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

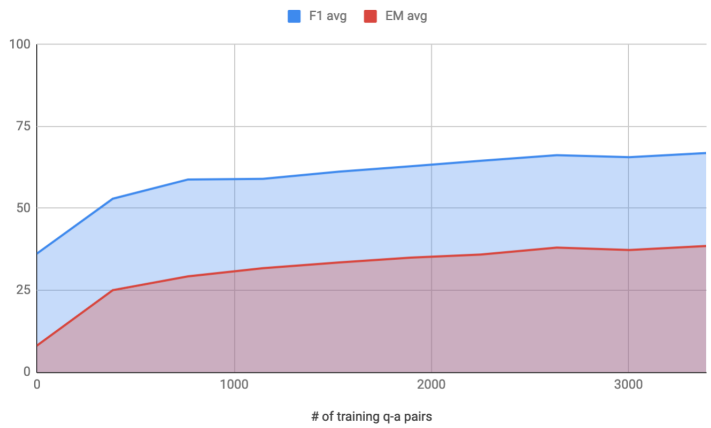

Figure 3: Performance of supervised algorithm (BERT; Section 4) as function of various sizes of observed training data. When no training data provided to the systems (left-most side of the figure), the performance measures amount to random guessing.

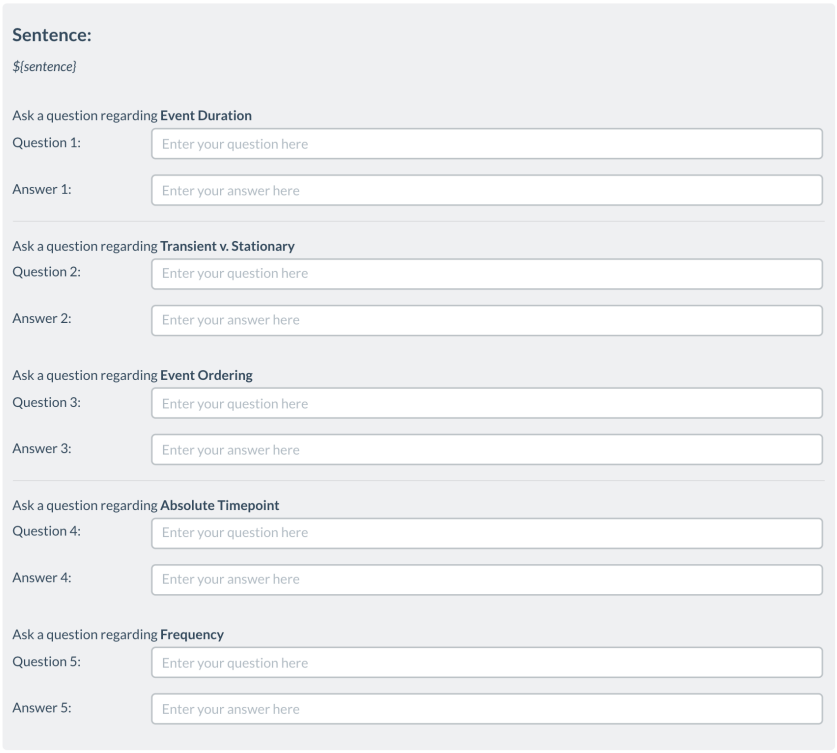

Figure 4: Step 1

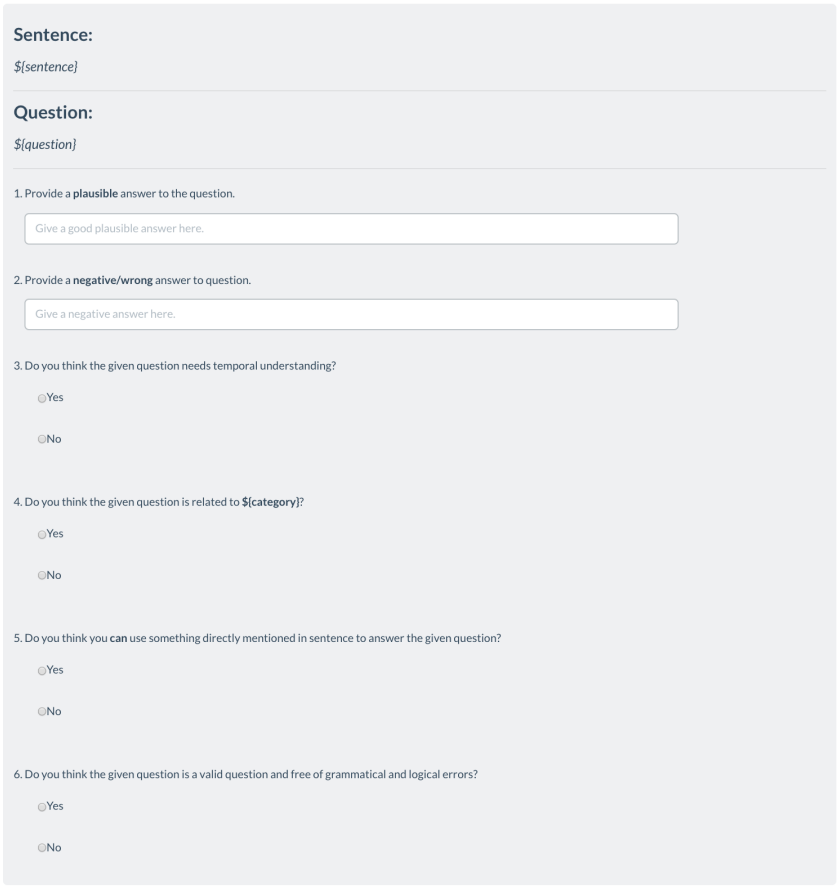

Figure 5: Step 2

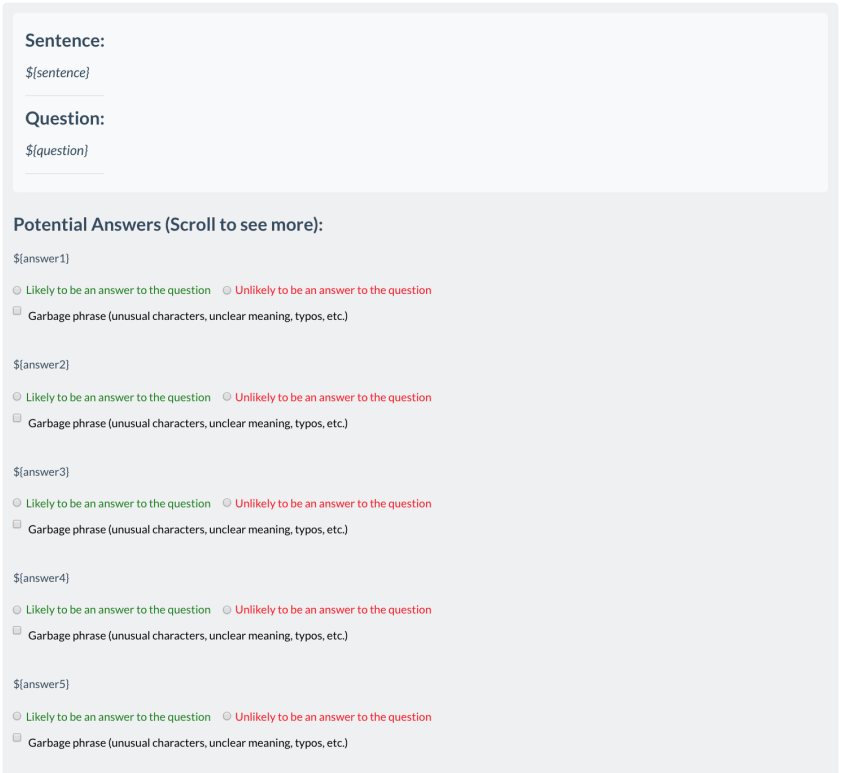

Figure 6: Step 3