Quick and (not so) Dirty: Unsupervised Selection of Justification Sentences for Multi-hop Question Answering

Vikas Yadav  Steven Bethard  Mihai Surdeanu

vikasy, bethard, msurdeanu@email.arizona.edu || University of Arizona, Tucson

Task and Contributions

Task:
- Explainable Multi-hop Question Answering (QA) - experiments on MultiRC and ARC.
- Justification selection for improving interpretability of a question answering system.

Contributions:
- An unsupervised justification selection approach which increases relevance (R) and coverage (C) of query terms in justifications with reduced redundancy (O)
- Unsupervised state-of-the-art results on justification selection task with similar or better performance over supervised BERT in various domains.
- State-of-the-art QA performance on MultiRC and ARC by coupling AutoROCC with BERT.

Approach

Score
Goal is to maximize relevance (R) and coverage (C) while reducing the redundancy amongst justification sentences (O).

\[ C(X) = \frac{\sum_{i=1}^{N} |D(t_i) \cap \{t(t_i) \cap t(s_i)\}|}{|X|} \quad (1) \]

Overlap (O) - Ratio of common terms between individual justifications

\[ S(P_i) = \frac{R}{\epsilon + O(P_i)} \left( e^{+C(A)} + e^{+C(Q)} \right) \quad (2) \]

Examples

Question: To which organ system do the esophagus, liver, pancreas, small intestine, and colon belong? (A) reproductive system (B) excretory system (C) digestive system (D) endocrine system

ROCC-selected justification sentences:
- vertebrate digestive system has oral cavity, teeth and pharynx, esophagus and stomach, small intestine, pancreas, liver and the large intestine.
- digestive system consists liver, stomach, large intestine, small intestine, colon, rectum, anus

Score

Goal is to maximize relevance (R) and coverage (C) while reducing the redundancy amongst justification sentences (O).

\[ C(X) = \frac{\sum_{i=1}^{N} |D(t_i) \cap \{t(t_i) \cap t(s_i)\}|}{|X|} \quad (1) \]

Overlap (O) - Ratio of common terms between individual justifications

\[ S(P_i) = \frac{R}{\epsilon + O(P_i)} \left( e^{+C(A)} + e^{+C(Q)} \right) \quad (2) \]

Analysis

Table: Performance on the MultiRC dataset, under various configurations.

<table>
<thead>
<tr>
<th>Method</th>
<th>F1m</th>
<th>F1a</th>
<th>EM0</th>
<th>Justice</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT + All passages</td>
<td>70.3</td>
<td>67.3</td>
<td>22.8</td>
<td></td>
</tr>
<tr>
<td>BERT + Science textbook</td>
<td>73.0</td>
<td>69.6</td>
<td>22.8</td>
<td></td>
</tr>
<tr>
<td>BERT + Fiction</td>
<td>70.5</td>
<td>67.8</td>
<td>13.3</td>
<td></td>
</tr>
<tr>
<td>BERT + News</td>
<td>69.7</td>
<td>67.9</td>
<td>16.9</td>
<td></td>
</tr>
<tr>
<td>GPT-2 Wang et al. 2019</td>
<td>73.1</td>
<td>70.5</td>
<td>21.8</td>
<td></td>
</tr>
</tbody>
</table>

Table: Top results of the non-parametric ROCC vs. a supervised sentence selection model, evaluated on the gold justification sentences from MultiRC.

Codes - https://github.com/vikas95/AutoROCC