GooAQ: Open Question Answering with Diverse Answer Types

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal
Hannaneh Hajishirzi, Chris Callison-Burch

EMNLP-Fingings’21
State of “Open” Question Answering

• “Open” question answering [Roberts et al., 2020; Lewis et al., 2021]
  • Also referred to as open-response, open-domain, or direct answer QA

- **how many calories burned 30 minutes crossfit?**
- **LM**
- **260 calories**

• Existing work
  • Built datasets and models for answering questions without specified context
    - NQ-open [Lee et al., 2019]
    - TriviaQA [Joshi et al., 2017]
    - ARC-DA [Bhakthavatsalam et al., 2021]
    - ELI5 [Fan et al., 2019] short answers long answers
what are the ingredients used in making black soap?

{ 9 oz Coconut Oil, 20 oz Palm Oil, 3.5 oz Shea Butter, ... }

what is the gravitational force of uranus?

8.87 m/s²

what is the difference between an assignment and a delegation?

The difference is that an assignment can't increase another party's obligations. Delegation, on the other hand, ....

what is the square feet of an acre?

43560

what is the time difference between south africa and mauritius?

Time-conversion
GooAQ: Overview

- Introducing **GooAQ**
  - Open QA benchmark
  - Contains 3 million questions
  - Questions with a variety of types
  - All collected using a coherent process

### Answer types and Count

<table>
<thead>
<tr>
<th>Answer types</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short answers</td>
<td>275k</td>
</tr>
<tr>
<td>unit conversion</td>
<td>45k</td>
</tr>
<tr>
<td>time conversion</td>
<td>2.5k</td>
</tr>
<tr>
<td>knowledge</td>
<td>32k</td>
</tr>
<tr>
<td>snippet (short)</td>
<td>196k</td>
</tr>
<tr>
<td>Snippet</td>
<td>2.7M</td>
</tr>
<tr>
<td>Collection</td>
<td>329k</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td><strong>3.1M</strong></td>
</tr>
</tbody>
</table>
GooAQ: Overview

- Introducing GooAQ

(a) Question Extraction

similar to Berant et al. 2013

(b) Answer Extraction
GooAQ: Overview

• Introducing GooAQ
  • Open QA benchmark
  • Contains 3 million questions
  • Questions with a variety of types
  • All collected using a coherent process

<table>
<thead>
<tr>
<th>Answer types</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short answers</td>
<td>275k</td>
</tr>
<tr>
<td>▼ unit conversion</td>
<td>45k</td>
</tr>
<tr>
<td>▼ time conversion</td>
<td>2.5k</td>
</tr>
<tr>
<td>▼ knowledge</td>
<td>32k</td>
</tr>
<tr>
<td>▼ snippet (short)</td>
<td>196k</td>
</tr>
<tr>
<td>Snippet</td>
<td>2.7M</td>
</tr>
<tr>
<td>Collection</td>
<td>329k</td>
</tr>
<tr>
<td>Overall</td>
<td>3.1M</td>
</tr>
</tbody>
</table>

Differences across different answer types?
Questions: Answer Type Behavior

Whether different answer types:

(Q1) benefit similarly from *pre-training*?
(Q2) benefit similarly from *labeled data*?
(Q3) benefit similarly from *larger models*?

Differences across different answer types?
Empirical Findings: Summary

Whether different answer types:

(Q1) benefit similarly from **pre-training**?
- short: 😞
- snippet: ✅
- collection: ✅

(Q2) benefit similarly from **labeled data**?
- short: ✅
- snippet: 😞
- collection: ✅

(Q3) benefit similarly from **larger models**?
- short: ✅
- snippet: ✅
- collection: ✅

**Pre-training** is surprisingly effective on the snippet and collection answers.

**Labeled data** is more helpful for short answer questions.

**Model size** is more helpful for snippet & collection answers.
Summary

- GooAQ: a collection of 3 million question-answer pairs with a diverse set of answers.

- Benchmark state-of-the-art models on different answer types.

- Valuable resource for training models.

- Code & Data: https://github.com/allenai/gooaq
Google answers

• **Idea:** extract answer boxes provided by Google.

• **Challenge:** what to query?
  • Use the queries from the previous section.

• **Challenge:** How to extract the queries?
  • Option 1: use [https://serpapi.com](https://serpapi.com)
  • Option 2: crawl Google!
Mining Google queries

• Mined ~5M queries and saved their HTML pages
• For each HTML page, we check if it contains an answer box.
• If so, we extract them:

  Question: ...
  Answer type: ...
  Short answers: ...
  Long answer: ...

Google questions: examples

Question: who is the president of usa?

Answer type: short ans

Short answers: Donald Trump

Long answer: -
Google questions: examples

Question: do canadian citizens need a visa for ecuador?

Answer type: snippet

Short answers: -

Long answer: Ecuador tourist visa is not required for citizens of Canada for a stay up to 90 days. ...
Question: how many calories burned 30 minutes crossfit?

Answer type: snippet w/ short answer

Short answers: 260 calories

Long answer: According to an online calorie estimator from the American Council on Exercise, a 115-pound person running for 30 minutes at a slow-to-moderate pace (a 10-minute mile) would burn about 260 calories: the same amount people who did CrossFit typically torched in 20 minutes, according to the research.
Question: what are the steps for decision making?

Answer type: list

Short answers: -

Long answer:

- Step 1: Identify the decision. You realize that you need to make a decision.
- Step 2: Gather relevant information.
- Step 3: Identify the alternatives.
- Step 4: Weigh the evidence.
- Step 5: Choose among alternatives.
- Step 6: Take action.
- Step 7: Review your decision & its consequences.
Question: what are the ingredients used in making black soap?

Answer type: set

Short answers: -

Long answer: {
9 oz Coconut Oil,
20 oz Palm Oil,
3.5 oz Shea Butter,
0.6 oz Coconut Carbon,
0.5 - 1.5 oz Fragrance or Essential Oil,
14 oz Water.
}
Question: what is the gravitational force of uranus?
Answer type: knowledge lookup
Short answers: 8.87 m/s²
Long answer: --
Question: how do you say 20 in hebrew?

Answer type: translate

Short answers: י''ש

Long answer: -
Question: 1 dollar is equivalent to how many rupees?

Answer type: currency conversion

Short answers: 75.52 Indian Rupee

Long answer: -
What is the difference between the United States and Canada?

There is no time difference between Washington, DC and Ottawa, ON, Canada.

Both United States and Canada have multiple time zones.

Question: what is the difference between the United States and Canada?

Answer type: time conversion

Short answers: -

Long answer: There is no time difference between Washington, DC and Ottawa, ON, Canada.
Google answers: question type statistics

# of questions with any answer: 3,146,425
# of questions with short answer: 308,697
# of questions with long answer: 3,033,898
**Experiment #5**

- Take a 200k subset of Google Q/As with short answers (ans length < 5 words)

<table>
<thead>
<tr>
<th>ID</th>
<th>Model</th>
<th>Trained on</th>
<th>Dev Score</th>
<th>Test Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>Giant (T5 – 11B)</td>
<td>NQ-open</td>
<td>34.5</td>
<td>35.1</td>
</tr>
<tr>
<td>D0</td>
<td>Dwarf (T5 – small)</td>
<td>NQ-open</td>
<td>25.3</td>
<td>24.5</td>
</tr>
<tr>
<td>D1</td>
<td>Dwarf (T5 – small)</td>
<td>Google Queries w/ G’s predictions + NQ-open</td>
<td>26.3</td>
<td>26.3</td>
</tr>
<tr>
<td>D2</td>
<td>Dwarf (T5 – small)</td>
<td>Google Queries w/ G’s predictions</td>
<td>10.6</td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td>Dwarf (T5 – small)</td>
<td>Google Queries w/ G’s predictions → NQ-open</td>
<td>26.3</td>
<td>26.0</td>
</tr>
<tr>
<td>D4</td>
<td>Dwarf (T5 – small)</td>
<td><strong>Filtered</strong> Google Queries w/ G’s predictions + NQ-open</td>
<td>26.4</td>
<td>26.4</td>
</tr>
</tbody>
</table>
Questions for you

• What other experiments we could have done here?

• How else would you sell/use this resource?
  • There are a lot of data that can be used for procedural reasoning.
  • Could be used for pre-training a model for QA.