

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 the bias and accuracy of a model in a zero-shot and fine-tuned manner. The bias

is tested by calculating plausibility [1] and faithfulness [2] to the explanations provided as well as manual inspection of handpicked examples. By analyzing the baseline models and in particular their explanations we aim to identify linguistic biases in their predictions. Using those identified linguistic biases we want to remove heavily biased examples for the training dataset and train a model on the de-biased data. The resulting model is compared to the baseline models.

Which methods will we apply? A pretrained BERT model [3] is used for the zero-shot tasks as well as basis for the fine-tuning tasks. Integrated Gradients [4], LIME [5] and Partition SHAP Values [6] are used as explanation methods. Two approaches to removing the biased examples are tested [Dont take the easy way out]: The biased data can be reweighted based on how biased it is such that biased examples influence the fine-tuning less or a more depending on how biased the example is. A different approach is based on ensembling the fine-tuning model during training with a biased model so that the model does not need to learn those.

How to probe or fine-tune for your task? For the zero-shot task two methods are tried. The next-sentence-prediction head of the pretrained BERT model is used or discourse relation markers between the premise and hypothesis are predicted using a pretrained DisSent [7] model. Each discourse marker is assigned to either entailment, neutral or contradiction and the most probable discourse marker is used to assign a prediction of the model.

For fine-tuning the premise and hypothesis are both fed into a LM separated by a marker token and a text classification head is trained based on the embedding obtained from the model to predict the class of the combined text. That is either entailment, neutral or contradiction.

3 Models and Data Sets

Select suitable data and resources To test the bias of the model we plan to use e-SNLI [1] for large-scale calculation of plausability and faithfulness of the explanations the models provide. For testing the accuracy and manually analyzing the bias of the models we plan to use SICK [2], as it is less biased than SNLI and MultiNLI. To fine-tune the model we use MultiNLI, as it is less biased than SNLI but is sufficiently large.

Additional resources to improve learning? Additionally the training set of SICK can be used to introduce data with less bias. Additional tests could be introduced to test specific biases identified by manual analysis.

4 Experiments

To evaluate the performance of the model on natural language inference we use the F1-Score and Matthews correlation coefficient [3]. We evaluate the faithfulness and plausability with the same metrics as described in [ferret paper].

The models used for probing are PLMs based on DisSent, which also provide the pretrained basis for the finetuned models.