Natural Language Inference: SNLI, MultiNLI, and Adversarial NLI

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding

SNLI

1. Bowman et al. 2015
2. All the premises are image captions from the Flickr30K corpus (Young et al. 2014).
3. All the hypotheses were written by crowdworkers.
4. Some of the sentences reflect stereotypes (Rudinger et al. 2017).
5. 550,152 train examples; 10K dev; 10K test
6. Mean length in tokens:
   - Premise: 14.1
   - Hypothesis: 8.3
7. Clause-types:
   - Premise S-rooted: 74%
   - Hypothesis S-rooted: 88.9%
8. Vocab size: 37,026
9. 56,951 examples validated by four additional annotators.
   - 58.3% examples with unanimous gold label
   - 91.2% of gold labels match the author's label
   - 0.70 overall Fleiss kappa
10. Leaderboard: `https://nlp.stanford.edu/projects/snli/`

Crowdsourcing methods

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** an **false** description of the photo.

Photo caption **A little boy in an apron helps his mother cook.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption.

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be.

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."*

Write a sentence which contradicts the caption.

Problems (optional) *If something is wrong with the caption that makes it difficult to understand, do your best above and let us know here.*

Examples

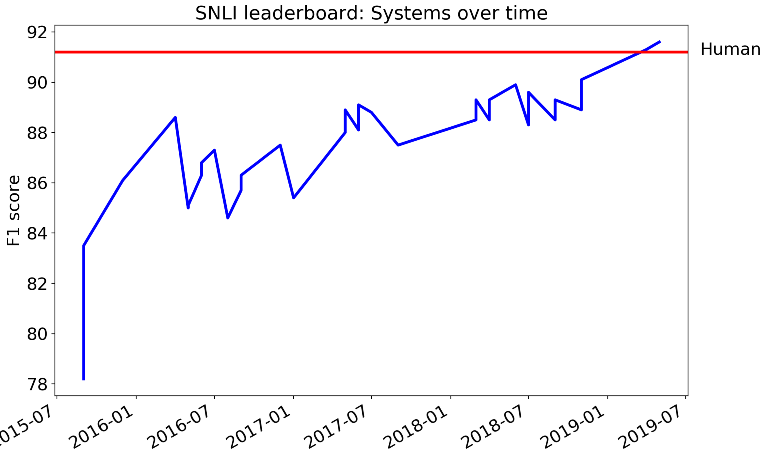
Premise	Relation	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	**contradiction** c c c c c	The man is sleeping
An older and younger man smiling.	**neutral** n n e n n	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	**contradiction** c c c c c	A man is driving down a lonely road.
A soccer game with multiple males playing.	**entailment** e e e e e	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	**neutral** n n e c n	A happy woman in a fairy costume holds an umbrella.

Event coreference

Premise	Relation	Hypothesis
A boat sank in the Pacific Ocean.	contradiction	A boat sank in the Atlantic Ocean.
Ruth Bader Ginsburg was appointed to the Supreme Court.	contradiction	I had a sandwich for lunch today

If premise and hypothesis *probably* describe a different photo, then the label is contradiction

Progress on SNLI



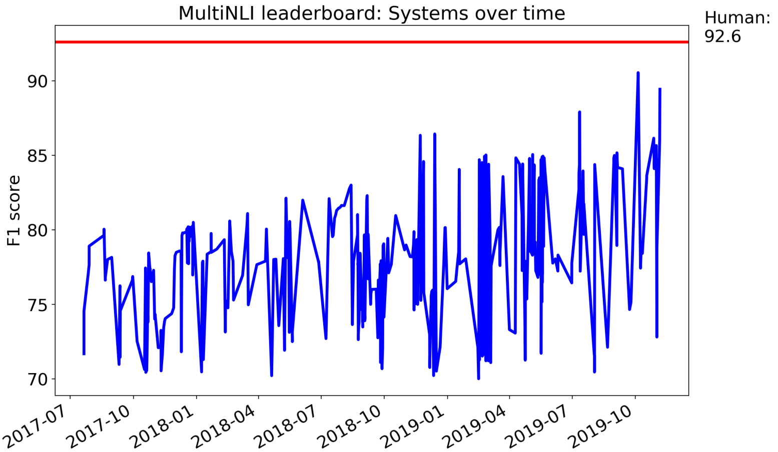
MultiNLI

1. Williams et al. 2018
2. Train premises drawn from five genres:
   - Fiction: works from 1912–2010 spanning many genres
   - Government: reports, letters, speeches, etc., from government websites
   - The *Slate* website
   - Telephone: the Switchboard corpus
   - Travel: Berlitz travel guides
3. Additional genres just for dev and test (the mismatched condition):
   - The 9/11 report
   - Face-to-face: The Charlotte Narrative and Conversation Collection
   - Fundraising letters
   - Non-fiction from Oxford University Press
   - *Verbatim*: articles about linguistics
4. 392,702 train examples; 20K dev; 20K test
5. 19,647 examples validated by four additional annotators
   - 58.2% examples with unanimous gold label
   - 92.6% of gold labels match the author's label
6. Test-set labels available as a Kaggle competition.
7. Project page: `https://www.nyu.edu/projects/bowman/multinli/`

MultiNLI annotations

	Matched	Mismatched
ACTIVE/PASSIVE	15	10
ANTO	17	20
BELIEF	66	58
CONDITIONAL	23	26
COREF	30	29
LONG_SENTENCE	99	109
MODAL	144	126
NEGATION	129	104
PARAPHRASE	25	37
QUANTIFIER	125	140
QUANTITY/TIME_REASONING	15	39
TENSE_DIFFERENCE	51	18
WORD_OVERLAP	28	37
	767	753

Progress on MultiNLI



Adversarial NLI dataset (ANLI)

1. Nie et al. 2019b
2. 162,865 labeled examples
3. The premises come from diverse sources.
4. The hypotheses are written by crowdworkers with the explicit goal of fooling state-of-the-art models.
5. This effort is a direct response to the results and findings for SNLI and MultiNLI that we just reviewed.

ANLI dataset creation

1. The annotator is presented with a premise sentence and a condition (entailment, contradiction, neutral).
2. The annotator writes a hypothesis.
3. A state-of-the-art model makes a prediction about the premise–hypothesis pair.
4. If the model's prediction matches the condition, the annotator returns to step 2 to try again.
5. If the model was fooled, the premise–hypothesis pair is independently validated by other annotators.

Additional ANLI details

Round	Model	Training data	Context sources	Examples
R1	BERT-large (Devlin et al. 2019)	SNLI + MultiNLI	Wikipedia	16,946
R2	ROBERTa (Liu et al. 2019)	SNLI + MultiNLI + NLI-FEVER + R1	Wikipedia	45,460
R3	ROBERTa (Liu et al. 2019)	SNLI + MultiNLI + NLI-FEVER + R2	Various	100,459
				162,865

- The train sets mix cases where the model's predictions were correct and incorrect. The majority of the model predictions are correct, though.
- The dev and test sets contain only cases where the model's prediction was incorrect.

Dynabench

Dynabench: Rethinking Benchmarking in NLP

Douwe Kiela[†], Max Bartolo[‡], Yixin Nie[⋆], Divyansh Kaushik[§], Atticus Geiger[¶], Zhengxuan Wu[¶], Bertie Vidgen[‖], Grusha Prasad[⋆⋆], Amanpreet Singh[†], Pratik Ringshia[†], Zhiyi Ma[†], Tristan Thrush[†], Sebastian Riedel[†‡], Zeerak Waseem[††], Pontus Stenetorp[‡], Robin Jia[†], Mohit Bansal[⋆], Christopher Potts[¶] and Adina Williams[†]

[†] Facebook AI Research; [‡] UCL; [⋆] UNC Chapel Hill; [§] CMU; [¶] Stanford University
[‖] Alan Turing Institute; [⋆⋆] JHU; [††] Simon Fraser University
dynabench@fb.com

https://dynabench.org

Dynabench

Dyna Bench

Rethinking AI Benchmarking

Dynabench is a research platform for dynamic data collection and benchmarking. Static benchmarks have well-known issues: they saturate quickly, are susceptible to overfitting, contain exploitable annotator artifacts and have unclear or imperfect evaluation metrics.

This platform in essence is a scientific experiment: can we make faster progress if we collect data dynamically, with humans and models in the loop, rather than in the old-fashioned static way?

Read more

https://dynabench.org

Other NLI datasets

- The GLUE benchmark (diverse tasks including NLI; Wang et al. 2018): https://gluebenchmark.com
- NLI Style FEVER (Nie et al. 2019a): https://github.com/easonnie/combine-FEVER-NSMN/blob/master/other_resources/nli_fever.md
- OCNLI: Original Chinese Natural Language Inference (Hu et al. 2020): https://github.com/CLUEbenchmark/OCNLI
- Turkish NLI (Budur et al. 2020): https://github.com/boun-tabi/NLI-TR
- XNLI (multilingual dev/test derived from MultiNLI; Conneau et al. 2018): https://github.com/facebookresearch/XNLI
- Diverse Natural Language Inference Collection (DNC; Poliak et al. 2018): http://decomp.io/projects/diverse-natural-language-inference/
- MedNLI (derived from MIMIC III; Romanov and Shivade 2018) https://physionet.org/content/mednli/1.0.0/
- SciTail (derived from science exam questions and Web text; Khot et al. 2018): http://data.allenai.org/scitail/

References I

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Stroudsburg, PA. Association for Computational Linguistics.

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. Data and Representation for Turkish Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *AAAI*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. To appear in NAACL 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. ROBERTa: A robustly optimized BERT pretraining approach. ArXiv:1907.11692.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019a. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019b. Adversarial NLI: A new benchmark for natural language understanding. UNC CHapel Hill and Facebook AI Research.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

References II

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.