Identifying inherent disagreement in natural language inference

Xinliang Frederick Zhang and Marie-Catherine de Marneffe

The Ohio State University

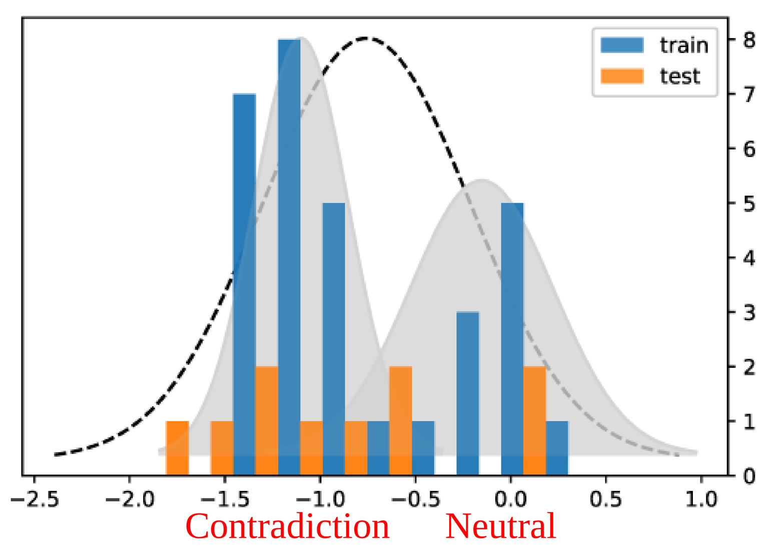
THE OHIO STATE UNIVERSITY

Inherent Disagreements in NLI

Pavlick and Kwiatkowski (2019)

Premise: A homeless man being observed by a man in business attire.

Hypothesis: Two men are sleeping in a hotel.



Data: CommitmentBank

de Marneffe et al (2019)

Premise:

Meg realized she'd been a complete fool. She could have said it differently. If she'd said Carolyn had borrowed a book from Clare and wanted to return it they 'd have given her the address.

Environment (Conditional)

Hypothesis: Carolyn had borrowed a book from Clare.

Disagreement [3, 3, 3, 2, 0, -3, -3, -3]

Entail Neutral Contradict

Finer-Grained Labels to Capture Disagreement

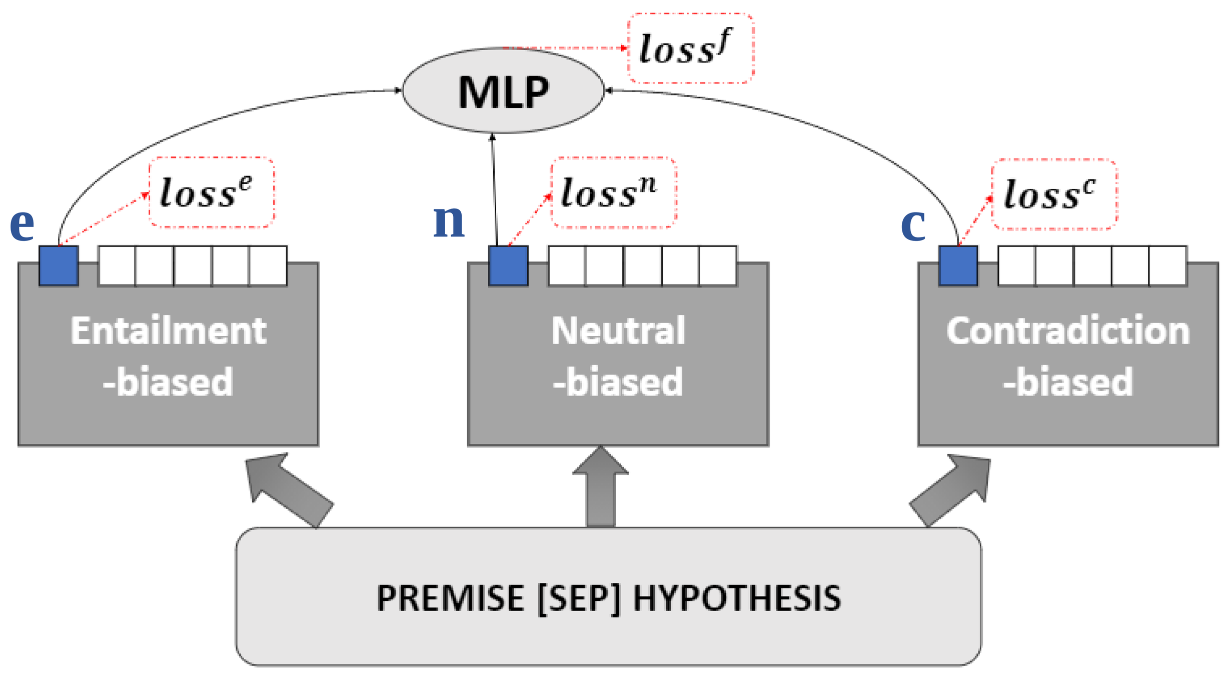
Entailment: 80% of annotations ∈ [1,3] OR $\sigma \leq 1$ and $\mu > 1$.

Neutral: 80% of annotations is 0 OR $\sigma \leq 1$ and $-0.5 \leq \mu \leq 0.5$.

Contradiction: 80% of annotations ∈ [-3, -1] OR $\sigma \leq 1$ and $\mu < -1$.

Disagreement: Items that do not fall in any of the three categories above.

Model: Artificial Annotators



$$P(y|\mathbf{x}) = \mathrm{softmax}(\mathbf{W_s}\tanh(\mathbf{W_t}[\mathbf{e};\mathbf{n};\mathbf{c}]))$$

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.

	Dev		Test					
	Acc.	F1	Acc.	F1	Entail	Neutral	Contradict	Disagree
Always 0	55.00	39.03	45.42	28.37	0.00	0.00	0.00	62.46
CBOW	55.25	40.54	45.09	28.37	0.00	0.00	0.69	62.17
Heuristic	65.00	62.08	54.17	50.60	22.54	**52.94**	64.46	58.20
Vanilla BERT	63.71	63.54	62.50	61.93	59.26	49.64	69.09	61.93
Joint BERT	64.47	64.28	62.61	62.07	59.77	47.27	67.36	63.21
AAs (ours)	**65.15**	**64.41**	**65.60***	**64.97***	**61.07**	51.27	**70.89**	**66.49***

Baselines 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.

	negation	modal	conditional	question	negR
Heuristic	51.29	48.02	37.69	44.64	54.16
V. BERT	60.91	73.98	44.84	53.02	61.91
J. BERT	60.94	73.95	46.02	51.68	63.67
AAs	**65.96**	**80.18**	**48.05**	**54.95**	**68.00**

F1 for CB test set under embedding environments and "I don't know/believe/think" ("negR").

AAs perform better across the board. Models achieve good results when there is enough data.

Correct inference by Heuristic?	Yes (130)		No (110)	
	Acc.	F1	Acc.	F1
V. BERT	80.00	80.45	41.51	42.48
J. BERT	79.74	80.04	42.73	44.15
AAs	**84.37**	**84.85**	**46.97**	**48.75**

BERT-based models performance on test items correctly predicted by (Yes) vs. items missed (No) by linguistic rules.

AAs learn linguistic patterns and context-dependent inference better.

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

Heuristics: D V. BERT: D

J. BERT: D AAs: N {N, N, N}

Neutral [3, 0, 0, 0, 0, 0, 0, 0, 0]

Premise: B: Yeah, and EDS is very particular about this, hair cuts, A:Wow. B: I mean it was like you can't have, you know, such and such facial hair, no beards, you know, and just really detailed. A: A: I don't know that that would be a good environment to work in.

Hypothesis: that would be a good environment to work in

Heuristics: C V. BERT: C

J. BERT: D AAs: C {C, C, C}

Disagreement [2, 0, 0, 0, 0, -1, -2, -3]

Predict \ Gold	E	N	C	D	Total
E	37	2	0	13	52
N	1	10	0	3	14
C	0	0	34	13	47
D	20	7	20	80	127
Total	58	19	54	109	240

Confusion matrix of AAs for the test set.

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