Identifying inherent disagreement in natural language inference

Xinliang Frederick Zhang and Marie-Catherine de Marneffe

The Ohio State University

NAACL 2021

Natural Language Inference (NLI)

THE OHIO STATE UNIVERSITY

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

Hypothesis: Two men are sleeping in a hotel.

Contradiction

Neutral

Entailment

Inherent Disagreements in Human Textual Inferences (Pavlick and Kwiatkowski, 2019)

Natural Language Inference (NLI)

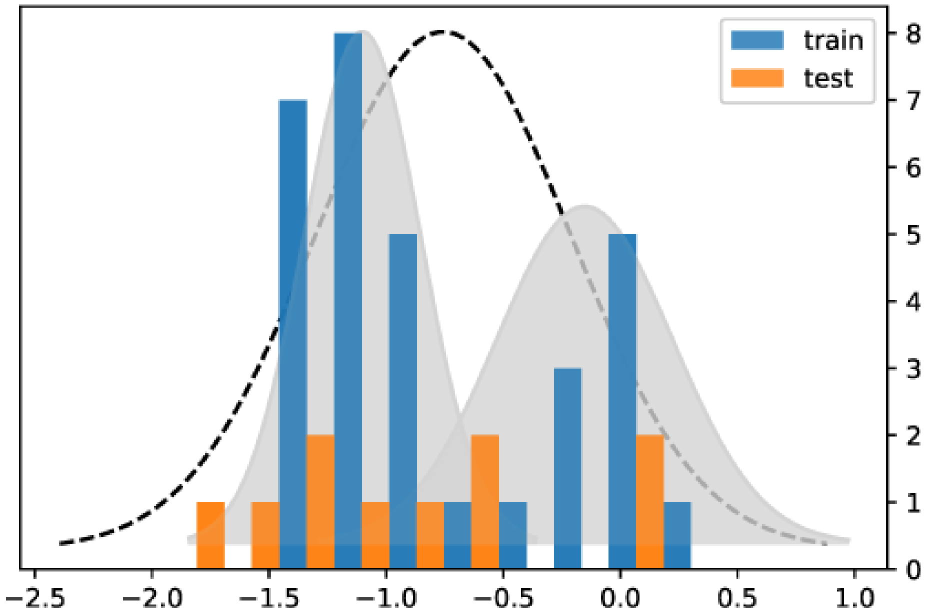
THE OHIO STATE UNIVERSITY

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

Hypothesis: Two men are sleeping in a hotel.



Contradiction Neutral Entailment



Contradiction Neutral


Inherent Disagreements in Human Textual Inferences (Pavlick and Kwiatkowski, 2019)

Data: CommitmentBank

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: I don't know that that would be a good environment to work in.

Hypothesis: that would be a good environment to work in

Label? [2, 0, 0, 0, 0, -1, -2, -3]

The CommitmentBank: Investigating projection in naturally occurring discourse (de Marneffe et al., 2019)

Data: CommitmentBank

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: I don't know that that would be a good environment to work in.

Hypothesis: that would be a good environment to work in

Label? [2, 0, 0, 0, -1, -2, -3]

The CommitmentBank: Investigating projection in naturally occurring discourse (de Marneffe et al., 2019)

Finer-grained labels for NLI

The Ohio State University

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: I don't know that that would be a good environment to work in.

Hypothesis: that would be a good environment to work in

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

Entailment

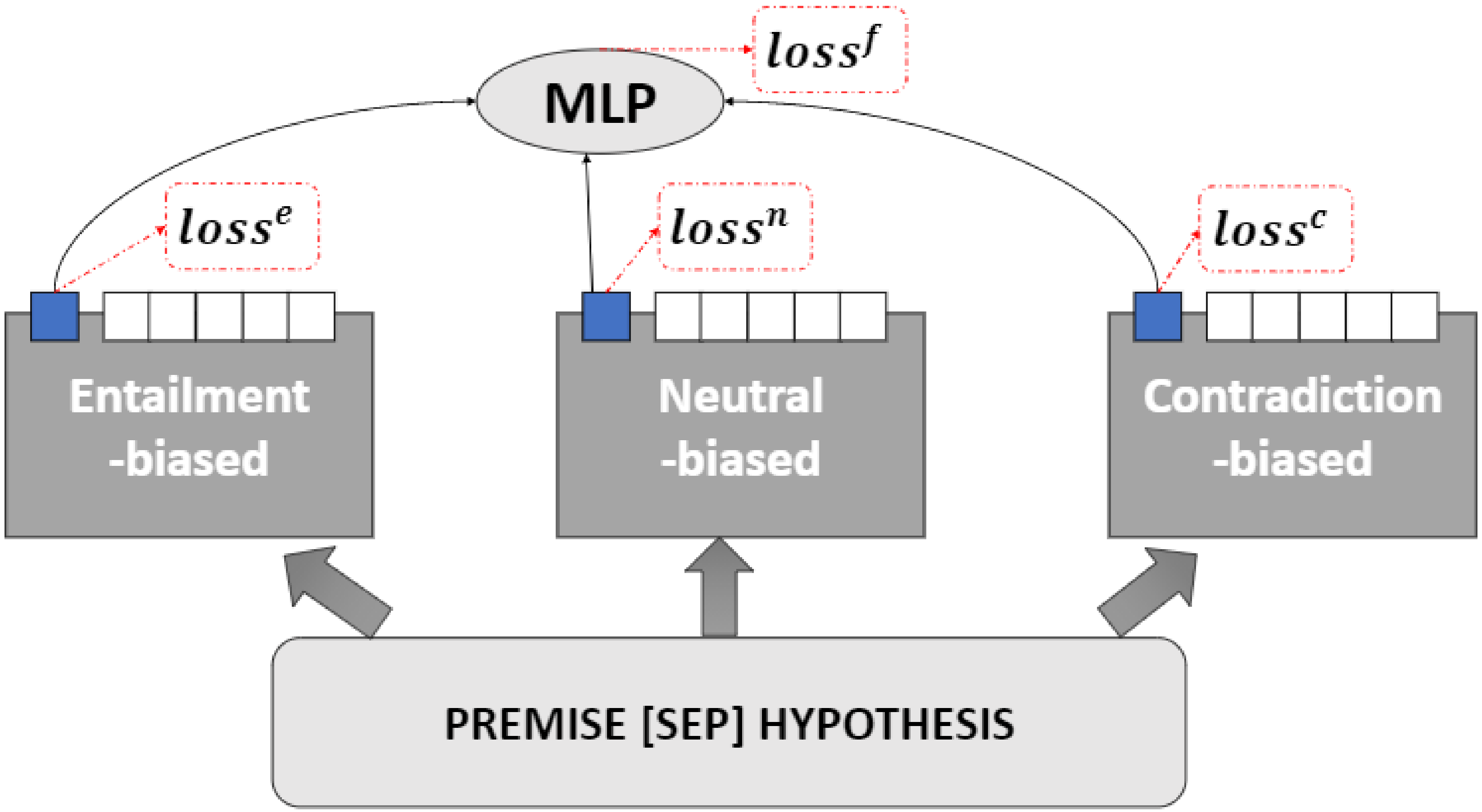
Neutral

Contradiction

Disagreement

The CommitmentBank: Investigating projection in naturally occurring discourse (de Marneffe et al., 2019)

Model: Artificial Annotators (AAs)



$$P(y|\mathbf{x}) = \mathrm{softmax}(\mathbf{W}_\mathbf{s} \tanh(\mathbf{W}_\mathbf{t}[\mathbf{e};\mathbf{n};\mathbf{c}]))$$

AAs perform better across the board

	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 learn linguistic patterns and context-dependent inference better

THE OHIO STATE UNIVERSITY

Correct inference by Heuristic?	Correctly predicted (130)		Missed (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 vs. items missed by linguistic rules.

Error analysis

The Ohio State University

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: 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]

Towards robust NLI

THE OHIO STATE UNIVERSITY

Our Artificial Annotators are a start in this direction but still far from succeeding (~ 66%).

A method which captures accurately the number of modes in the annotation distribution would lead to a better model.

The Ohio State University

Thanks!

Code is available at:

https://github.com/FrederickXZhang/FgNLI

Contact: zhang.9975@osu.edu