When does data augmentation improve generalization?

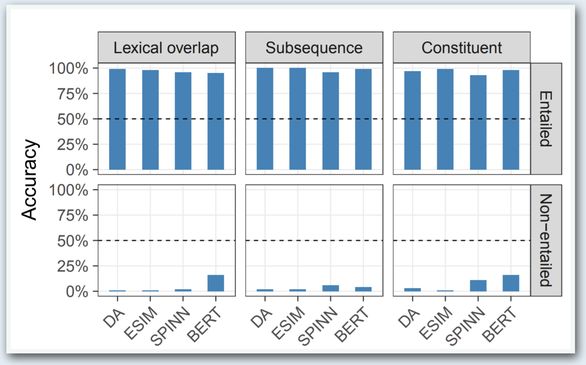
Rohan Jha, Charles Lovering, and Ellie Pavlick

BROWN

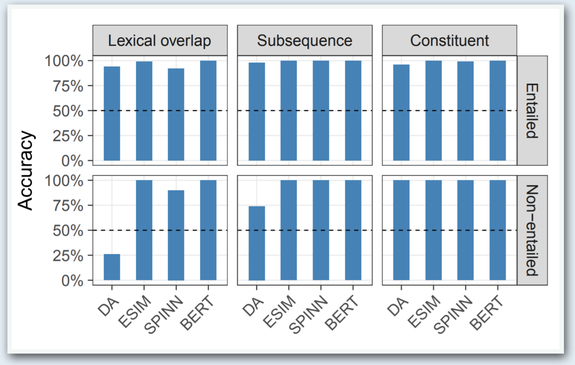
Language Understanding and Representation Laboratory

Motivation

BERT trained on MNLI.



Without data augmentation.



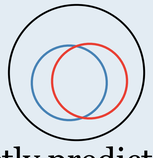
With data augmentation.

Questions

1. How many *counter-examples* are sufficient to prevent a learner from adopting a heuristic? How does the difficulty of representing the "right" feature affect the learner's preference for adopting heuristics?
2. Do we need both positive and negative *counter-examples*?
3. Does increasing the training size for the same number of counter-examples reduce or increase the test error?

Set Up

RNN trained on synthetic data.



The **distractor property** is easily learned but imperfectly predicts the label.

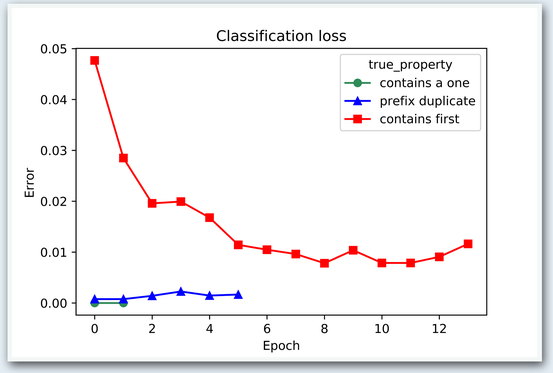
- Example: Contains A

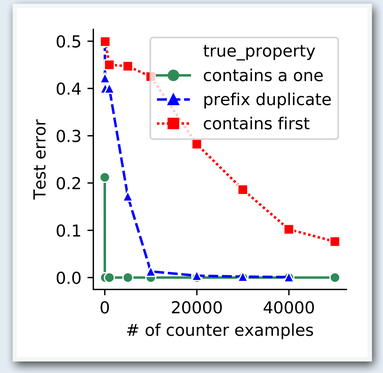
The **true property** is harder to learn but exactly corresponds with the label.

- Example: Contains an adjacent duplicate

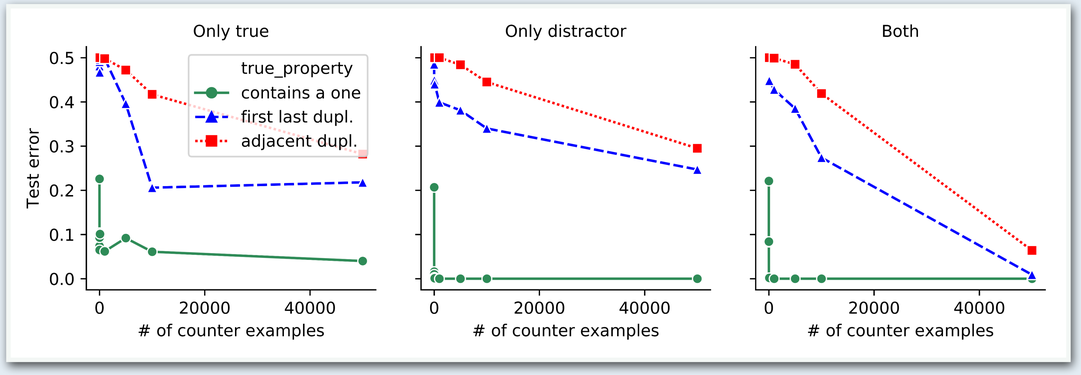
Input	Label	Example	Interpretation, in terms of NLI
both	1	ACCHG	Lexical overlap and entailed
true	1	DLNMM	Entailed but no lexical overlap
neither	0	KPTOT	Neither entailed nor lexical overlap
distractor	0	PLTUA	Lexical overlap but not entailed

Results

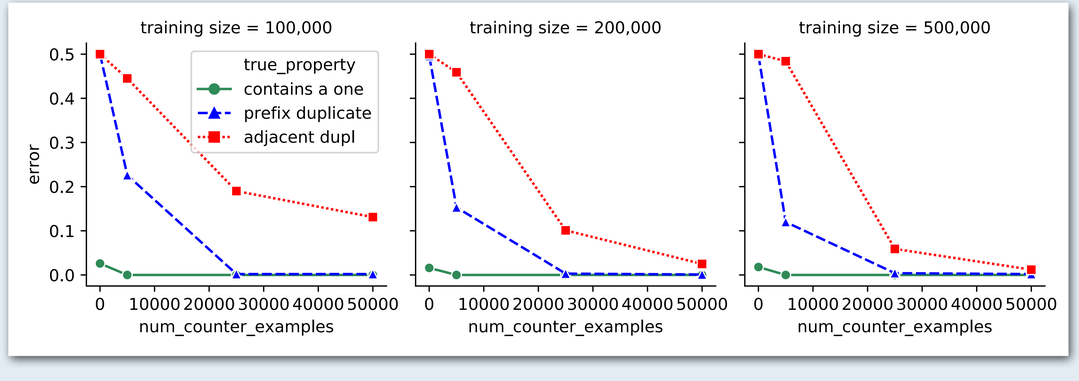




1. Hardness matters for the number of necessary *counter-examples*.



2. Both classes of *counter-examples* are helpful in data augmentation.



3. Increasing the training size for the same number of *counter-examples* reduces the test error.