Annotation Artifacts in Natural Language Inference Data

Gururangan et all. 2018

Livy Real
Grupo de Linguística Computacional (GLiC)
Universidade de São Paulo
12 de novembro de 2018

NLI task

Given a pair of sentences, a premise p and a hypothesis h, the goal is to determine whether or not p semantically entails h.

Entailment h is definitely true given p

Neutral h might be true given p

Contradiction h is definitely not true given p

Large datasets required

SNLI (Bowman et al. (2015))

MultiNLI (Williams et al., 2018)

Method: crowd workers are presented with a premise p drawn from some corpus (e.g., image captions), and are required to generate three new sentences (hypotheses) based on p: neutral / entailment / contradiction

The point

hypotheses generated by this crowdsourcing process contain artifacts that can help a classifier detect the correct class without ever observing the premise

Premise A woman selling bamboo sticks talking to two men on a loading dock.

Entailment There are **at least three people** on a loading dock.

Neutral A woman is selling bamboo sticks t**o help provide for her family.**

Contradiction A woman is **not** taking money for any of her sticks.

The experiment

- fastText (Joulin et al., 2017), an off-the-shelf text classifier that models text as a bag of words and bigrams, to predict the entailment label of the hypothesis without seeing the premise
-

Model	SNLI	MultiNLI	
		Matched	Mismatched
majority class	34.3	35.4	35.2
`fastText`	**67.0**	**53.9**	**52.3**

“Artifacts” analysis

Entailment. - generic words such as animal, instrument, and outdoors, replace exact numbers with approximates (some, at least, various), and to remove explicit gender. entailed hypotheses are generally shorter (8.8% of entailed hypotheses in SNLI are fully contained within their premise, while only 0.2% of neutrals and contradictions)

Neutral. Modifiers (tall, sad, popular) and superlatives (first, favorite, most) are affiliated with the neutral class, add cause and purpose clauses. neutral hypotheses tend to be long

Contradiction. Negation words such as nobody, no, never and nothing; sleeping contradicts any activity, and naked contradicts any description of clothing

Re-evaluation

Model	SNLI			MultiNLI Matched			MultiNLI Mismatched		
	Full	*Hard*	*Easy*	*Full*	*Hard*	*Easy*	*Full*	*Hard*	*Easy*
DAM	84.7	69.4	92.4	72.0	55.8	85.3	72.1	56.2	85.7
ESIM	85.8	71.3	92.6	74.1	59.3	86.2	73.1	58.9	85.2
DIIN	86.5	72.7	93.4	77.0	64.1	87.6	76.5	64.4	86.8

Avoiding artifacts

“a better solution might not eliminate the artifacts altogether, but rather **balance** them across labels. Future strategies for reducing annotation artifacts might involve experimenting with the prompts or **training** given to crowd workers, e.g., to encourage a wide range of strategies, or incorporating baseline or adversarial systems that **flag examples that appear to use over-represented heuristics**.”

Conclusions

1. Many datasets contain annotation artifacts (SICK, SNLI, MultiNLI, etc)
2. Supervised models leverage annotation artifacts - supervised models will exploit shortcuts in the data for gaming the benchmark, if such exist
3. Annotation artifacts inflate model performance
4. Hard / Easy SNLI/MultiNLI
5. Encourage development of additional challenging benchmarks that expose the true performance levels of state-of-the-art NLI models