SQuAD 2.0 Project
Liyang Sun, Konrad Morzkowski, Phillip Hoovestol

Problem and Data

- Problem of general question answering
- Current top solutions based on transformer architectures (BERT)
- SQuAD 2.0 dataset
- Difference with old SQuAD: unanswerable questions
- Train/Dev/Test split: 129 941/5951/5915

SQuAD 2.0 Project
The Stanford Question Answering Dataset

Approach

Char-BiDAF
- Adding character based embeddings.
- Learned via 1-D convolutional network.

Bert
- Based on the transformer encoder
- Pre-trained on masked language modeling and next sentence prediction tasks.
- Variable threshold for none answer predictions

Ensemble
- Combining different individual models’ outputs
- Plurality vote method
- Models’ ranking as tie-breaker
- Specific selection rules, by taking into account models’ diversity
- Predict no-answer anytime BERT with threshold +3.0 predicted no-answer.

AoA
- Question to document and document to question attention vectors computed.
- Used to scaled BERT’s output
- Performed worse than the baseline
- Scaling of BERT’s inputs unnecessary (more sophisticated attention)

BERT
- Null score difference threshold: substantial role in performance

Ensemble
- Ranking tie-breaker and no-answer technique improves the model
- Better performance with diverse individual models, even if those models individually perform worse.

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev F1</th>
<th>Dev EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>61.44</td>
<td>57.92</td>
</tr>
<tr>
<td>BiDAF (Character embeddings)</td>
<td>62.24</td>
<td>59.05</td>
</tr>
<tr>
<td>BERT (threshold -1.0)</td>
<td>76.12</td>
<td>72.99</td>
</tr>
<tr>
<td>BERT (threshold -3.0)</td>
<td>75.85</td>
<td>73.03</td>
</tr>
<tr>
<td>BERT (threshold +3.0)</td>
<td>72.33</td>
<td>68.87</td>
</tr>
<tr>
<td>Naive Ensemble</td>
<td>78.47</td>
<td>73.77</td>
</tr>
<tr>
<td>Ensemble with ranking</td>
<td>76.61</td>
<td>73.87</td>
</tr>
<tr>
<td>Ensemble with ranking, models selection, and null rule</td>
<td>77.07</td>
<td>74.19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Test F1</th>
<th>Test EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble with ranking, models selection, and null rule</td>
<td>77.40</td>
<td>74.49</td>
</tr>
<tr>
<td>BERT (threshold -1.0)</td>
<td>76.69</td>
<td>73.61</td>
</tr>
</tbody>
</table>

Analysis

AoA
- Performed worse than the baseline
- Scaling of BERT’s inputs unnecessary (more sophisticated attention)

BERT
- Null score difference threshold: substantial role in performance

Ensemble
- Ranking tie-breaker and no-answer technique improves the model
- Better performance with diverse individual models, even if those models individually perform worse.

<table>
<thead>
<tr>
<th>Model</th>
<th>Type 1 error</th>
<th>Type 2</th>
<th>Type 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT (threshold -3.0)</td>
<td>647</td>
<td>637</td>
<td>416</td>
</tr>
<tr>
<td>BERT (threshold -1.0)</td>
<td>451</td>
<td>782</td>
<td>477</td>
</tr>
<tr>
<td>BERT (threshold +3.0)</td>
<td>240</td>
<td>1025</td>
<td>513</td>
</tr>
</tbody>
</table>

Conclusion

Takeaways
- Solid grasp on current state-of-the-art question answering models.
- General understanding of neural networks implementation and evaluation.
- Rewarding to be able to identify error categories and improve upon them by simple modeling changes.

Future Work
- Training our models longer to enhance performance
- Synthetic self-training to generate relevant examples for training.

References