Making Language Models Robust Against Negation

MohammadHossein Rezaei Eduardo Blanco

University of Arizona

mhrezaei@arizona.edu

NAACL 2025 (main)

1. Overview and Pre-training Tasks

The computer screen stayed blank. It didn't display any images.

Next Sentence Prediction (NSP)

The computer screen stayed blank. + It didn't display any images. ✓ A coherent continuation

Reverse polarity

The computer screen stayed blank. + It displayed some images. ✗ Not a coherent continuation

Next Sentence Polarity Prediction (NSPP)

The computer screen stayed blank. *Does next sentence have negation?* ✓ n't is a negation cue

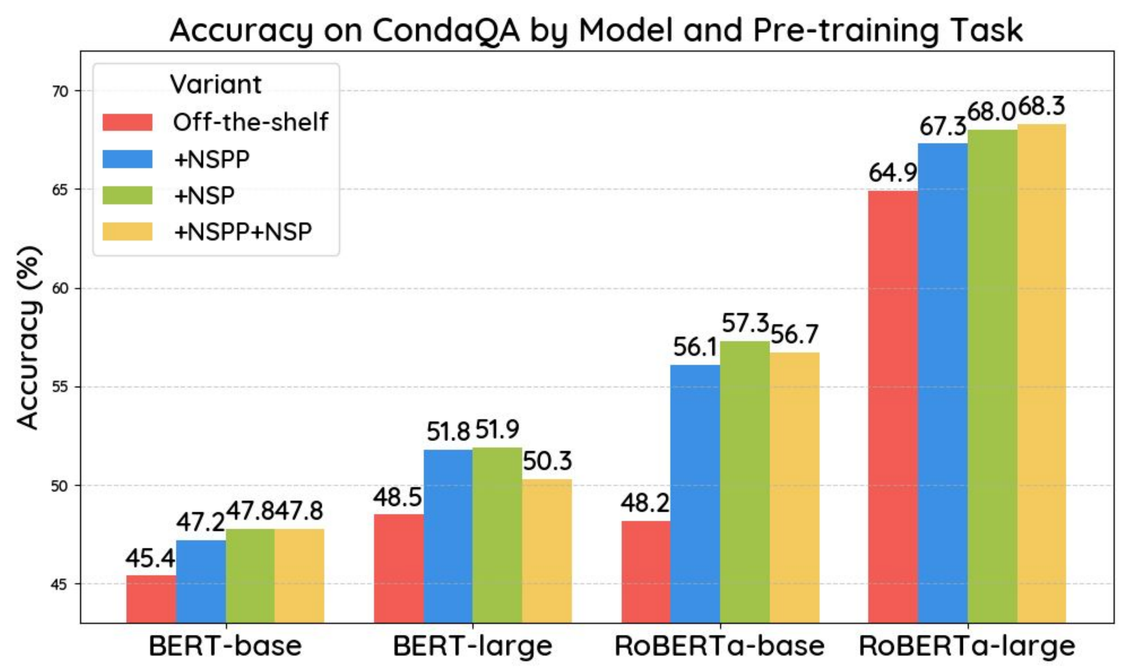
2. Introduction

- **Motivation**:
  - Negation is present in 25% of English sentences
  - LLMs struggle when negation is involved
  - BERT predicts "dog" in both of the following:
    - A beagle is a type of [MASK]
    - A beagle is **not** a type of [MASK]
- **Contributions**:
  - **Two** novel self-supervised **pre-training tasks**
  - A dataset (~6.4M samples) for these tasks
  - Experimental results with BERT and RoBERTa on **CondaQA** and **eight other benchmarks**

3. Pipeline

- **Data Collection**
  - Collect consecutive sentence pairs
  - Filter pairs containing negation cues
  - Collect an equal number of affirmative pairs
- **Reversing Polarity**
  - **Definition**:
    - Remove negation cues if present
    - Add negation to affirmative sentences
  - Use **linguistic rules** instead of LLMs, which resist factual contradiction due to safety guardrails
- **Further Pre-training**
  - Further pre-train **BERT** and **RoBERTa** on:
    - Next Sentence Prediction (NSP)
    - Next Sentence Polarity Prediction (NSPP)
    - Jointly on both tasks
  - Train on varying data sizes until val loss plateaus

4. Experimental Results



Results on **CondaQA**, the largest question-answering corpus that requires **reasoning over negation**. Further pre-training on any of our tasks **statistically significantly** outperforms off-the-shelf LM

We also report performance improvements on **NLI** corpora (RTE, SNLI, MNLI), **NLU** cropora (QNLI, WiC, WSC), **LAMA**, and **LAMA-neg** benchmarks

5. Conclusion

- **Takeaways**
  - Further pre-training on our tasks is **beneficial**
  - The **NSP** task is better than the NSPP task
  - Further pre-training on both tasks is not beneficial
- **Future work**
  - Expand to more and larger models
  - Use more negation cues in pre-training
  - Further pre-train on larger corpora