Identifying inherent disagreement in natural language inference

Xinliang Frederick Zhang and Marie-Catherine de Marneffe
The Ohio State University

NAACL 2021
Premise: A homeless man being observed by a man in business attire.

Hypothesis: Two men are sleeping in a hotel.

Contradiction  Neutral  Entailment

Inherent Disagreements in Human Textual Inferences (Pavlick and Kwiatkowski, 2019)
Premise: A homeless man being observed by a man in business attire.

Hypothesis: Two men are sleeping in a hotel.

Contradiction Neutral Entailment
Data: CommitmentBank

Premise: B: Yeah, and EDS is very particular about this, hair cuts, A: Wow. B: I mean it was like you can’t have, you know, such and such facial hair, no beards, you know, and just really detailed. A: I don’t know that that would be a good environment to work in.

Hypothesis: that would be a good environment to work in

Label? [ 2, 0, 0, 0, 0, -1, -2, -3 ]
Data: CommitmentBank

Premise: B: Yeah, and EDS is very particular about this, hair cuts, A: Wow. B: I mean it was like you can’t have, you know, such and such facial hair, no beards, you know, and just really detailed. A: I don’t know that that would be a good environment to work in.

Hypothesis: that would be a good environment to work in

Label? [2, 0, 0, 0, 0, -1, -2, -3]

The CommitmentBank: Investigating projection in naturally occurring discourse (de Marneffe et al., 2019)
Finer-grained labels for NLI

Premise: B: Yeah, and EDS is very particular about this, hair cuts, A: Wow. B: I mean it was like you can’t have, you know, such and such facial hair, no beards, you know, and just really detailed. A: I don’t know that that would be a good environment to work in.

Hypothesis: that would be a good environment to work in

Disagreement [ 2, 0, 0, 0, 0, -1, -2, -3 ]

The CommitmentBank: Investigating projection in naturally occurring discourse (de Marneffe et al., 2019)
Model: Artificial Annotators (AAs)

\[
P(y|x) = \text{softmax}(W_s \tanh(W_t[e; n; c]))
\]
### Baselines and AAs perform better across the board

<table>
<thead>
<tr>
<th></th>
<th>Dev</th>
<th></th>
<th>Test</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>F1</td>
<td>Acc</td>
<td>Entail</td>
<td>Neutral</td>
<td>Contradict</td>
<td>Disagree</td>
</tr>
<tr>
<td>Always 0</td>
<td>55.00</td>
<td>39.03</td>
<td>45.42</td>
<td>28.37</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>CBOW</td>
<td>55.25</td>
<td>40.54</td>
<td>45.09</td>
<td>28.37</td>
<td>0.00</td>
<td>0.00</td>
<td>0.69</td>
</tr>
<tr>
<td>Heuristic</td>
<td>65.00</td>
<td>62.08</td>
<td>54.17</td>
<td>50.60</td>
<td>22.54</td>
<td><strong>52.94</strong></td>
<td>64.46</td>
</tr>
<tr>
<td>Vanilla BERT</td>
<td>63.71</td>
<td>63.54</td>
<td>62.50</td>
<td>61.93</td>
<td>59.26</td>
<td>49.64</td>
<td>69.09</td>
</tr>
<tr>
<td>Joint BERT</td>
<td>64.47</td>
<td>64.28</td>
<td>62.61</td>
<td>62.07</td>
<td>59.77</td>
<td>47.27</td>
<td>67.36</td>
</tr>
<tr>
<td>AAs (ours)</td>
<td><strong>65.15</strong></td>
<td><strong>64.41</strong></td>
<td><strong>65.60</strong>*</td>
<td><strong>64.97</strong>*</td>
<td><strong>61.07</strong></td>
<td>51.27</td>
<td><strong>70.89</strong></td>
</tr>
</tbody>
</table>

Baselines and AAs overall performance on CB dev and test sets, and F1 scores of each class on the test set (average of 10 runs). * indicates a statistically significant difference (t-test, p≤0.01).
AAs learn linguistic patterns and context-dependent inference better.

<table>
<thead>
<tr>
<th>Correct inference by Heuristic?</th>
<th>Correctly predicted (130)</th>
<th>Missed (110)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc.</td>
<td>F1</td>
</tr>
<tr>
<td>V. BERT</td>
<td>80.00</td>
<td>80.45</td>
</tr>
<tr>
<td>J. BERT</td>
<td>79.74</td>
<td>80.04</td>
</tr>
<tr>
<td>AAs</td>
<td>84.37</td>
<td>84.85</td>
</tr>
</tbody>
</table>

BERT-based models performance on test items correctly predicted by vs. items missed by linguistic rules.
Error analysis

Premise:    B: Yeah, and EDS is very particular about this, hair cuts, A: Wow. B: I mean it was like you can’t have, you know, such and such facial hair, no beards, you know, and just really detailed. A: I don’t know that that would be a good environment to work in.

Hypothesis: that would be a good environment to work in

Heuristics: C   V. BERT: C
J. BERT: D   AAs: C {C, C, C}
Disagreement [2, 0, 0, 0, 0, -1, -2, -3]
Towards robust NLI

Our Artificial Annotators are a start in this direction but still far from succeeding (~ 66%).

A method which captures accurately the number of modes in the annotation distribution would lead to a better model.
Thanks!

Code is available at:

https://github.com/FrederickXZhang/FgNLI

Contact: zhang.9975@osu.edu