A large annotated corpus for learning natural language inference (Bowman et al., 2015)

Presentation for L101: Machine Learning for Language Processing
Raymond Wang
25 October 2022
<table>
<thead>
<tr>
<th>p (premise)</th>
<th>h (hypothesis)</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhausted looking firemen are walking.</td>
<td>Firemen are walking.</td>
<td>Entailment</td>
</tr>
<tr>
<td>A man walking proudly down the street.</td>
<td>The man is part of the gay pride parade.</td>
<td>Neutral</td>
</tr>
<tr>
<td>Two ladies are reading through binders.</td>
<td>The girls are watching a movie.</td>
<td>Contradiction</td>
</tr>
</tbody>
</table>

SNLI (Bowman et al., 2015)
Existing datasets

- Stanford Natural Language Inference Corpus
- 570k pairs of sentences
- Written by humans, labelled by humans

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Natural</th>
<th>Validated</th>
</tr>
</thead>
<tbody>
<tr>
<td>FraCaS</td>
<td>300</td>
<td>-</td>
<td>Yes</td>
</tr>
<tr>
<td>RTE</td>
<td>7k</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SICK</td>
<td>10k</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SNLI</td>
<td>570k</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DG</td>
<td>728k</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Levy</td>
<td>1,500k</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PPDB</td>
<td>100,000k</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Bowman, Samuel, EMNLP, 2015
Coreference

Bowman et al., 2015
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. Example: For the caption “Two dogs are running through a field,” you could write “There are animals outdoors.”

- Write one alternate caption that might be a **true** description of the photo. Example: For the caption “Two dogs are running through a field,” you could write “Some puppies are running to catch a stick.”

- Write one alternate caption that is **definitely** a **false** description of the photo. Example: For the caption “Two dogs are running through a field,” you could write “The pets are sitting on a couch.” This is different from the maybe correct category because it’s impossible for the dogs to be both running and sitting.

### Data set sizes:
- Training pairs: 550,152
- Development pairs: 10,000
- Test pairs: 10,000

### Sentence length:
- Premise mean token count: 14.1
- Hypothesis mean token count: 8.3

### Parser output:
- Premise ‘S’-rooted parses: 74.0%
- Hypothesis ‘S’-rooted parses: 88.9%
- Distinct words (ignoring case): 37,026

Figure 2: The distribution of sentence length.
Validation

Around 10% of data were validated by 4 more annotators.

Rate of agreement is extremely high -> corpus is sufficiently high quality to pose a challenging but realistic task.

Bowman et al., 2015
Table 5: 3-class accuracy, training on either our data or SICK, including models lacking cross-bigram features (Feature 6), and lacking all lexical features (Features 4–6). We report results both on the test set and the training set to judge overfitting.

Table 7: LSTM 3-class accuracy on the SICK train and test sets under three training regimes.

**Model Results**

Lexicalised classifier

Sentence embeddings

Bowman et al., 2015
Issues

• Captions
• Short sentences
• Lexicalised and unigrams
• Aggregate performance into accuracy
• No world knowledge; word embedding and senses
• h shares focus with p
• Validation
<table>
<thead>
<tr>
<th>p</th>
<th>h</th>
<th>$\text{? (e, n, c)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A young woman with long orange hair is sitting on a city bench.</td>
<td>A young woman is sitting on a bench in the park.</td>
<td>Undecidable (2, 2, 1)</td>
</tr>
<tr>
<td>Two elderly men having a conversation, snow covered grass in the background.</td>
<td>The men are drinking coffee and having some cookies.</td>
<td>Contradiction (0, 2, 3)</td>
</tr>
<tr>
<td>two people working in water next to field</td>
<td>Two people are planting rice.</td>
<td>Neutral (0, 3, 2)</td>
</tr>
<tr>
<td>A man in red stands with his child at the beach.</td>
<td>A man wearing red standing with his child at a beach overlooking the ocean.</td>
<td>Neutral (2, 3, 0)</td>
</tr>
<tr>
<td>Man wearing black t-shirt sitting at a computer desk.</td>
<td>The man is working on the computer.</td>
<td>Entailment (3, 2, 0)</td>
</tr>
</tbody>
</table>
• MNLI (Williams et al., 2018)
• Transformer based models (see BERT Devlin et al., 2018; RoBERTa Liu et al., 2019, etc.)
• Evaluation metrics (see GLUE Wang et al., 2019)
References


