TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension

Mandar Joshi, Eunsol Choi, Daniel S. Weld, Luke Zettlemoyer

Presenter: Zhuolun Xiang
Background

• Question Answering (QA) Formulation
  • Answer a question $q$ given evidences $D$
  • Dataset of tuples $\{(q_i, a_i, D_i) | i = 1, ..., n\}$
  • $a_i$ is a substring of $D_i$

• Example

Answer: Fitness
Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney’s first video release “Callanetics: 10 Years Younger In 10 Hours” outsold every other fitness video in the US.
Overview

• TriviaQA
  • Over 650K question-answer-evidence triples
  • First dataset where questions authored by trivia enthusiasts
  • Evidence documents from Web search and Wiki pages
  • A high percentage of the questions are challenging

• Dataset samples

<table>
<thead>
<tr>
<th>QuestionId</th>
<th>Question</th>
<th>Answer</th>
<th>Web</th>
<th>Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>qw_3199</td>
<td>Miami Beach in Florida borders which ocean?</td>
<td>Atlantic</td>
<td>Evidence</td>
<td>Evidence</td>
</tr>
<tr>
<td>bt_1255</td>
<td>What was the occupation of Lovely Rita according to the song by the Beatles</td>
<td>Traffic Warden</td>
<td>Evidence</td>
<td>Evidence</td>
</tr>
<tr>
<td>og_77</td>
<td>Who was Poopdeck Pappys most famous son?</td>
<td>Popeye</td>
<td>Evidence</td>
<td>Evidence</td>
</tr>
<tr>
<td>wh_1026</td>
<td>The Nazi regime was Germany's Third Reich; which was the first Reich?</td>
<td>HOLY ROMAN EMPIRE</td>
<td>Evidence</td>
<td>Evidence</td>
</tr>
</tbody>
</table>
Overview

• TriviaQA
  • Over 650K question-answer-evidence triples
• First dataset where questions authored by trivia enthusiasts
• Evidence documents from Web search and Wiki pages
• A high percentage of the questions are challenging

Question:

Miami Beach in Florida borders which ocean?

Answer: Atlantic

The left column contains Web search results. The right column contains question entity pages from Wikipedia.

Web search results

Dania Beach Station | U.S. Customs and Border Protection
https://www.cbp.gov/-/miami-sector-florida/dania-beach-station
Miami Sector Florida. Dania Beach Station; ... The Dania Beach Border Patrol Station area of responsibility is the ... bordered by the Atlantic Ocean on the east and ...

Hotel Ocean, Miami Beach, FL - Booking.com
www.booking.com/hotel/us/ocean-miami-beach.html
Facing the Atlantic Ocean on South Beach's famous Ocean Drive, this Miami hotel is less than a 2-minute walk from the Art Deco ... Fort Lauderdale Hollywood Airport ...

What states border the Atlantic Ocean? | Reference.com
https://www.reference.com/geography/states-border-atlantic-ocean-1...
... Delaware, Florida, Georgia, Louisiana, Maine ... The states in the U.S. that border the Atlantic Ocean include ... Daytona Beach, West Palm Beach, Miami and ...

Miami Beach Vacation Rentals & Condo Rentals - Airbnb
https://www.airbnb.com/s/Miami-Beach-FL
It was my birthday and I chose to spend it in South Beach Miami. I was looking for a beach ... the beach, Ocean ... Lincoln Road, Miami Beach Golf ...

Hotel Ocean – Hotel Ocean
hotelocean.com
Escape to the Luxury of Our Miami Beach Oceanfront Hotel. The Hotel Ocean is situated ideally in ... The Hotel Ocean is in just the right spot to be convenient ...

Question entity pages

Miami Beach, Florida - Wikipedia
https://en.wikipedia.org/wiki/Miami_Beach,_Florida
Florida - Wikipedia
https://en.wikipedia.org/wiki/Florida
Dataset Collection

• Gather question-answer pairs from 14 trivia websites
  • Remove short questions
• Collect evidence from Web search and Wikipedia
  • Web search
    • Pose questions on Bing
    • Exclude trivia websites
    • Crawl top 10 results
  • Wikipedia
    • Use TAGME to find Wikipedia entities in the question
    • Add these pages as evidence
Dataset Collection

• Gather question-answer pairs
  • Remove short questions
• Collect evidence from Web
  • Web search
    • Pose questions on Bing
    • Exclude trivia websites
    • Crawl top 10 results
  • Wikipedia
    • Use TAGME to find Wikipedia entities in the question
    • Add these pages as evidence
Dataset Collection

• Gather question-answer pairs from 14 trivia websites
  • Remove short questions
• Collect evidence from Web search and Wikipedia
  • Web search
    • Pose questions on Bing
    • Exclude trivia websites
    • Crawl top 10 results
  • Wikipedia
    • Use TAGME to find Wikipedia entities in the question
    • Add these pages as evidence
Dataset Collection

- Gather question-answer pairs from 14 trivia websites
  - Remove short questions
- Collect evidence from Web search and Wikipedia
  - Web search
    - Pose questions on Bing
    - Exclude trivia websites
    - Crawl top 10 results
  - Wikipedia
    - Use TAGME to find Wikipedia entities in the question
    - Add these pages as evidence

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of QA pairs</td>
<td>95,956</td>
</tr>
<tr>
<td>Number of unique answers</td>
<td>40,478</td>
</tr>
<tr>
<td>Number of evidence documents</td>
<td>662,659</td>
</tr>
<tr>
<td>Avg. question length (word)</td>
<td>14</td>
</tr>
<tr>
<td>Avg. document length (word)</td>
<td>2,895</td>
</tr>
</tbody>
</table>
Dataset Analysis

- Question-answer pairs
  - Avg length = 14
  - Manually analyze 200 sampled questions

<table>
<thead>
<tr>
<th>Property</th>
<th>Example annotation</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. entities / question</td>
<td>Which politician won the <strong>Nobel Peace Prize</strong> in 2009?</td>
<td>1.77 per question</td>
</tr>
<tr>
<td>Fine grained answer type</td>
<td>What <strong>fragrant essential oil</strong> is obtained from Damask Rose?</td>
<td>73.5% of questions</td>
</tr>
<tr>
<td>Coarse grained answer type</td>
<td><strong>Who</strong> won the Nobel Peace Prize in 2009?</td>
<td>15.5% of questions</td>
</tr>
<tr>
<td>Time frame</td>
<td>What was photographed for the first time in <strong>October 1959</strong></td>
<td>34% of questions</td>
</tr>
<tr>
<td>Comparisons</td>
<td>What is the appropriate name of the <strong>largest</strong> type of frog?</td>
<td>9% of questions</td>
</tr>
</tbody>
</table>
Dataset Analysis

• Question-answer pairs
  • Avg length = 14
  • Manually analyze 200 sampled questions

<table>
<thead>
<tr>
<th>Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical</td>
<td>4.17</td>
</tr>
<tr>
<td>Free text</td>
<td>2.98</td>
</tr>
<tr>
<td>Wikipedia title</td>
<td>92.85</td>
</tr>
<tr>
<td>Person</td>
<td>32</td>
</tr>
<tr>
<td>Location</td>
<td>23</td>
</tr>
<tr>
<td>Organization</td>
<td>5</td>
</tr>
<tr>
<td>Misc.</td>
<td>40</td>
</tr>
</tbody>
</table>

Property of answers
Dataset Analysis

• Question-answer pairs
  • Avg length = 14
  • Manually analyze 200 sampled questions

• Evidences
  • 75.4%/79.7% of Web/Wiki evidences contain answers
  • Human test achieves 75.3/79.6 accuracy on Web/Wiki domains

• Answer 40% of questions needs information from multiple sentences
Experiments: Baseline Methods

- Random entity baseline (Wiki domain only)
  - Entities in Wiki pages form candidate answer set
  - Randomly pick one that not occur in question
- Entity classifier
  - Ranking problem over candidate answers
  - Function learnt using LambdaMART (Wu et al., 10)
- Neural model
  - Use BiDAF model (Seo et al., 17)
Experiments

• Metrics
  • Exact match (EM) and F1 score
  • For numerical and freeform answer: single given answer as ground truth
  • For Wiki entity: use Wiki aliases as well

• Setup
  • Random partition into train (80%)/development (10%)/test (10%)

<table>
<thead>
<tr>
<th></th>
<th>Questions</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Documents</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wikipedia</td>
<td></td>
<td>61,888</td>
<td>7,993</td>
<td>7,701</td>
</tr>
<tr>
<td></td>
<td></td>
<td>110,648</td>
<td>14,229</td>
<td>13,661</td>
</tr>
<tr>
<td>Web</td>
<td></td>
<td>76,496</td>
<td>9,951</td>
<td>9,509</td>
</tr>
<tr>
<td></td>
<td></td>
<td>528,979</td>
<td>68,621</td>
<td>65,059</td>
</tr>
<tr>
<td>Wikipedia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>verified</td>
<td>Questions</td>
<td></td>
<td>297</td>
<td>584</td>
</tr>
<tr>
<td></td>
<td>Documents</td>
<td></td>
<td>305</td>
<td>592</td>
</tr>
<tr>
<td>Web</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>verified</td>
<td>Questions</td>
<td></td>
<td>322</td>
<td>733</td>
</tr>
<tr>
<td></td>
<td>Documents</td>
<td></td>
<td>325</td>
<td>769</td>
</tr>
</tbody>
</table>
Experiments

- Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Domain</th>
<th>Distant Supervision</th>
<th>Verified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EM</td>
<td>F1</td>
</tr>
<tr>
<td>Random</td>
<td>Wiki</td>
<td>12.67</td>
<td>22.45</td>
</tr>
<tr>
<td>Classifier</td>
<td>Wiki</td>
<td>23.42</td>
<td>27.67</td>
</tr>
<tr>
<td>BiDAF</td>
<td>Wiki</td>
<td>40.26</td>
<td>45.74</td>
</tr>
<tr>
<td>Classifier</td>
<td>web</td>
<td>24.63</td>
<td>29.08</td>
</tr>
<tr>
<td>BiDAF</td>
<td>web</td>
<td>41.08</td>
<td>47.40</td>
</tr>
</tbody>
</table>

- Human baseline: 79.7% on Wiki, 75.4% on web
Conclusion

• TriviaQA
  • 650K question-answer-evidence triples
  • Questions authored by trivia enthusiasts
  • Evidence documents from Web search and Wiki pages
  • Experiments show TriviaQA is a challenging testbed

Thanks!