Annotation Artifacts in Natural Language Inference Data

Suchin Gururangan *           Swabha Swayamdipta *
Omer Levy           Roy Schwartz           Samuel R. Bowman           Noah A. Smith

* equal contribution

**Natural Language Inference**

Inter-annotator agreement # annotators something about guidelines?

- There are animals outdoors.
- Some puppies are running to catch a stick.
- The pets are sitting on a couch.
- Two dogs are running through a field.

Premise

SNLI (Bowman et. al, 2015)
MultiNLI (Williams et. al., 2017)

**Annotation Artifacts**

- Generalization
- Shortening
- Modifiers
- Purpose Clauses
- Negation
- Cats

**What are NLI models really learning?**

- Easy
- Hard

NLI models learn lexical relationships rather than entailment semantics.

**Takeaways**

1. These results are consistent with numerous findings of issues with NLP datasets.
2. Annotation artifacts can be addressed by improving annotation protocols to correct for these biases.
3. We release the Hard benchmarks on Kaggle and the SNLI website.