Evaluating BERT for natural language inference: A case study on the CommitmentBank

Nanjiang Jiang, Marie-Catherine de Marneffe

The Ohio State University

Work funded by NSF

BERT SOTA-ed natural language inference, but can it infer speaker commitment in English?

Premise: *A: Okay. So Frank, what, uh, type of, uh, budget do you or your family have? B: Well, uh I don't know that we really have a budget.*

Environment Clause-embedding Verb

Hypothesis: *he and his family really have a budget.*

Label: -1.82 → Contradiction

de Marneffe et al (2019)

Annotation Artifacts?

Hypothesis: no artifacts, unlike most NLI datasets

Premise: top PMI for 1/4-grams Gururuangan et al (2019)

Entailment		Neutral		Contradiction	
unigram	4-gram	unigram	4-gram	unigram	4-gram
g	might have known that	clever	Do you think that	mean	. B : I
herself	. You could say	Nicky	. Do you think	five	don ' t think
perhaps	Mr . An ##tar	pressure	you know , what	care	B : Uh ,
wrong	. An ##tar '	Base	' ' I hope	guy	: I don '
notice	An ##tar ' s	radio	' I hope you	jury	, I mean ,

Potential biases in CommitmentBank (CB) reflect linguistic generalizations

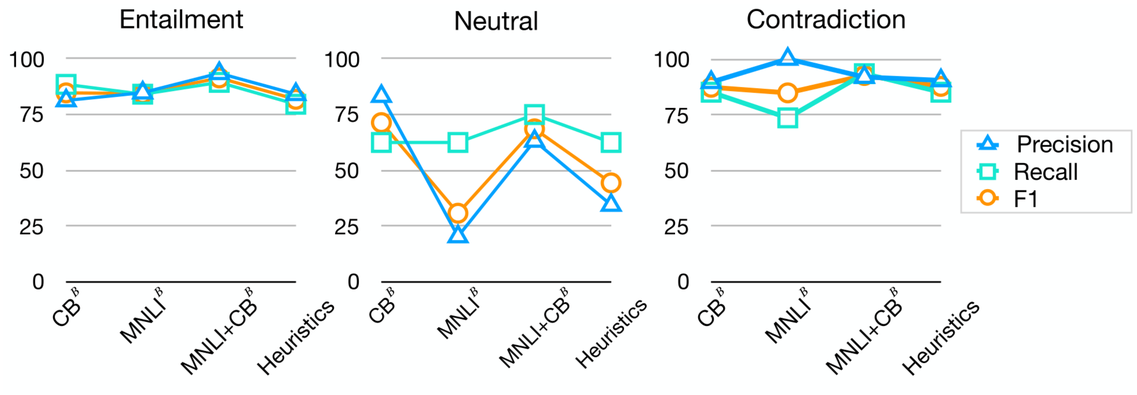
Models

Fine-tuned BERT vs.
Linguistically-motivated Heuristics

- negation, neg-raising → contradictions
- modals, factive verbs → entailments
- all others → neutral

	CB		MultiNLI	
	Acc.	F1	Acc.	F1
CBOW	69.2	47.6	71.2	71.2
MNLI^B	77.6	66.7	**83.6**	**83.6**
Heuristics	81.2	71.3	-	-
CB^B	85.2	81.2	41.1	30.6
MNLI+CB^B	**91.2**	**85.3**	72.3	74.4
Human	98.9	95.8	92.0/92.8	-

Accuracy/F1 on CB test set and MultiNLI dev set



F1 on CB test set by class

- **Heuristics and MNLI^B are strong baselines**
- **Tuning on CB hurts original MNLI performance**
- **All models do worst on neutral**

What does BERT learn from fine-tuning with CB?

	Items obeying linguistic generalizations (203)	Items requiring pragmatic inferences (47)
CB^B	**88.7**	55.8
MNLI^B	68.9	55.7
CB+MNLI^B	**89.4**	**68.5**

BERT trained on CB are better at identifying heuristics
All models struggle with pragmatic inferences

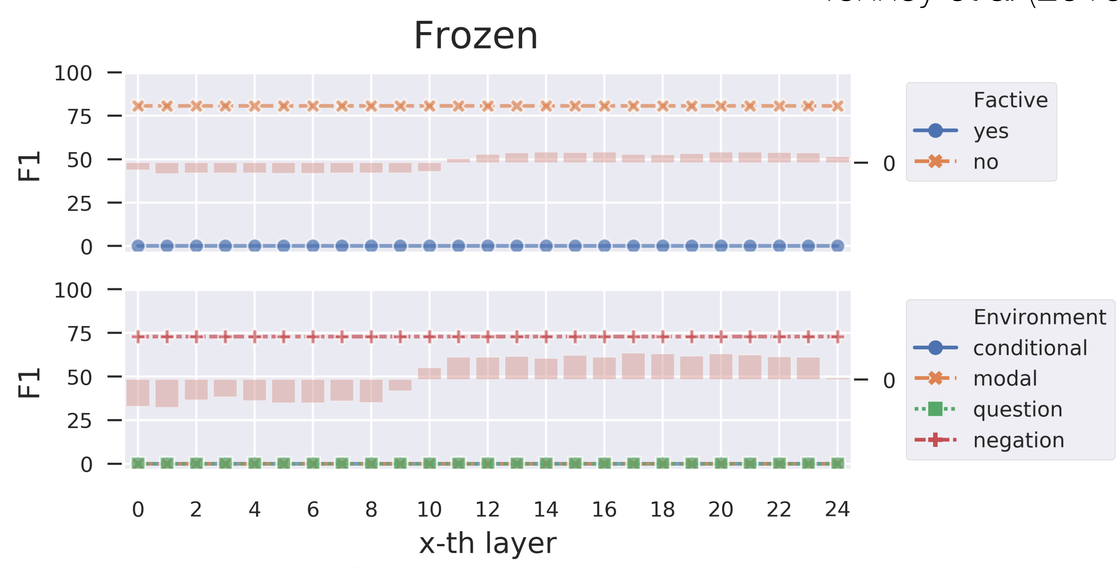
Premise: *B: Yeah, it's called VCX or something like that. Also called Delta Clipper, which is a decent name for something like that. A: Wow. Well, I don't know. you think you'd, uh, go up in space if you had a chance?*
Hypothesis: *speaker B would* go up in space if he had a chance
Heuristics: neutral
MNLI+CB^B: contradiction
Gold: neutral

Premise: ``*Rather a long shot, wasn't it? Twenty years?*'' *How do you know the baby was born here?*
Hypothesis: *the baby was born here*
Heuristics: entailment
MNLI+CB^B: entailment
Gold: neutral

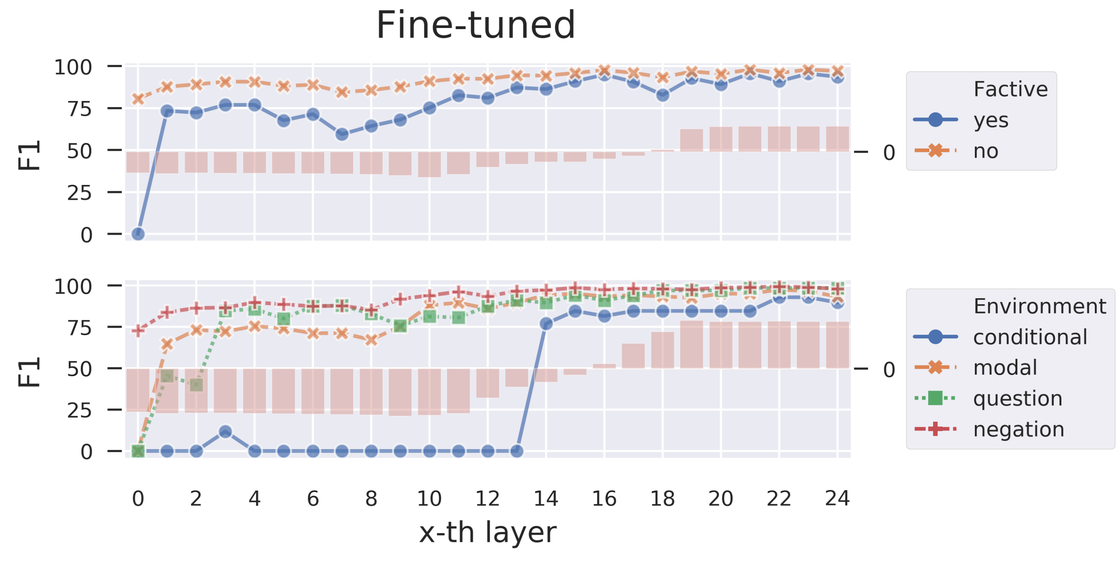
Examples where CB+MNLI^B still fails

Feature Probing

Does BERT representation encode the linguistic features? Tenney et al (2019)



Frozen MNLI+CB^B does not encode features in heuristics



Factivity is processed later than Environment, in line with language acquisition Hacquard et al (2019)