A Meta-framework for Spatiotemporal Quantity Extraction from Text

Qiang Ning, Ben Zhou, Hao Wu, Haoruo Peng, Chuchu Fan, Matt Gardner
amazon | Penn | HOORAY DATA | NEWS BREAK | Massachusetts Institute of Technology | Microsoft

1.Spatiotemporal Quantity Extraction

We propose a meta-framework STEQE for quantity extractions, which includes quantity span, type, spatial and temporal groundings.

2.Our Contribution

(A) We argue that spatiotemporal quantity extraction is **important** but **challenging**.

Important: provide rapid response to evolving situations (covid-19) and automatic extraction for sociopolitical analysis (elections).

Challenging: rarely studied by previous work, existing models are not sufficient.

(B) A meta-framework STEQE which includes problem definitions and a reusable extraction pipeline for domains and tasks.

(C) Annotated datasets on three example sociopolitical events: Covid-19, Black Lives Matter protests, and 2020 California wildfires.

(D) Experiments on existing models' performance on spatiotemporal quantity extraction, analysis and future directions.

4.Data Collection and Statistics

We use the STEQE pipeline to annotate three example datasets on Covid-19, BLM protests, and California wildfires for experiments and showcasing the pipeline. All domains go through the same annotation pipeline.

Domain	#Q Typing	#Q Space	#Q Time	#Q Test
Covid	1.5k	3.4k	4.3k	500
BLM	4k	1.5k	1.6k	500
Wildfire	2k	2k	1.6k	500

5. Experiment Setup

Typing: accuracy

Spatial Grounding: Exact Match (city/state level)

Temporal Grounding:

- overall num ending on DCT (binary)
- non-overall num start/end EM

End-to-end evaluation:

- Spatial EM-city + Temporal binary

Supervision: in-domain / all-domain

Naïve Baselines:

- Typing: predict most popular type.
- Spatial: predict the nearest location mention in text, relative to the quantity span.
- Temporal: predict overall num ending on DCT.
- End2end: same as the spatial and temporal.

Our Model: T5 with labels as output sequences

	T5 input	T5 gold output
Typing Model	..to <MARKER> 440 square kilometers..	Physical measurements
Spatial Model	..to <MARKER> 440 square kilometers..	US, CA, Los Angeles

Input Document

Document Creation Time: 2020-08-27

Title: 104 New USC Student Coronavirus Cases

Text: Log Angeles, CA – The number of coronavirus cases confirmed among USC students continued rising Thursday, with the university announcing **[104]** new cases over the past four days.

Extraction

Type: Confirmed Coronavirus Cases

Spatial Grounding: United States -> California -> Los Angeles -> USC

Temporal Grounding: [2020-08-23, 2020-08-26]

Figure 1: An example document, quantity extraction, type recognition and spatiotemporal grounding.

3. STEQE Formulation

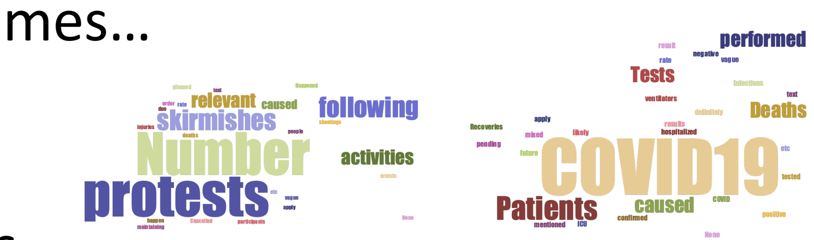
Raw Document **Quantity Recognition**

[DCT] [Title t_1, t_2, t_3, …] [Body Text b_1, b_2, …, $\boxed{q_1, q_2, \ldots q_m,}$ … b_n]

Type **Location** **Time Span**

Quantity Recognition: Finds a special type of numbers that are associated with **events**

- Non-quantity numbers: Date/Time, Duration, Entity names…
- Exclude article words and ordinal numbers



Quantity Typing: Domain-specific semantic types

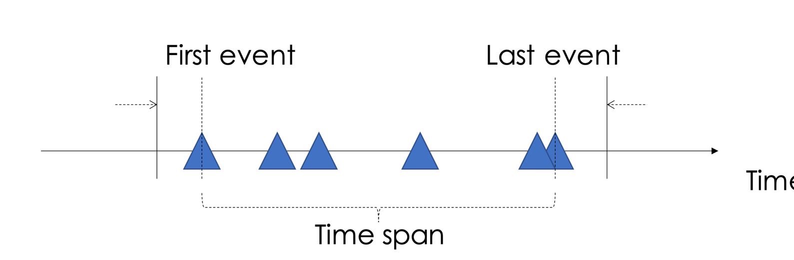
- Assume single-typing, ignore rate and money quantities

Spatial Grounding: Hierarchical country->state->city->free-span

South Portland, Maine – A facility for people with cognitive disabilities reports having **[six]** cases

Temporal Grounding:

- Best start/end timepoint estimation based on events
- Overall number: there have been **[3 million]** cases so far



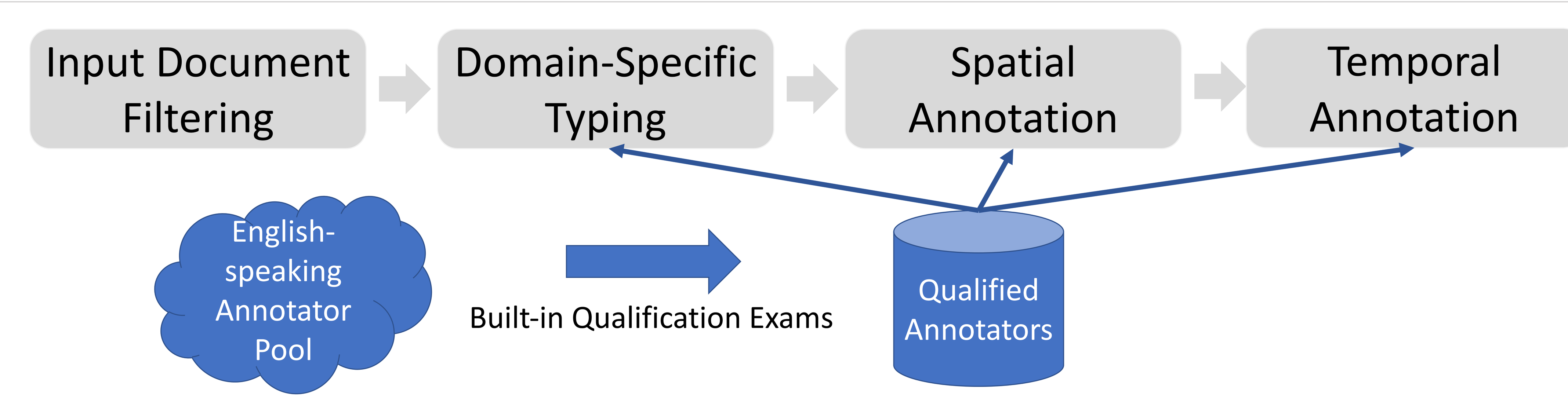
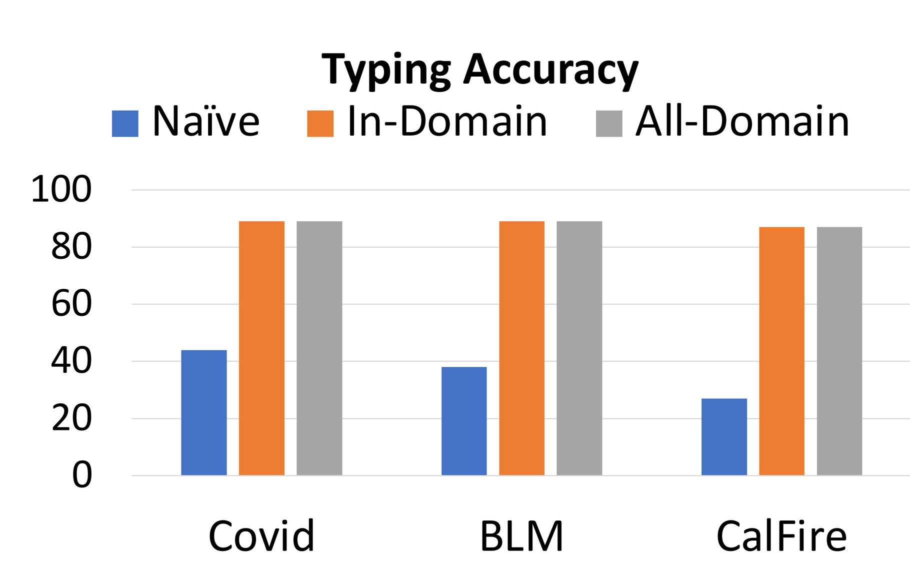
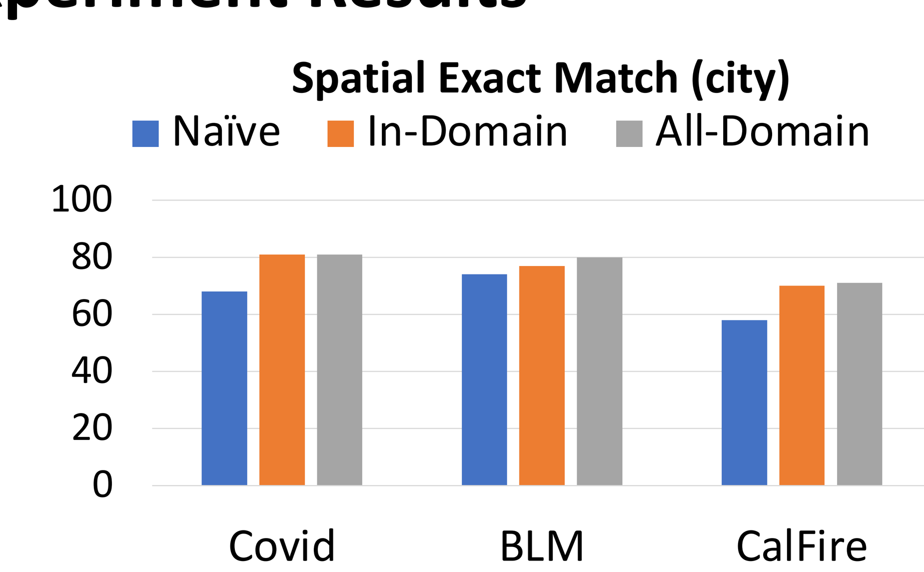
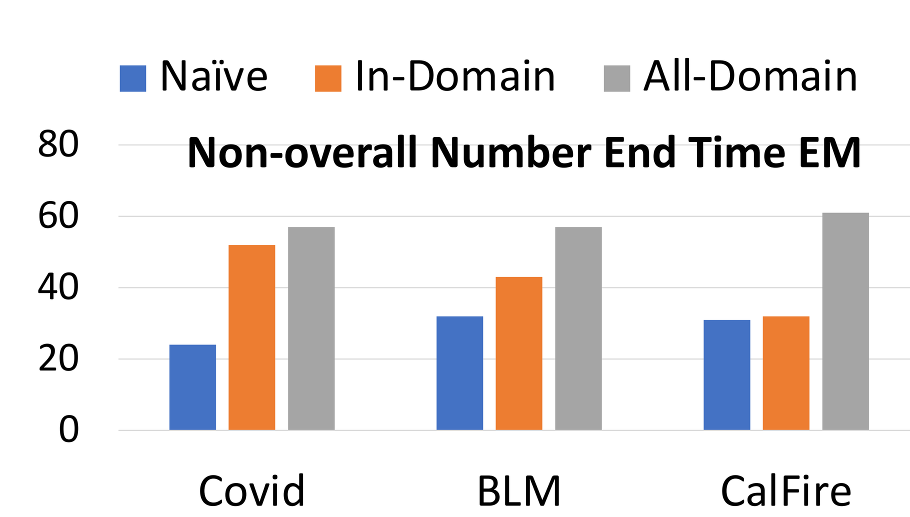


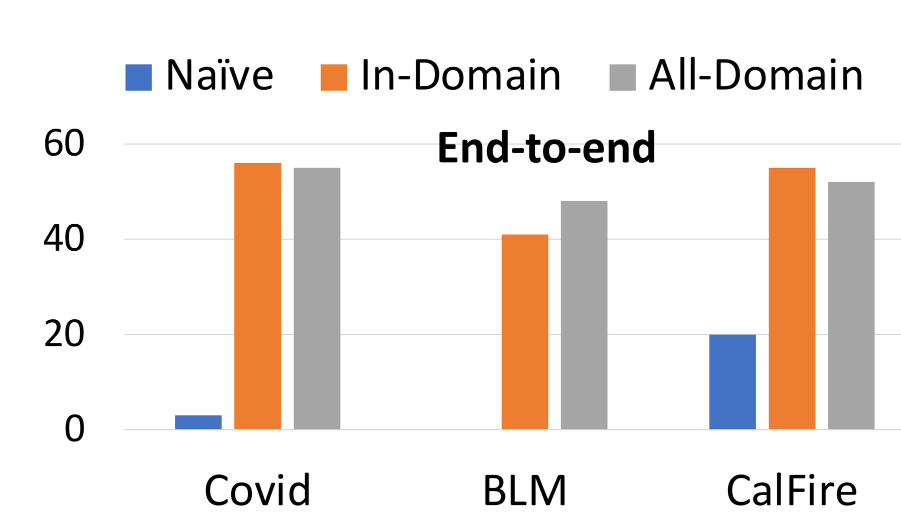
Figure 2: STEQE Shareable Annotation Pipeline

6. Experiment Results









https://github.com/steqe
Code & Data

Takeaways:

- T5 achieves usable performance in some settings but has room for improvement.
- All-domain supervision is generally better, especially in more difficult settings.