Inherent Disagreements in NLI

Premise: A homeless man being observed by a man in business attire.
Hypothesis: Two men are sleeping in a hotel.

Data: CommitmentBank

Premise: Meg realized she’d been a complete fool. She could have said it differently.
Hypothesis: Carolyn had borrowed a book from Clare and wanted to return it.

Error Analysis

Premise: ‘She was about to tell him that was his own stupid fault and that she wasn’t here to wait on him – particularly since he had proved to be so inhospitable. But she bit back the words. Perhaps if she made herself useful he might decide she could stay - for a while at least just until she got something else sorted out.
Hypothesis: she could stay

Models: Artificial Annotators

P(y|x) = \text{softmax}(W_{a} \tanh(W_{e} \{e; n; c\}))

Five Baselines

“Always 0”: Always predict Disagreement.
CBOW: Each item is represented as the average of its tokens’ GLOVE vectors.

Heuristic baseline: Linguistics-driven rules, e.g., conditional environment discriminates for disagreement items.

Vanilla BERT: Straightforwardly predict among 4 finer-grained NLI labels.
Joint BERT: Two BERT models are jointly trained. One identifies disagreement item; the other one carries out systematic inference.

Baseline and AAs overall performance on CB dev and test sets, and F1 scores of each class on the test set (average of 10 runs). * indicates a statistically significant difference (t-test, p<0.01).

AAs do worse on Neutral items due to lack of Neutral training data.

The best performance (~66%) is still far from achieving robust NLU.

Models achieve good results when there is enough data.

AAs learn linguistic patterns and context-dependent inference better.

A method capturing accurately # of modes in the annotation distribution would lead to a better model.