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

<table>
<thead>
<tr>
<th>Premise</th>
<th>Relation</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A turtle danced.</td>
<td>entails</td>
<td>A turtle moved.</td>
</tr>
<tr>
<td>Every reptile danced.</td>
<td>neutral</td>
<td>A turtle ate.</td>
</tr>
<tr>
<td>Some turtles walk.</td>
<td>contradicts</td>
<td>No turtles move.</td>
</tr>
</tbody>
</table>
## Adversarial testing

<table>
<thead>
<tr>
<th>Premise</th>
<th>Relation</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A little girl kneeling in the dirt crying.</td>
<td>entails</td>
<td>A little girl is very sad.</td>
</tr>
<tr>
<td>Adversarial</td>
<td>entails</td>
<td>A little girl is very unhappy.</td>
</tr>
</tbody>
</table>

Train

Glockner et al. 2018
### Adversarial testing

<table>
<thead>
<tr>
<th>Premise</th>
<th>Relation</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A <strong>woman</strong> is pulling a <strong>child</strong> on a sled in the snow.</td>
<td>entails</td>
<td>A child is sitting on a sled in the snow.</td>
</tr>
<tr>
<td>A <strong>child</strong> is pulling a <strong>woman</strong> on a sled in the snow.</td>
<td>neutral</td>
<td></td>
</tr>
</tbody>
</table>

Nie et al. 2019


