“Going on a vacation” takes longer than “Going for a walk”: A Study of Temporal Commonsense Understanding

Ben Zhou     Daniel Khashabi*     Qiang Ning*     Dan Roth

*Currently affiliated with AI2
Temporal Common Sense

- Humans assume information when reading
  - Not explicitly mentioned
  - Related to time

- Happens all the time
  - To better understand the storyline and beyond
My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.
My friend Bill **went to Duke University in North Carolina.** With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.
My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.
My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.
My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.

<table>
<thead>
<tr>
<th>Event</th>
<th>Duration</th>
<th>Frequency</th>
<th>Ordering</th>
<th>Stationarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>College</td>
<td>about 4 years, start at the age of 18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bill in North Carolina</td>
<td>about 4 years</td>
<td></td>
<td></td>
<td>Stationarity</td>
</tr>
<tr>
<td>Duke in North Carolina</td>
<td>always (expected)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Join Google</td>
<td>after college graduation</td>
<td></td>
<td>Ordering</td>
<td></td>
</tr>
<tr>
<td>NBA Finals</td>
<td>every year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visit Alma Mater</td>
<td>0-2 times per year, 0-2 days each time</td>
<td></td>
<td>Frequency</td>
<td></td>
</tr>
<tr>
<td>Attend basketball games</td>
<td>a few hours</td>
<td></td>
<td>Duration</td>
<td></td>
</tr>
</tbody>
</table>
My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.

* Human infer temporal common sense that helps them to better understand the story.

- Q: How old is Bill?
  - A: Around 25.
  - R: $3 + 4 + 18$

- Q: How long will take Bill to fly to Duke?
  - A: A few (1-5) hours.
  - R: Duke is always in NC, Bill is now in CA

- Q: How often would he visit Duke in the future?
  - A: A few (<5) times a year.

- Q: Which one happened first, went or joined?
  - A: Went.
Our Contribution

- **MC-TACO 🌮 (multiple choice temporal common-sense):**
  - A dataset that focuses on temporal commonsense
  - **Input:**
  - **Task:** Decide whether each answer is plausible.
  - **Metrics:**
    - Exact Match: able to label all candidate answers of a question
    - F1: The F1 score of "plausible"
  - **Statistics:**
    - 1,893 questions
    - 13,225 question-answer pairs
  - **Conclusion:** current systems are not enough to solve this.

---

**Example:**

He went to Duke University. How long did it take him to graduate?

- **Gold:** 4 years
- **Prediction:**
  - 4 years: ✗
  - 10 days: ✔
  - 3.5 years: ✗
  - 16 hours: ✔

**Metrics:***

- Exact Match: 0.0
- F1: 66.7

**Reading Comprehension:** able to answer any questions regarding a piece of text

**Exact Match:** able to label all candidate answers of a question
MC-TACO: Construction

- **Step 0: Source Sentence Generation**
  - Randomly samples sentences

- **Step 1: Question Generation**
  - Ask people to write questions
    - A) temporal
    - B) non-extractive
      - To require commonsense
  - Ask for one “plausible” answer

He joined Google as a software engineer after graduating from college.

- How long did he stay in college?
  - Duration: 4 years

- Will he work at Google for the rest of his life?
  - Duration: No
**MC-TACO: Construction**

- **Step 2: Question Verification**
  - 2 additional verifications on each question
  - 100% agreement
  - We also ask for
    - 1 “plausible” answer
    - 1 “implausible” answer

---

He joined **Google** as a software engineer after graduating from college.

How long did he stay in college? **Yes**

What did he do after college? **Yes**

Temporal? **Yes**

Non-extractive? **Yes**

---
MC-TACO: Construction

- Step 3: Candidate Answer Expansion
  - Seed answers from step 1+2
  - Expand candidates automatically
    - Perturbations
    - Information Retrieval

He joined Google as a software engineer after graduating from college.

- How long did he stay in college?
  - 4 years
  - 6 years
  - 11 days

- What happened after he started working?
  - He started making money.
  - He started a factory.
  - He contributed to public services.

…
…
Step 4: Answer Labeling
- Each answer is labeled by 4 different annotators
- Either “likely” or “unlikely”
- Enforce 100% agreement
  - Eliminate marginal answers with “intermediate” probability

He joined Google as a software engineer after graduating from college.

How long did he stay in college?
- 4 years
- 6 years
- 11 days

What happened after he started working?
- He started making money.
- He started a factory.
- He contributed to public services.
Results

ESIM: Enhanced LSTM for Natural Language Inference (Chen et al., 2016)
GloVe: Global Vectors for Word Representation (Pennington et al., 2014)
ELMo: Deep contextualized word representations (Peters et al., 2018)
BERT: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2019)
RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)

Naïve Best    ESIM + GloVe    ESIM + ELMo    BERT    BERT + Unit Normalization    RoBERTa (post publication)

F1: 49.8  50.3  54.9  66.1  69.9  72.3
Exact Match: 17.4  20.9  26.4  39.6  42.7  43.6
Human F1: 3 weeks -> 0.75 months
Human Exact Match: 49.8  50.3  54.9  66.1  69.9  72.3

26% improvement over +GloVe
13% drop
40% drop
26% improvement over +GloVe
13% drop
3 weeks -> 0.75 months
Surface Association
Summary

- Define 5 temporal commonsense phenomena
- Present MC-TACO, a QA dataset focused on temporal commonsense
- Show that existing systems are not enough to solve it
- Encourage further research
- Thanks!

GitHub (data, baseline, evaluator)
https://github.com/CogComp/MCTACO/

Leaderboard
https://leaderboard.allenai.org/mctaco/