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.
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
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

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

<table>
<thead>
<tr>
<th>Premise</th>
<th>Relation</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A boat sank in the Pacific Ocean.</td>
<td>contradiction</td>
<td>A boat sank in the Atlantic Ocean.</td>
</tr>
<tr>
<td>Ruth Bader Ginsburg was appointed to the Supreme Court.</td>
<td>contradiction</td>
<td>I had a sandwich for lunch today</td>
</tr>
</tbody>
</table>

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

SNLI leaderboard: Systems over time

- F1 score
- SNLI
- MultiNLI
- ANLI
- Dynabench
- Other NLI datasets

Human
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

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

|                  | 767     | 753        |
Progress on MultiNLI

MultiNLI leaderboard: Systems over time

Human: 92.6
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

<table>
<thead>
<tr>
<th>Round</th>
<th>Model</th>
<th>Training data</th>
<th>Context sources</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>BERT-large</td>
<td>SNLI + MultiNLI</td>
<td>Wikipedia</td>
<td>16,946</td>
</tr>
<tr>
<td></td>
<td>(Devlin et al. 2019)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>ROBERTa</td>
<td>SNLI + MultiNLI + NLI-FEVER + R1</td>
<td>Wikipedia</td>
<td>45,460</td>
</tr>
<tr>
<td></td>
<td>(Liu et al. 2019)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>ROBERTa</td>
<td>SNLI + MultiNLI + NLI-FEVER + R2</td>
<td>Various</td>
<td>100,459</td>
</tr>
<tr>
<td></td>
<td>(Liu et al. 2019)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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: 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
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?

https://dynabench.org
Other NLI datasets

- The GLUE benchmark (diverse tasks including NLI; Wang et al. 2018): https://gluebenchmark.com
- 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
- 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


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.


