TellMeWhy: A Dataset for Answering Why-Questions in Narratives

Yash Kumar Lal
Nate Chambers
Raymond Mooney
Niranjan Balasubramanian
Knowing *why* is important for reasoning about events.

**Story:** Rudy was convinced that bottled waters all tasted the same. He went to the store and bought several popular brands. He went back home and set them all on a table. He spent several hours tasting them one by one. He came to the conclusion that they actually did taste different.

Q: Why did Rudy go to the store and buy several popular brands?

A: He wanted to taste test.

Knowing why can help

- Explain his motivation
- Visualize events in a narrative (Kintsch & Dijk, 1978)
- Understand plans and goals (Schank & Abelson, 1975)

No large scale datasets for *why* questions *over narratives*
TellMeWhy: A collection of why questions over narratives

ROCStories

**Story:** Rudy was convinced that bottled waters all tasted the same. *He went to the store and bought several popular brands.* He went back home and set them all on a table. He spent several hours tasting them one by one. He came to the conclusion that they actually did taste different.

**Q:** Why did Rudy go to the store and buy several popular brands?

**A:** He wanted to taste test.
TellMeWhy: Dataset Characteristics

- **30k questions**
- **Answers not always in text.**
- **Diverse answers possible.**

Breakdown:

<table>
<thead>
<tr>
<th>Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implicit Answers</td>
<td>29%</td>
</tr>
<tr>
<td>Explicit Answers</td>
<td>71%</td>
</tr>
<tr>
<td>Token Overlap b/w Turker Answers</td>
<td>26%</td>
</tr>
</tbody>
</table>
Benchmarking Large LM-based QA Models on TellMeWhy.

Models

T5 (Raffel et al, 2020)
UnifiedQA (Khashabi et al, 2020)

Fine-tuning large QA models isn’t enough.
Automatic evaluation is inadequate.

- Automatic metrics not enough (unsurprisingly).
- BertScore is the most reliable (relatively).
Takeaways

A challenging why-question dataset.

Large LMs are not capable of answering why questions.

Harness for standardized human evaluation for TellMeWhy.
TellMeWhy: A Dataset for Answering Why-Questions in Narratives

View on GitHub  Download in JSON format  Download in CSV format

Contact: ylal@cs.stonybrook.edu