# Temporal Event Reasoning using Multi-Source Auxiliary Learning Objectives

Xin Dong[1], Tanay Kumar Saha[2], Ke Zhang[2], Joel Tetreault[2], Alejandro Jaimes[2], and Gerard de Melo[3]

[1]Rutgers University [2]Dataminr Inc

[3]Hasso Plattner Institute / University of Potsdam

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Contact:
xd48@rutgeres.edu
gdm@demelo.org
www.demelo.org

## Motivation

Temporal event reasoning is vital in modern information-driven applications operating on news articles, social media, financial reports, etc.

- Question Answering samples from TORQUE

**Passage**: They were traveling in an up-armored high-mobility, multi-purpose, wheeled vehicle when this occurred. Those injured were evacuated by air to a nearby forward operating base for treatment.

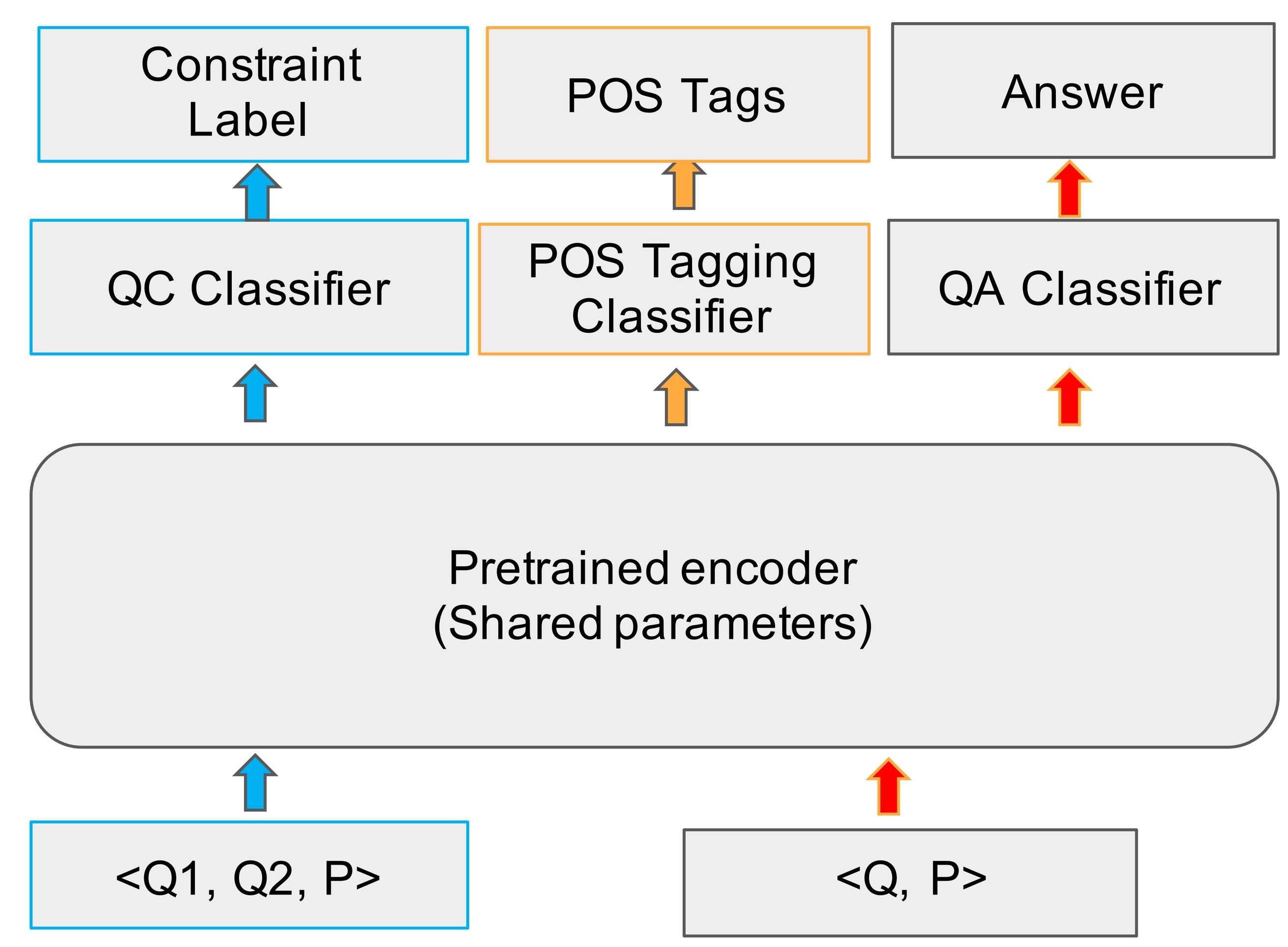
| Questions | Answers |
|---|---|
| What events have already finished? | traveling, occurred, evacuated |
| What will happen in the future? | No answer. |
| What events happened during their travel? | occurred, evacuated |
| What events have begun but has not finished? | treatment |
| What happened after it occurred? | evacuated, treatment |
| What happened before the injured were treated? | traveling, occurred, evacuated |

- Temporal information inferring from POS Tag

| Example | POS Tag | Temporal Information |
|---|---|---|
| People have **predicted** his demise so many times... | VBN: verb, past participle | event has happened |
| Security Council **passed** a resolution ... | VBD: verb, past tense | event happened |

## Method

Injects additional temporal knowledge into the pre-trained model from two sources:
(i) part-of-speech tagging
(ii) question constraints.
e.g., the set of answers to "What events have already finished?" and "What will happen in the future?" should typically be disjoint.




## Results and Analysis

### Experimental Settings

**TORQUE** : a reading com- prehension dataset of temporal ordering questions and answers. It provides 3.2k passages (~50 tokens/passage), 24.9k events (7.9 events/passage), and 21.2k user-provided questions. For end-to-end training, the task is modeled as a binary classification problem that requires predicting for each token in the passage whether it is an answer.

**MATRES** : a temporal relation (TempRel) extraction benchmark, consisting of 275 documents with entity relationships labeled as Before, After, Equal, or Vague.

**Metrics**: TORQUE is evaluated in terms of F1 score, Exact Match (EM), and Consistency (C). The latter is defined as the percentage of contrast groups for which a model's predictions have F1 ≤ 80% for all questions in a group. The contrast groups provided by TORQUE consist of questions with contrasting changes to the temporal keywords, e.g., "What happened after the snow started?" versus "What happened before the snow started?". For MATRES, we report standard micro-averaged F1 scores.

### TORQUE (Question Answering Setup)

| Method | F1 | EM | C |
|---|---|---|---|
| RoBERTa-Large [11] | 75.2 | 51.1 | 34.5 |
| RoBERTa-Large | | | |
| + Question CC | 75.7 | **51.3** | <u>36.2</u> |
| + POS Tagging | <u>75.8</u> | 50.7 | 35.6 |
| + POS Tagging + Question CC | **76.0** | <u>51.2</u> | **36.7** |

Results on TORQUE Dataset.

### MATRES (Relation Extraction Setup)

| Method | F1 |
|---|---|
| Want et al. [16] | 78.8 |
| RoBERTa-Large | 80.1 |
| + TORQUE | 80.6 |
| + TORQUE (Question CC) | 80.4 |
| + TORQUE (POS Tagging ) | 80.7 |
| + TORQUE (POS Tagging + Question CC) | **81.1** |

Results on MATRES Dataset.

### Influence of Amount of Training Data for TORQUE.

| Ratio | 30% | | | 50% | | | 100% | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **F1** | **EM** | **C** | **F1** | **EM** | **C** | **F1** | **EM** | **C** |
| RoBERTa-Large | 57.3 | 37.9 | 20.1 | 73.3 | 46.3 | 32.0 | 75.2 | 51.1 | 34.5 |
| Our Approach | 68.5 | 39.4 | 25.1 | 74.3 | 48.5 | 34.5 | 76.0 | 51.2 | 36.7 |
| *Improvement* (%) | 19.5% | 4.0% | 24.8% | 1.4% | 4.8% | 7.8% | 1.1% | 0.2% | 6.4% |

Results on TORQUE with different ratios of training data.