Natural Language Inference: Dataset artifacts and adversarial testing

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding

Hypothesis-only baselines

- In his project for this course (2016), Leonid Keselman observed that hypothesis-only models are strong.
- Other groups have since further supported this (Poliak et al. 2018; Gururangan et al. 2018; Tsuchiya 2018; Belinkov et al. 2019)
- SNLI hypothesis-only baselines typically 65–70% vs. chance at 33%
- Likely due to artifacts:
  - Specific claims are likely to be premises in entailment cases.
  - General claims are likely to be hypotheses in entailment pairs.
  - Specific claims are more likely to lead to contradiction.

NLI dataset artifacts

1. **Artifact**: A dataset bias that would make a system susceptible to adversarial attack even if the bias is linguistically motivated.
2. Tricky example: negated hypotheses signal contradiction
   - Linguistically motivated: negation is our best way of establishing relevant contradictions.
   - An artifact because we would curate a dataset in which negation correlated with the other labels but led to no human confusion.

Known artifacts in SNLI and MultiNLI

- These datasets contain words whose appearance nearly perfectly correlates with specific labels [1, 2].
- Entailment hypotheses over-represent general and approximating words [2].
- Neutral hypotheses often introduce modifiers [2].
- Contradiction hypotheses over-represent negation [1, 2].
- Neutral hypotheses tend to be longer [2].

1 = Poliak et al. 2018, 2 = Gururangan et al. 2018

Artifacts in other tasks

- Visual Question Answering: Kafle and Kanan 2017; Chen et al. 2020
- Story Completion: Schwartz et al. 2017
- Reading Comprehension/Question Answering: Kaushik and Lipton 2018
- Stance Detection: Schiller et al. 2020
- Fact Verification: Schuster et al. 2019

Adversarial testing

Premise	Relation	Hypothesis
A turtle danced.	entails	A turtle moved.
Every reptile danced.	neutral	A turtle ate.
Some turtles walk.	contradicts	No turtles move.

Adversarial testing

	Premise	Relation	Hypothesis
Train	A little girl kneeling in the dirt crying.	entails	A little girl is very sad.
Adversarial		entails	A little girl is very unhappy.

Glockner et al. 2018

Adversarial testing

	Premise	Relation	Hypothesis
Train	A **woman** is pulling a **child** on a sled in the snow.	entails	A child is sitting on a sled in the snow.
Adversarial	A **child** is pulling a **woman** on a sled in the snow.	neutral	

Nie et al. 2019

References I

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. ArXiv:2003.06576.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of NLI models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6867–6874.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2020. Stance detection benchmark: How robust is your stance detection? ArXiv:2001.01565.

Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. ArXiv:1908.05267.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. Story cloze task: UW NLP system. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 52–55, Valencia, Spain. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.