Annotation Artifacts in Natural Language Inference Data

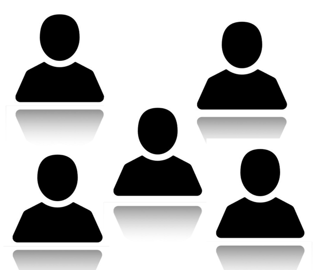
Suchin Gururangan* Swabha Swayamdipta*
Omer Levy Roy Schwartz Samuel R. Bowman Noah A. Smith
* equal contribution

UWNLP
Carnegie Mellon University Language Technologies Institute
AI2
NYU

Natural Language Inference

Two dogs are running through a field.

Premise



There are animals outdoors.
Entailment

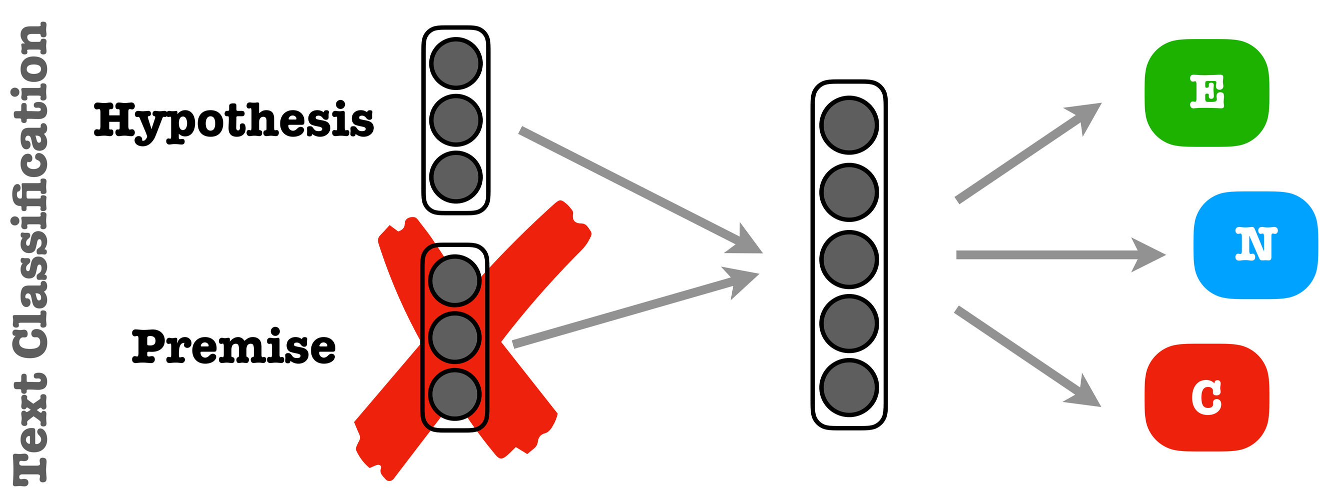
Some puppies are running to catch a stick.
Neutral

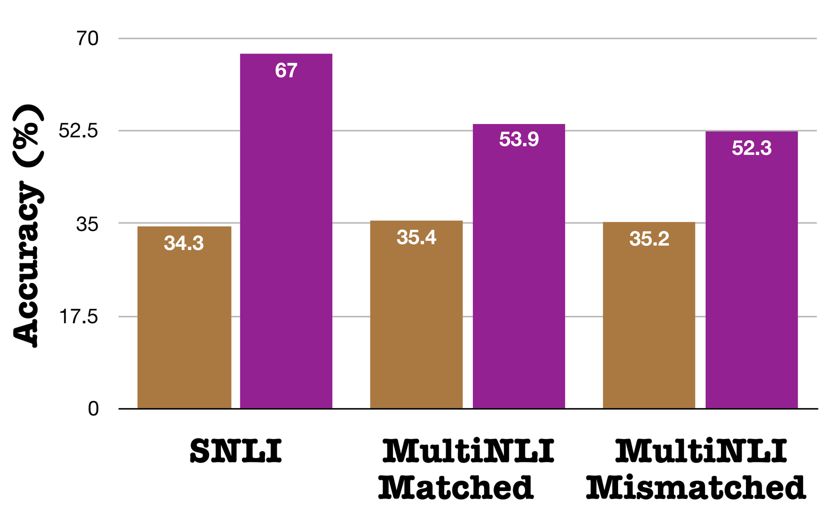
The pets are sitting on a couch.
Contradiction

SNLI (Bowman et. al, 2015) 570 K
MultiNLI (Williams et. al., 2017) 433 K

SNLI premises are Flickr captions.
MultiNLI premises are collected from diverse genre.
Hypotheses are crowdsource-generated.

Kicking out Premises...





Over 50% of NLI examples can be correctly classified **without** ever observing the premise!

Entailment Artifacts

Some men and boys are playing frisbee in a grassy area.

Premise

Generalization

People play frisbee outdoors.
Entailment

A person in a red shirt is mowing the grass with a green riding mower.

Premise

Shortening

A person in red is cutting the grass on a riding mower.
Entailment

Neutral Artifacts

A middle-aged man works under the engine of a train on rail tracks.

Premise

Modifiers

A man is doing work on a black Amtrack train.
Neutral

A group of female athletes are huddled together and excited.

Premise

Purpose Clauses

They are huddled together because they are working together.
Neutral

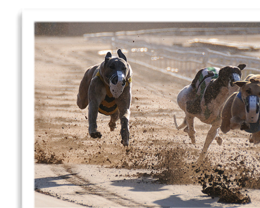
Contradiction Artifacts

Older man with white hair and a red cap painting the golden gate bridge on the shore with the golden gate bridge in the distance.

Premise

Negation

Nobody wears a cap.
Contradiction



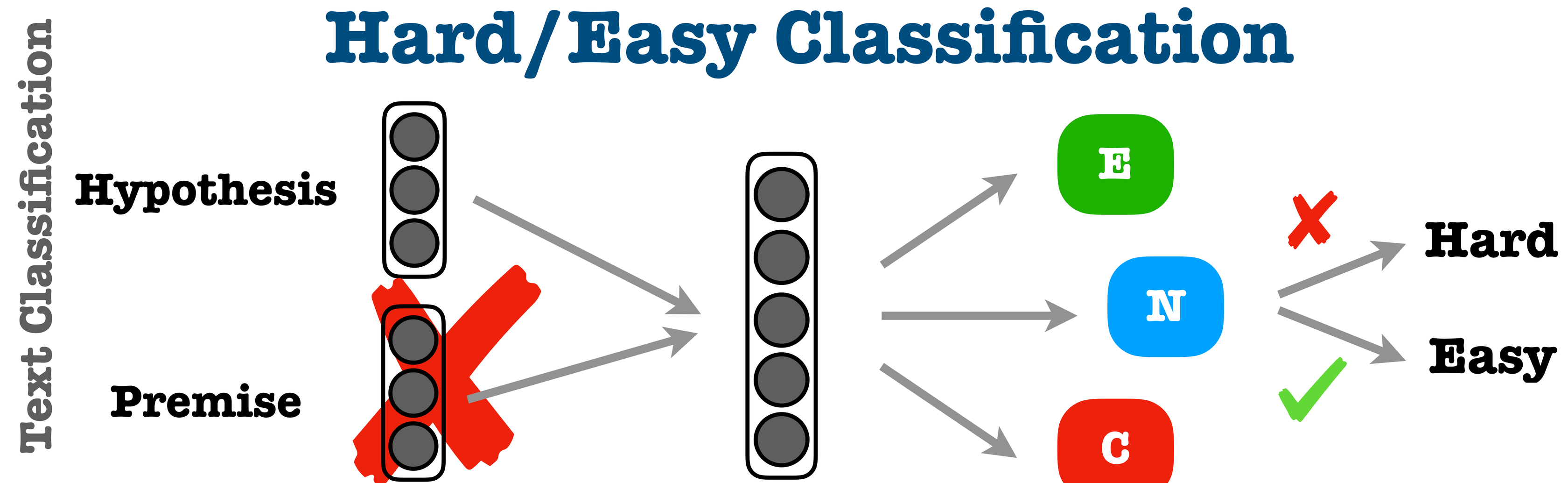
Three dogs racing on racetrack.

Premise

Cats!

Three cats race on a track.
Contradiction

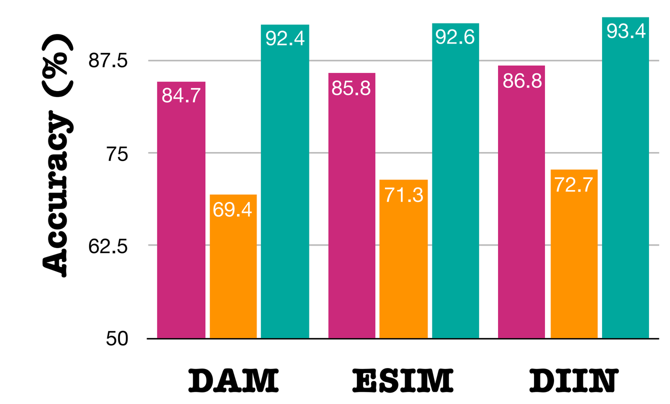
Hard/Easy Classification



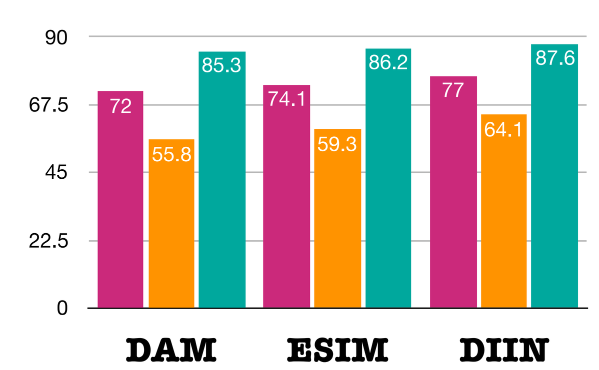
What are NLI models really learning?

DAM - Decomposable Attention Model (Parikh et. al. 2016)
ESIM - Enhanced Sequential Inference Model (Chen et. al., 2017)
DIIN - Densely Interactive Inference Network (Gong et. al. 2018)

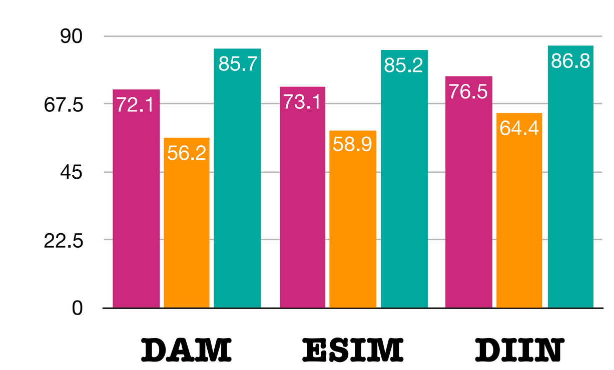
SNLI



MultiNLI Matched



MultiNLI Mismatched



NLI models learn from lexical cues rather than entailment semantics.

Takeaways

- ★ Results consistent with numerous findings of issues with NLP datasets like ROC (Cai et. al., 2017) and VQA (Agrawal et. al., 2016).
- ★ Annotation artifacts could be addressed by improving annotation protocols preemptively to correct for common biases.

Resources

★ **Hard benchmarks for MultiNLI:**
www.kaggle.com/c/multinli-matched-open-hard-evaluation/

★ **Hard benchmark for SNLI:**
www.nlp.stanford.edu/projects/snli/snli_1.0_test_hard.jsonl