**Improving Event Duration Question Answering by Leveraging Existing Temporal Information Extraction Data**

Felix Giovanni Virgo, Fei Cheng, Sadao Kurohashi
Graduate School of Informatics, Kyoto University

---

**Introduction**

**Duration Question Answering:** McTACO (Zhou et al., 2019)

If you have ever heard, "Eat a good breakfast", thats why. How long does it take to eat breakfast?

- 15 minutes **plausible**
- several days **not plausible**
- 20 minutes **plausible**

The performance of modern pre-trained NLP models for this task is still far behind humans due to limited training data.

There are plenty of auxiliary resources containing duration information, e.g. UDS-T dataset (Vashishta et al., 2019), that can be used to improve McTACO.

However, a straightforward two-stage fine-tuning is less likely to succeed since there are discrepancy between the two tasks:

- UDS-T : Duration Unit Classification
- McTACO : Duration Question Answering

We need to bridge the discrepancy between the two tasks.

---

**Duration Task Recasting**

We bridge the discrepancy by recasting Duration Information Extraction dataset into Question Answering.

**UDS-T Duration Classification**

Their worker even cleaned 3 of my windows and changed a lightbulb for me.

**UDS-T Duration QA**

Their worker even cleaned 3 of my windows and changed a lightbulb for me.

How long does it take for their worker to clean 3 of my windows?

- 2 hours **plausible**
- a few hours **plausible**
- several years **not plausible**
- 4 months **not plausible**

---

**Statistics**

The duration distribution of our recast data is relatively similar to McTACO, except for minutes and years.

**Number of QA pairs**

<table>
<thead>
<tr>
<th>Model</th>
<th>train + dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>McTACO-duration</td>
<td>1.1k</td>
<td>3.0k</td>
</tr>
<tr>
<td>UDS-DurationQA</td>
<td>39.9k + 4.9k</td>
<td>4.8k</td>
</tr>
</tbody>
</table>

---

**Experiments**

**Two-stage Fine-tuning**

1. Irrelevant Contexts Removal
2. Question Generation
3. Candidate Answer Generation
   - Positive answer generation
   - Negative answer generation

---

**Results**

Model | EM  | F1  |
------|-----|-----|
RoBERTa-large → McTACO-duration: 40.45 67.42 |
RoBERTa-large → UDS-T (duration cls.) → McTACO-duration: 39.49 64.95 |
RoBERTa-large → UDS-DurationQA (unit only) → McTACO-duration: 42.78 66.97 |
RoBERTa-large → UDS-DurationQA → McTACO-duration: 45.86 70.52 |

---

**Additional Experiments**

Are the setups that leverage two-stage fine-tuning more effective than multi-task learning?

<table>
<thead>
<tr>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-stage Fine-tuning</td>
<td>45.86</td>
<td>70.52</td>
</tr>
<tr>
<td>Multi-task Learning</td>
<td>41.72</td>
<td>66.93</td>
</tr>
</tbody>
</table>

How does our proposed method compare to a SOTA pretrained temporal common sense language model?

<table>
<thead>
<tr>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TACOLM (Zhou et., 2020) → McTACO</td>
<td>34.60</td>
<td>-</td>
</tr>
<tr>
<td>BERT-base → McTACO</td>
<td>33.76</td>
<td>60.98</td>
</tr>
<tr>
<td>BERT-base → UDS-DurationQA → McTACO</td>
<td>36.52</td>
<td>63.22</td>
</tr>
</tbody>
</table>