BERT++: Reading Comprehension on SQuAD

Lirong Yuan

Problem

SQuAD 2.0
- Read Comprehension (RC):
  - Goal: find an answer in a paragraph or a document.
  - Required skills: logical reasoning, commonsense reasoning, understand analogy, causal relations, clause relations, and so on.
- Stanford Question Answering Dataset (SQuAD 2.0):
  - 500+ Wikipedia articles.
  - 100,000+ answerable question-answer pairs.
  - 50,000 unanswerable questions.

BERT
- Language representation model that could be fine-tuned with an additional layer to create models.
- End-to-end model for SQuAD:
  - Embedding layer
  - Linear layer
  - Loss function
  - Predictions
- Limitations:
  - Lack of generalization and real understanding.
  - System should abstain from answering when the question is unanswerable.

Approach

HybridBERT
- Classification Layer: predict answerability of questions
- Loss function:

\[ L_{\text{joint}} = -\log \left( \frac{1}{\delta} e^x + \frac{\delta}{1-\delta} e^{\sum_j \alpha_j} \right) \]

BERT+
- Model architecture:
  - Input data
  - Neural reader
  - Answer verifier
  - Loss functions
  - Predictions
- Neural reader: Predict plausible answer for even unanswerable questions
- Answer verifier: Check whether the candidate answer is legitimate.

Results

Standard Evaluation metrics
- Exact match (EM):
  - Whether the model output matches the ground truth answers exactly.
- F1:
  - Harmonic mean of precision and recall (less strict than the EM score).

Leaderboard

<table>
<thead>
<tr>
<th>System</th>
<th>Dev PCE LeaderBoard EM</th>
<th>Test PCE LeaderBoard EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT (baseline)</td>
<td>71.915</td>
<td>74.487</td>
<td></td>
</tr>
<tr>
<td>HybridBERT (single model)</td>
<td>72.474</td>
<td>75.158</td>
<td></td>
</tr>
<tr>
<td>BERT++ (single large-used model)</td>
<td>76.077</td>
<td>78.523</td>
<td>71.956 74.560</td>
</tr>
<tr>
<td>BERT++ (single large-uncased model)</td>
<td>77.657</td>
<td>80.180</td>
<td></td>
</tr>
<tr>
<td>BERT++ (ensemble model)</td>
<td>78.233</td>
<td>80.530</td>
<td>75.300 77.696</td>
</tr>
</tbody>
</table>

Distribution of no-answer probabilities
- Baseline model
- BERT++:

Impact of passage length, question length, answer length

Analysis

Sample of errors for unanswerable questions
- Missing Information:
  - Question can be answered, but the information is not in the context.
- False Premise:
  - Question asserts a fact that contradicts information in the context.
- Topic Error:
  - Question references a related but different entity in the context.
- Content negation:
  - Question asks for the opposite information mentioned in the context.

Error Type:
- Passage
- Question
- Predicted Answer
- Missing Information
- Similarity, it is not known if L (the set of all problems that can be solved in logarithmic space) is strictly contained in P or equal to P. Again, there are many complexity classes between the two, such as NL and NC, and it is not known if they are distinct or equal classes.
- False premise
  - James Wolfe defeated Montcalm at Quebec in a battle that claimed the lives of both commanders, and victory at Fort Niagara successfully cut off the French frontier forts further to the west and south.
  - Who was defeated by Montcalm at Quebec?
- Topic error
  - Mercury is the working fluid in the mercury vapor turbine. Low boiling hydrocarbons can be used in a binary cycle.
  - What is the typical working fluid in a vapor turbine?
- Content negation
  - In 1918 Roger de Tourney travelled to the Iberian Peninsula to carve out a state for himself from Moorish lands, but failed...
  - Who carved out a state for himself from Moorish lands?

Conclusions

- BERT++ is better at modeling answerability of questions.
- Achieved 75.300 EM, 77.696 F1 on test set and competitive rank.

Future work

- Enhance language representation: integrate world knowledge to learn more complete semantic representation of larger language units.
- Better sentence representation: encode sentence structures with neural dependency parser.
- Build targeted models on data sets that have: short passages, very short and very long questions, or long answers and ensemble.

Acknowledgments
I would like to thank Chris Manning and all the TAs for the wonderful classes and patient help during office hours. I would also like to thank Microsoft Azure for providing resources to train the models.

References