QUOREF: A Reading Comprehension Benchmark with Questions Requiring Coreferential Reasoning

Pradeep Dasigi (A12), Nelson F. Liu ( ), Ana Marasović (A12), Noah A. Smith (W, A12), Matt Gardner (A12)

24K span-selection questions on English Wikipedia passages
Questions require coreference resolution (est. 78%)
Answers with multiple spans (10%)
Built with strong adversary in loop

Pronoun resolution* (69%)
Anna and Declan [...] discover the three van thieves going through Anna’s luggage. Declan fights them. [...] Who does Declan get into fight with? the three van thieves

Nominal resolution* (54%)
Aside from Constantinople and Trebizond, other Byzantine cities also featured tzykanisteria, most notably Sparta, Ephesus, and Athens. [...] Which cities had tzykanisteria? Constantinople; Trebizond; Sparta; Ephesus; Athens

Multiple resolutions* (32%)
The now upbeat collective keep the toucan, nicknaming it “Amigo” [...] When authorities show up to catch the bird, Pete and Liz spirit him away by hiding him in her dress. What is the name of the character who hides in Liz’s dress? Amigo

Common-sense* (10%)
Amos [...] with his mother Fania and father Arieh. One of his mother’s friends is killed while hanging up laundry during the war [...] Fania fall into depression [...] kills herself by overdose [...]. How does Arieh’s wife die? kills herself by overdose

* estimated from 100 training instances

Baseline Performance

→ Passage-only models show extent of artifacts
→ Best model << human performance (23 F1 points)
→ Human performance est. from 400 test instances

Error Analysis: Distance

Error Analysis: Main Entity

→ Main entity is the first or the most frequent entity in passage ⇒ artifact of Wikipedia passages
→ XLNet-QA exploits answers being main entities

Crowdsourcing Details

1. Write questions that link multiple mentions of the same entity
2. Follow good and bad examples

Dataset Stats

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questions</td>
<td>19399</td>
<td>2418</td>
<td>2537</td>
</tr>
<tr>
<td>Paragraphs</td>
<td>3771</td>
<td>454</td>
<td>477</td>
</tr>
<tr>
<td>Paragraph vocab</td>
<td>57648</td>
<td>18226</td>
<td>18885</td>
</tr>
<tr>
<td>Question vocab</td>
<td>19803</td>
<td>5579</td>
<td>5624</td>
</tr>
<tr>
<td>% multi-span answers</td>
<td>10.2</td>
<td>9.1</td>
<td>9.7</td>
</tr>
</tbody>
</table>

Dataset and leaderboard @ allennlp.org/quoref