Project LIT - Outline

Erik Imgrund, Niklas Loeser, Andre Trump

February 17, 2023

1 Task Description

What is the aim - in view of the specific phenomenon under consideration? Identify linguistic biases of current models for NLI. Improve LMs by removing biases from the training process.

Hypotheses Current language models are biased for NLI in the zero-shot and fine-tuned settings. Removing biases from the dataset improves results in a less biased model. A less biased model results in worse accuracy.

2 Method

How will we approach this aim? First as a baseline we will test

Which methods will/did you apply?

How to probe or fine-tune for your task?

3 Models and Data Sets

Select suitable data and resources Additional resources to improve learning?
Use or create specific tests for targeted evaluation (e.g., foiling, masking, ...)
Dataset statistics: classes, distributions, ...

4 Experiments

4.1 Evaluation Metrics

4.2 Specific for Probing and Fine-Tuning

Models to use

Experiment Configuration

4.3 Experimental Settings

Data Variations, Experiment Variation

Baselines and Model Variants

5 Analysis

5.1 Confusion Analysis

5.2 Visualizations

5.3 Interpretation Methods