Transformations**
   To expose statistical biases learned by a model.

Original Input: The men are experts when it comes to electronics .

Lexical overlap: are comes electronics experts it men the to when .

Gradient based: the best are ora suit can comes to beans.

Statistical: two men are looking at in computer park .

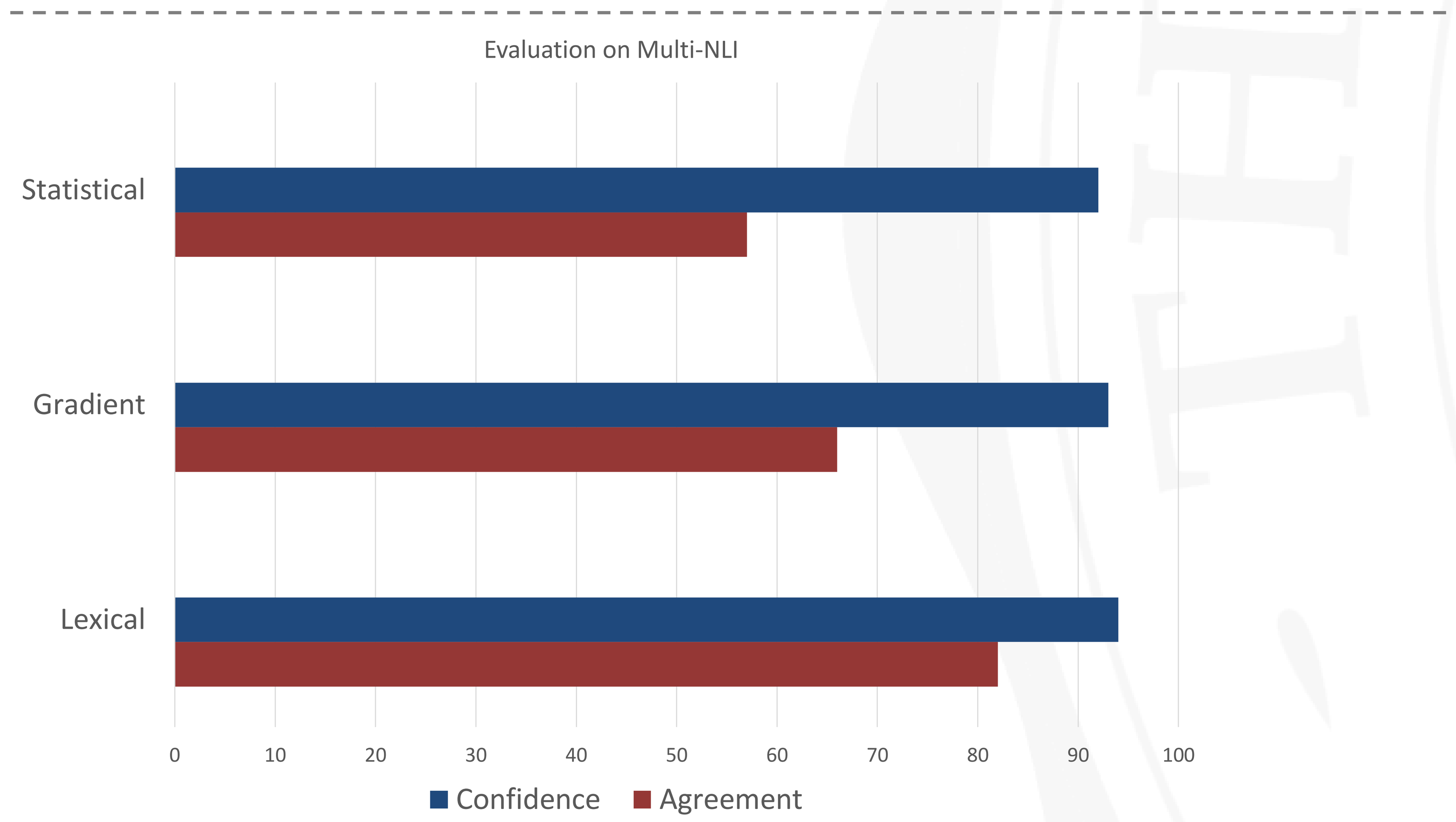
Generated examples are *meaningless* to humans (validated using Mechanical Turk)

4. Results: Models understand language?

Response to unnatural examples measured using two metrics:

- `Agreement Score`: % of retained predictions
- `Average Confidence`: Based on probability of predicted label

Results shown on MultiNLI. More results In paper.



Observation: Large percentage of predictions remain unchanged even after making inputs meaningless.

Do models really understand natural language?

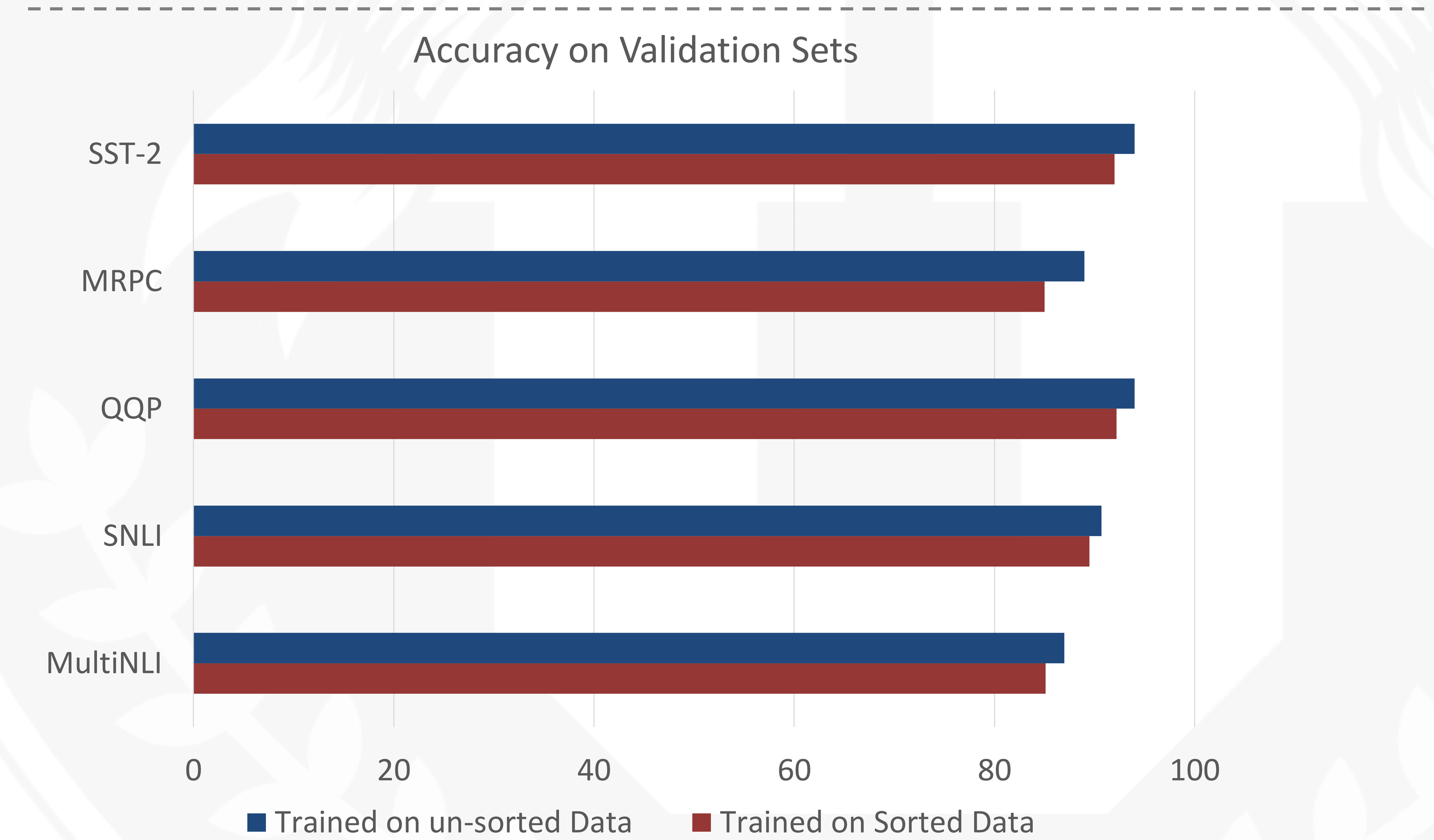
No. Models even have hard time *differentiating* natural text from unnatural one.

5. Models learn spurious correlations?

Our hypothesis: Such behavior occurs because models learn spurious (unwanted) correlations from the dataset.

Following experiment confirms the hypothesis:

- Make all examples in training set meaningless (by sorting etc.)
- Train model on meaningless examples
- Evaluate on well-formed (natural) examples



Accuracy almost same for both models

Observation: Sorting the words has no impact on what model learns and predicts.

Model learns correlations based on word identity without considering word order information.

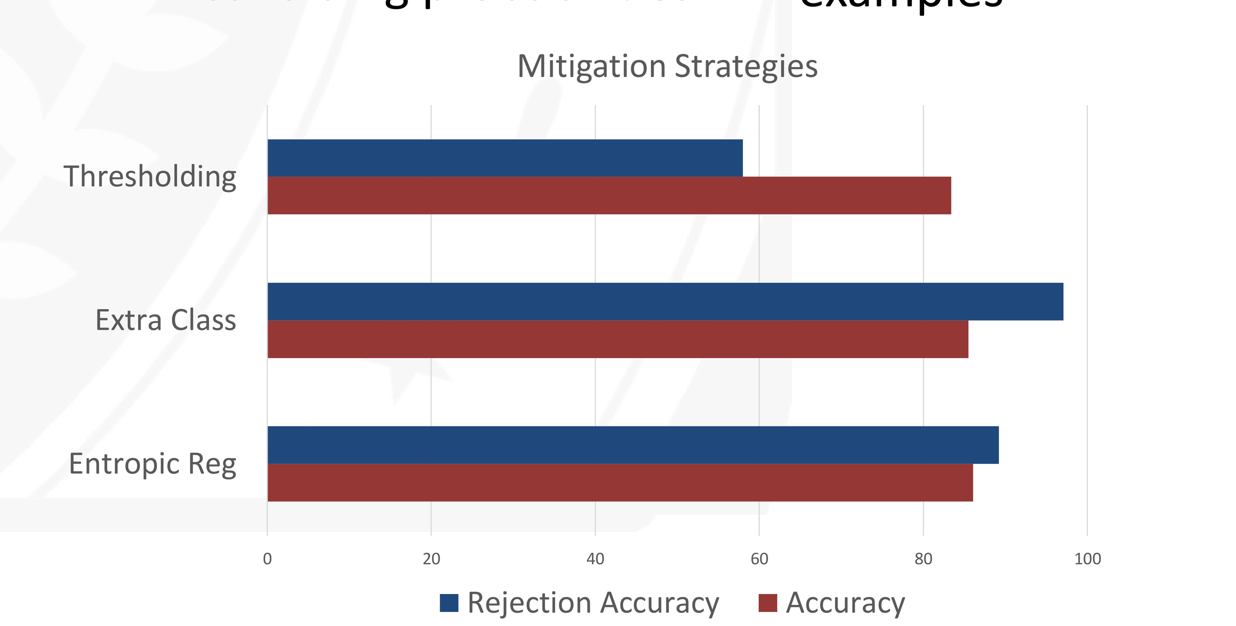
6. Mitigation Strategies

Three mitigation strategies:

- Entropic regularization
- Invalid as an extra class
- Thresholding probabilities

Two metrics:

- Accuracy on original examples
- Rejection Accuracy on unnatural examples



Adding meaningless examples generated using destructive transformations as extra class works best.