

A Supplemental Material

A.1 Perturbing Candidate Answers

Here we provide a few missing details from *Step 3* of our annotations (Section 3). In particular, we create collections of common temporal expressions (see Table 3) to detect whether the given candidate answer contains a temporal expression or not. If a match is found within this list, we use the mappings to create perturbations of the temporal expression.

Adjectives	Frequency	Period	Typical time	Units
early:late late:early morning:late night night:early morning evening:morning everlasting:periodic initial:last first:last last:first overdue:on time belated:punctual long-term:short-term delayed:early punctual:belated	always:sometimes:never occasionally:always:never often:rarely usually:rarely rarely:always constantly:sometimes never:sometimes:always regularly:occasionally:never	night:day day:night	now:later today:yesterday tomorrow:yesterday tonight:last night yesterday:tomorrow am:pm pm:am a.m.:p.m. p.m.:a.m. afternoon:morning morning:evening night:morning after:before before:after	second:hour:week:year seconds:hours:weeks:years minute:day:month:century minutes:days:months:centuries hour:second:week:year hours:seconds:weeks:years day:minute:month:century days:minutes:months:centuries week:second:hour:year weeks:seconds:hours:years month:minute:day:century months:minutes:days:centuries year:second:hour:week years:seconds:hours:weeks century:minute:day:month centuries:minutes:days:months

Table 3: Collections of temporal expressions used in creating perturbation of the candidate answers. Each mention is grouped with its variations (e.g., “first” and “last” are in the same set).

A.2 Performance as a function of training size

An intuition that we stated is that, the task at hand requires a successful model to bring in external world knowledge beyond what is observed in the dataset; since for a task like this, it is unlikely to compile an dataset which covers all the possible events and their attributes. In other words, the “traditional” supervised learning alone (with no pre-training or external training) is unlikely to succeed. A corollary to this observation is that, tuning a pre-training system (such as BERT (Devlin et al., 2019)) likely requires very little supervision.

We plot the performance change, as a function of number of instances observed in the training time (Figure 3). Each point in the figure share the same parameters and averages of 5 distinct trials over different random sub-samples of the dataset. As it can be observed, the performance plateaus after about 2.5k question-answer pairs (about 20% of the whole datasets). This verifies the intuition that systems can rely on a relatively small amount of supervision to tune to task, if it models the world knowledge through pre-training. Moreover, it shows that trying to make improvement through getting more labeled data is costly and impractical.

A.3 Annotation Interfaces

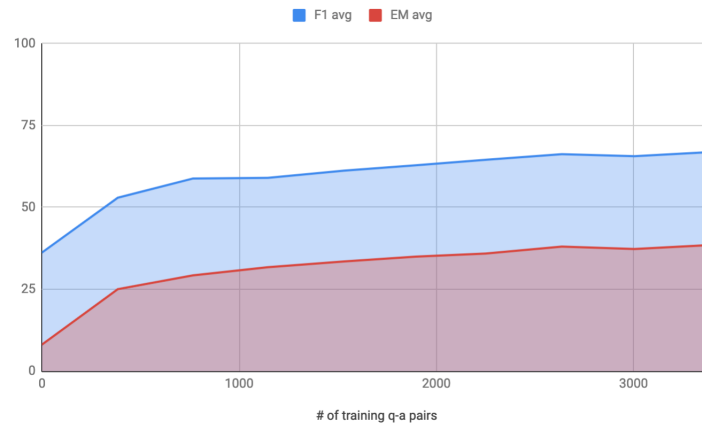


Figure 3: Performance of supervised algorithm (BERT; Section 4) as function of various sizes of observed training data. When no training data provided to the systems (left-most side of the figure), the performance measures amount to random guessing.

Sentence:

Ask a question regarding **Event Duration**

Question 1:

Answer 1:

Ask a question regarding **Transient v. Stationary**

Question 2:

Answer 2:

Ask a question regarding **Event Ordering**

Question 3:

Answer 3:

Ask a question regarding **Absolute Timepoint**

Question 4:

Answer 4:

Ask a question regarding **Frequency**

Question 5:

Answer 5:

Figure 4: Step 1

Sentence:

\$(sentence)

Question:

\$(question)

1. Provide a **plausible** answer to the question.

Give a good plausible answer here.

2. Provide a **negative/wrong** answer to question.

Give a negative answer here.

3. Do you think the given question needs temporal understanding?

☐ Yes

☐ No

4. Do you think the given question is related to \$(category)?

☐ Yes

☐ No

5. Do you think you **can** use something directly mentioned in sentence to answer the given question?

☐ Yes

☐ No

6. Do you think the given question is a valid question and free of grammatical and logical errors?

☐ Yes

☐ No

Figure 5: Step 2

Sentence:

\$(sentence)

Question:

\$(question)

Potential Answers (Scroll to see more):

\$(answer1)

☐ Likely to be an answer to the question

☐ Unlikely to be an answer to the question

☐ Garbage phrase (unusual characters, unclear meaning, typos, etc.)

\$(answer2)

☐ Likely to be an answer to the question

☐ Unlikely to be an answer to the question

☐ Garbage phrase (unusual characters, unclear meaning, typos, etc.)

\$(answer3)

☐ Likely to be an answer to the question

☐ Unlikely to be an answer to the question

☐ Garbage phrase (unusual characters, unclear meaning, typos, etc.)

\$(answer4)

☐ Likely to be an answer to the question

☐ Unlikely to be an answer to the question

☐ Garbage phrase (unusual characters, unclear meaning, typos, etc.)

\$(answer5)

☐ Likely to be an answer to the question

☐ Unlikely to be an answer to the question

☐ Garbage phrase (unusual characters, unclear meaning, typos, etc.)

Figure 6: Step 3