A Meta-framework for Spatiotemporal Quantity Extraction from Text

Anonymous ACL submission

Abstract

News events are often associated with *quantities* (e.g., the number of COVID-19 patients or the number of arrests in a protest), and it is often important to extract their type, time, and location from unstructured text in order to analyze these quantity events. This paper thus formulates the NLP problem of spatiotemporal quantity extraction, and proposes the first meta-framework for solving it. This meta-framework contains a formalism that decomposes the problem into several information extraction tasks, a shareable crowdsourcing pipeline, and transformer-based baseline models. We demonstrate the meta-framework in three domains—the COVID-19 pandemic, Black Lives Matter protests, and 2020 California wildfires—to show that the formalism is general and extensible, the crowdsourcing pipeline facilitates fast and high-quality data annotation, and the baseline system can handle spatiotemporal quantity extraction well enough to be practically useful. All resources of this paper will be released for future research on this topic.

1 Introduction

Events are often associated with *quantities* – how many COVID-19 patients are on ventilators, how many people are injured during protests, or how large is the extent of a wildfire. We often need to figure out the event type, and where and when it happened for these quantities for coherent discussion of public policy on sociopolitical events in rapidly evolving situations: *"19 deaths"* is different from *"19 recoveries;" "19 deaths in a small city yesterday"* apparently describes a more severe situation than *"19 deaths in the whole country last month."* However, until dedicated channels are established, these quantities are typically first reported on social media and local news articles, which then have to slowly make their way to some aggregate location for decision-makers to use.

DCT: Thursday, 08/27/2020
Title: Study Sessions, Dinners: 104 New USC Student Coronavirus Cases
Text: LOS 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...

Recognition: 104
Type: Confirmed cases
Spatial Grounding: US → California → Los Angeles → USC
Temporal Grounding: [08/23/2020, 08/26/2020]

DCT: Monday, 06/01/2020
Title: Black Lives Matter: 16 Organizations That Are Bailing Out Protestors
Text: ...Police officers have arrested [thousands] of demonstrators...

Recognition: thousands
Type: Arrests
Spatial Grounding: US
Temporal Grounding: Overall quantity ending on 06/01/2020

Figure 1: Given document creation time (DCT), title, and text, the STEQE problem is to do quantity recognition, typing, spatial grounding, and temporal grounding according to the proposed formalism (Sec. 2). Above are two examples from our COVID-19 dataset and BLM protest dataset.

Prior work on *events* focused on extracting event mentions, attributes, and relationships (ACE, 2005; Chen and Ji, 2009; Do et al., 2011; UzZaman et al., 2013; Glavaš et al., 2014; Zhou et al., 2019; Chen et al., 2021), and paid little attention to quantities associated with those events, which presents an opportunity to perform targeted information extraction on these *quantity events*.

This paper studies spatiotemporal quantity extraction (STEQE): finding quantities of certain types and extracting their associated times and locations. We develop a general meta-framework to help researchers overcome challenges and extend to new domains easily. Specifically, the contributions of this meta-framework are:

Task Formulation We draw on ideas from existing NLP tasks to create the first formalism that defines STEQE as four information extraction tasks:

quantity recognition, typing, spatial grounding, and temporal grounding. While each of these has analogues in the literature, our combination of them into a complete picture of quantity events is novel.

Annotation Collection We release a shareable and extensible crowdsourcing pipeline on CROWDAQ (Ning et al., 2020a) that facilitates fast and reliable data annotation. We show how this pipeline facilitates fast and high-quality annotations for three sociopolitical events: the COVID-19 pandemic, Black Lives Matter (BLM) protests, and 2020 California wildfires. These practical STEQE datasets are also released to foster future research.

Modeling We propose a T5 baseline model for its flexibility across tasks and easy domain transfer. This model shows that, while the end-to-end STEQE problem remains challenging in all domains, temporal grounding is typically the most difficult task, pointing out a research direction next.

2 STEQE

The STEQE problem aims to extract information about quantity events in text, consisting of four parts: determining which numerical expressions actually correspond to events (§2.1), the type of the event that a quantity is referring to (§2.2), where that event happened (§2.3), and the temporal extent to which the quantity refers (§2.4).

Note that for each of these subparts, there could have been other definition and formulation choices. We describe our formalism's design choices, and discuss why they could lead to better-defined learning problems and more reliable data collection, along with their limitations and how to extend our formalism for more advanced use cases.

2.1 Quantity Recognition

Similar to named entity recognition (NER) (Tjong Kim Sang and De Meulder, 2003), quantity recognition is defined as a text span detection problem. We discuss two questions regarding the definition of *quantities*: (1) how to distinguish between quantities and non-quantities; (2) how to define the span for quantities to avoid misalignment.

First, quantities are a special type of numbers that are associated with events, either in digits (e.g., *"123"*) or in words (e.g., *"one hundred twenty three"*). Some non-quantity examples:

1. Date and time: *"May 8, 2020"* and *"5:30 pm"*
2. Duration: *"3 months"* and *"60 years old"*
3. Part of an entity name: *"COVID-19"*, *"Porsche 911"*, and *"502 Main Street"*

Article words, *"a"* and *"an"*, require more attention. When we say *"a man died,"* the *"a"* does mean *"1"* death, while in *"a large number of people died,"* the *"a"* itself does not have the meaning of *"1,"* and we thus do not consider it a quantity.

Ordinal numbers can also indicate events, but their spatiotemporal extent can be understood differently: *"the fifth case in Seattle"* implies that there had been 5 cases, and the spatiotemporal extent of *"fifth"* can be that of the fifth case only, or all of the five cases. Ordinal-number events are rare in our study, so comparing to the extra annotation requirement, we decide to consider ordinal numbers as non-quantities, although the definition is easily extensible to cover them in the future.

Second, we need to define the boundaries of these quantity spans. For instance, in *"five cases in Seattle,"* should one mark *"five"* or *"five cases"*? What about *"4.8 billion"* and *"$4.8 billion"*? Similar to marking an event using its predicate only, our rule is to keep the span minimal while keeping the *numerical* semantics: we will mark *"five"* (i.e., drop *"case"*), *"4.8 billion"* (i.e., keep *"billion"*), and *"4.8 billion"* (i.e., drop *"$"*) in these examples. Minimising the span does not lose information about the quantity—only marking *"five"* in *"five cases"* does not prevent us from labeling its type and spatiotemporal extent in subsequent annotation tasks. Below are some arguably tricky cases, where quantities are in brackets.

1. Rate: *"[20 percent] of the tenants were infected"*, *"the positive rate is now [200] per [100,000]"*, *"[1000] tests per day"*
2. Approximation: *"[4 or 5] are missing"*
3. Range: *"the positive rate is [2 to 3 percent] / at least [2%] / at most [3%]"*

2.2 Quantity Typing

Again, similar to NER, recognized quantities can have an associated type from a predefined set of classes.[1] A clear event type is important for subsequent spatiotemporal grounding, but some quantities can have multiple types, and some can have multiple interpretations for their spatiotemporal extent. This work thus makes two design choices to mitigate these issues.

[1]The set of types in a STEQE problem will be domain-specific. We will explain the label set for typing for each of the 3 domains studied in this work later in §3.2.

Enforce single-typing In this work, we allow quantities to have only one single type. This ensures annotation quality since multiple types for a single quantity may complicate the spatiotemporal extent. For instance, in *"[three] men were hospitalized 5 days after being tested positive,"* the time span of hospitalization and that of tested positive are different. We enforce single-typing by providing an *order* of importance. For instance, hospitalization is more important than tested positive, so the spatiotemporal extent of *"three"* will be that of hospitalizations.

Ignore rate and money quantities Rate and money quantities are excluded in all of our typing labels, because their spatiotemporal extent can be interpreted in different ways. For instance, the spatiotemporal extent of *"a bill of $4.8 billion"* can be interpreted either as when and where this bill was passed, or as when and where the bill will be used; similarly, to define the time span of the rate quantity *"[20%] of the tenants were infected"*, we can either use the time span from the very first case to the last case that brought the infection rate from 0% to 20%, or use the time span when the infection rate was holding at 20%. For applications where one needs to spatiotemporally ground rate and money quantities, one could extend our instructions to clarify the ambiguities above.

2.3 Spatial Grounding

The spatial grounding problem of STEQE is to ground real-world events to a locale (see Fig. 5 in Appendix), avoiding complications in applications like human-robot interactions (e.g., *"turn left and go to the kitchen, and then pick up the fruit on the table"*). Thus we do not need to handle the nuances of relative spatial relationships like *"the kitchen is on our left"* and *"the table is in the kitchen."* We describe our formalism in terms of the format, granularity, and multi-location handling.

Format An important decision for spatial grounding is the *format*: we can use natural language to describe the locale, select text spans from the original text, or select from a map directory. In this work, we use a combination of all three for spatial grounding to balance between flexibility and consistency: we choose from a predefined set of questions to determine the country (U.S. vs non-U.S.) and state, use free text for the name of the city, and span selection for more granular locale information (e.g., *"a pork plant"*). We leave it for future work if one wants to extend to other countries, or if one can provide a detailed map directory.

Title: Six COVID-19 cases emerge in South Portland

Text: SOUTH PORTLAND, Maine -- A facility for people with cognitive disabilities reports having [six] COVID-19 cases...

Spatial grounding for [six]: US → Maine → South Portland → A facility for people with cognitive disabilities

Figure 2: The desired spatial grounding annotation is the most specific location mentioned in the text that contains all individual events.

Granularity We define the spatial grounding annotation to be the most specific location mentioned in the text that contains all individual events. For instance, in Fig. 2, the title mentions 6 cases in "South Portland," but later we will see that the 6 cases are all from "a facility for people with cognitive disabilities." The annotation should specify that facility instead of stopping at "South Portland." This design choice requires annotators to check the context in addition to the sentence containing the quantity, and is important for downstream tasks because it is likely that there are cases in South Portland but not in that facility.

Multi-location We handle events in multiple locations by broadening the granularity of the spatial location, as mentioned above. However, there are cases where the same quantity is explicitly mentioned with two or more separate locations:

1. *"Both Seattle and Tacoma had [10] new cases."*
2. *"Seattle and Tacoma together had more than [10] new cases."*

The *"10"* in both sentences above are associated with two cities, Seattle and Tacoma. The semantics are also different: being shared by two locales, or the events from both locales combine to make this quantity. In our pilot studies, we tried to consider these details in multi-location quantities, but found that they were very rare and crowd workers could not capture them reliably. We thus decide to ignore these cases in this work and only allow crowd workers to select a single location.

2.4 Temporal Grounding

The temporal grounding problem of STEQE is to ground each real-world quantity event to a single time span, which reduces the complexities in temporal semantics often encountered in prior datasets

(Pustejovsky et al., 2003; Cassidy et al., 2014; O'Gorman et al., 2016; Ning et al., 2018a, 2020b) and improves practicality.

Format A time span consists of two time points, and the key is the format for time points. In this work, we allow a time point to be UNKNOWN if the text is unclear. For a specific time point, there are two general ways to describe it: (1) use absolute date and time (e.g., *"Feb 1st, 2021"*); (2) use relative time Δ based on a reference point T (e.g., *"3 days before lockdown"*).

We have chosen the first method in this study, and when a time point is unclear based on the text, we allow annotators to simply select *"Unknown"*. The second method above is strictly more expressive, but also comes with many degrees of freedom: the reference point T can be either an absolute date and time T_{time} or another event T_{event} (e.g., *"lockdown"*), and the relative time difference Δ can be either a specific duration Δ_{spec} like *"3 days before/after"* or a rough description Δ_{rough} like *"a few days before/after."* In our pilot studies allowing for $T_{time} + \Delta_{rough}$, $T_{event} + \Delta_{spec}$, or $T_{event} + \Delta_{rough}$, we found the $T + \Delta$ method too flexible to achieve annotation agreement; in the meantime, using absolute date and time could reliably estimate those time spans in practice.

Granularity Given the nature of news events, it is often enough to be specific up to *days*. We define the time span of a quantity to be from the day of first event to the day of the last,[2] but this exact time span may not always exist in the text, so STEQE uses the best over-estimate of this gold time span based on information in the text (see Fig. 3).

This work also addresses common ambiguities. (1) Some time expressions are not critical and thus less specific in text, e.g., *"March 2020,"* for which we will simply use the entire span of that range, e.g., [03/01/2020, 03/31/2020]. (2) For time expressions like *"mid September"* and *"end of 2020"*, we choose the closest dates, e.g., *"09/15"* and *"12/31/2020"*. (3) Depending on the actual publication date and the content of an article, there can be different interpretations for *"today,"* thus leading to a one-day disagreement among people regarding time expressions like *"yesterday"* or *"in the last three days."* We allow our annotators to use their best judgment in these cases.

[2]If these events are durative, then accordingly, the time span should change to the day when the first event *started* to the day when the last event *ended*. We did not find it necessary to point this out in our data collection.

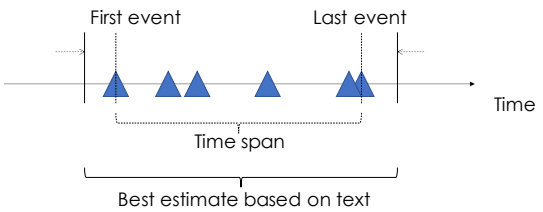

Figure 3: We define the time span of a quantity to start from the first event and end at the last; the desired temporal grounding annotation is the tightest estimate based on the text that covers all 6 events.

Multi-span Similar to spatial grounding, we handle events in multiple time spans by broadening the granularity of the time span, as mentioned above, and as with spatial grounding, we do not label multiple time spans separately in rare cases like *"10 arrests on Monday and Wednesday."*

Overall quantity A special type of temporal grounding phenomenon is *overall quantities*. Strictly speaking, this notion exists for spatial grounding as well (e.g., the overall COVID-19 case number around the world or the U.S.). While humans easily agree on the spatial extent of these overall quantities, their time spans are often ambiguous, especially the start time. For instance, in *"there have been [3 million] cases so far,"* the start time is supposed to be *"the beginning of the pandemic,"* but people do not always agree on when that was. The disagreement comes from (1) the pandemic started at different times in different regions of the world; (2) one may argue that the pandemic started either since the first confirmed case, or since the lockdown. This debate over start-time is not an NLP problem, so instead of inventing a new mechanism to resolve this, we simply allow "overall" as a label for the start time of a quantity.

3 Data Annotation

We have walked through the definition of the tasks in our STEQE framework, with discussions on various design choices. Next we explain how to collect annotations via this framework in practice.

3.1 Input Document Filtering

We worked with NewsBreak Inc.,[3] a local news aggregation company, to obtain raw newswire texts from publicly available news outlets. We then made

[3]https://www.newsbreak.com/

use of NewsBreak's internal tools to determine the topic of these news articles, i.e., whether an article is about COVID-19, Black Lives Matter protests in 2020, or the 2020 California wildfires. The data also comes with meta information including each article's source domain and publication time. Altogether, we obtain 1M articles on COVID-19 between 01/01/2020 and 12/31/2020, 100k on protests from 05/22/2020 to 12/31/2020, and 90k on California fires from 08/01/2020 to 12/31/2020 as source articles.

3.2 Domain-specific Typing

Following the general guidelines in §2.2, we used the following domain-specific types in this study.

1. COVID-19 pandemic: deaths caused by COVID-19, deaths likely caused by COVID-19, recoveries, confirmed cases, tests, tested negative, hospitalizations, patients on ventilators, and in ICUs.
2. BLM protests: protests, participants, order maintainers, arrests, deaths, injuries, and shootings.
3. California fires: fires, physical measurements, people impacted, items impacted, and resources.

These domain-specific types can be very specific (see those for the COVID-19 pandemic) or generic (see those for California fires), which demonstrates the flexibility of our framework.

3.3 Shareable CROWDAQ Pipeline

CROWDAQ (Ning et al., 2020a) is an open-source platform that standardizes data annotation pipelines and provides a customizable annotation interface, automated annotator qualification exams, progress monitoring, and annotation agreement monitoring.[4] CROWDAQ pipelines have four components: instruction, tutorial, exam, and main task: an annotator will read the instruction and tutorial, and then work on a set of multiple-choice exam questions. CROWDAQ automatically checks their scores and assigns qualifications. Qualified annotators will then be able to work on the main task. For each of the four tasks defined in Sec. 2, we have designed CROWDAQ pipelines that are general enough to be used for annotating in all domains.[5] We release the CROWDAQ pipelines for public use.[6]

[4] http://www.crowdaq.com/

[5] The only change for a new domain is instructions and exams for quantity typing, which have to be domain-specific.

[6] Please see the description at https://dev2.crowdaq.com/w/instruction/steqe/readme.

3.4 Data statistics

We first show statistics of our qualification exams in Table 1. We can see quantity recognition expectedly has the fewest hard questions and highest passing rate, and spatial and temporal grounding have more hard questions. Note that typing for California fires seems harder than typing for the other two domains, likely due to our choice of more generic types for California wildfires.

Qual ID	Qual Name	Hard (%)	Passed (%)
$\mathcal{Q}$	Recognition	18	94
$\mathcal{SG}$	Sp. Grd.	47	62
$\mathcal{TG}$	Temp. Grd.	50	57
$\mathcal{T}$-C	Typing (COVID)	27	60
$\mathcal{T}$-B	Typing (BLM)	36	60
$\mathcal{T}$-F	Typing (Fire)	50	53

Table 1: The difficulty of the qualification exams in this work. **Hard**: exam questions where less than 70% attempts were correct. **Passed**: the ratio of passed in all attempts. See Table 4 in the appendix for more details.

We then launched main annotation tasks on Amazon Mechanical Turk (MTurk) that were available only to qualified workers. We also required 3 different workers for each single annotation job and used majority voting to aggregate multiple workers' annotations. Since quantity recognition is a relatively easy task and our quantity recognition system based on BERT (Devlin et al., 2019) for the COVID domain was reliable enough to be applied to other domains, we did not further collect quantity recognition data. Table 2 and Table 5 (Appendix) show more statistics of these datasets.

Note that we did not enforce full annotation for all quantities (i.e., one quantity may only receive typing annotations, and another may only receive spatial annotations) to cover more documents (Ning et al., 2019a). Within those reported in Table 2, 500 quantities in each domain are fully labeled with both typing and spatiotemporal extent, and we use these as our test sets.

We paid \$0.05 for each job in quantity recognition, and \$0.15 for those in typing, spatial grounding, and temporal grounding; in the COVID-19 data collection, the average hourly pay of the top 5 annotation contributors was \$25 (typing), \$13 (spatial grounding), and \$12 (temporal grounding). In total, the cost of 3 datasets was \$11k (including 20% overhead paid to MTurk).

We developed our CROWDAQ pipeline for COVID-19 and applied it on other domains. When

Task	QID	#W	#Q	WAWA	Expert
Recog.					
- COVID	$\mathcal{Q}$	58	2.6k	92%	98%
Typing					
- COVID	$\mathcal{Q}$, $\mathcal{T}$-C	52	1.5k	95%	100%
- BLM	$\mathcal{Q}$, $\mathcal{T}$-B	74	4k	87%	94%
- Fire	$\mathcal{Q}$, $\mathcal{T}$-F	68	2k	91%	96%
Sp. Grd.					
- COVID	$\mathcal{T}$-C, $\mathcal{SG}$	91	3.4k	91%	98%
- BLM	$\mathcal{T}$-B, $\mathcal{SG}$	50	1.5k	80%	96%
- Fire	$\mathcal{T}$-F, $\mathcal{SG}$	63	2k	92%	90%
Temp. Grd.					
- COVID	$\mathcal{T}$-C, $\mathcal{TG}$	132	4.3k	86%	100%
- BLM	$\mathcal{T}$-B, $\mathcal{TG}$	57	1.6k	77%	96%
- Fire	$\mathcal{T}$-F, $\mathcal{TG}$	63	1.6k	82%	96%

Table 2: The required qualifications (QID), numbers of actual annotators (#W) and annotated quantities (#Q), worker agreement with aggregate (WAWA), and expert evaluation on 50 random samples after worker aggregation. The WAWA metric is for the "state" choice in spatial grounding, and the "overall number" judgment in temporal grounding (reported by CROWDAQ directly). The expert evaluation scores are all accuracy, except for F_1 for quantity recognition.

we received news articles in BLM protests and California wildfires from NewsBreak Inc., it only took us about 2 weeks to obtain the annotations used in this work, including designing domain-specific typing instructions and exams, launching tasks to MTurk, and waiting for crowd workers to finish. This fast and reliable data collection is appealing for responding to emerging events in the future.

4 Model

Quantity recognition is a typical span selection problem and we use the standard token classification model based on BERT (large, cased) (Devlin et al., 2019) that comes with HuggingFace (Wolf et al., 2020). For typing, spatial, and temporal grounding, we use the T5-large language model (Raffel et al., 2020) for its flexibility across tasks and easy domain transfer. We format data from each task to fit into T5's sequence to sequence (seq-to-seq) nature. Specifically, for each quantity, the input sequence to T5 is the string of the previous 3 sentences, the current sentence with a special marker token right before the quantity span, the next 3 sentences, the title, and document creation time (DCT). For typing, the output sequence is a single token representing each label mapped from a reserved vocabulary. For spatial grounding, the output sequence is the location names from the highest hierarchy to the lowest ended by an end-of-sentence (EOS) marker. For temporal grounding, the output sequence is the start time followed by the end time. Both times are either *"unknown"* or a date string in ISO 8601 format (e.g., *"2021-01-15"*). We view the start time of an overall quantity as *"unknown"*. To get complete date predictions, we enforce the decoding length to be at least 12 and use a date parser to find *"unknowns"* or dates.

5 Experiments

In our evaluation of **quantity recognition** using the aforementioned BERT model on a random set of 300 sentences (100 from each domain), we find the precision 99% for all domains, and the recall 95% (COVID), 87% (BLM), and 87% (Fire). The recall is slightly lower because of poor performance on article words (*"a"* and *"an"*). However, since most missed quantities are not associated with event types that we care about (e.g., *"[a] post office"* or *"[a] comment"*), the adjusted recall is 98% (COVID), 94% (BLM), and 93% (Fire) if we do not consider those irrelevant quantities.

Table 3 shows system performances on **typing**, **spatial**, and **temporal grounding** on extracted quantities. Our test set in each domain consists of 500 fully annotated quantities. The rest of the data is split into 80% for training and 20% for development, that we use to acquire the learning rate (5e-3) and batch size (32). We compare T5 with a naive method, which always predicts the majority type in each domain for "typing," the location mention closest to the quantity in text for "spatial grounding,"[7] and overall quantity ending on DCT for "temporal grounding." For **spatial grounding**, we report two exact match (EM) scores, up to the state-level and city-level, respectively. For **temporal grounding**, we report the accuracy for judging whether a quantity is *an overall quantity ending on DCT* ("Binary" in Table 3), and two EM scores for cases where the gold start time is a specific date ("S-N" for "Start-Nontrivial") and where the end time is not DCT ("E-N" for "End-Nontrivial").

T5 (in-domain) On quantity typing, T5 improves by a large margin over the naive baseline in all domains. The naive baseline performs reasonably well on spatial grounding at the state level (82-92% EM-state across three domains), but often fails to provide more granular information at the

[7] This assumes world knowledge of geo-hierarchies, e.g., "L.A." is in California.

System	Task	Typing	Spatial Grounding		Temporal Grounding			End-to-end
		Acc	EM-city	EM-state	Binary	S-N	E-N	EM-city, Binary
Naive	COVID	44	68	84	68	0	24	3
	BLM	38	74	82	32	0	32	0
	Fire	27	58	92	**86**	0	31	20
T5 (in-domain)	COVID	**89**	80	90	**74**	56	43	55
	BLM	**89**	77	89	58	51	43	43
	Fire	**87**	**71**	**94**	84	4	31	**56**
T5 (all domains)	COVID	**89**	**82**	**91**	**74**	**58**	**54**	**55**
	BLM	**89**	**81**	**91**	**66**	**67**	**58**	**49**
	Fire	**87**	70	93	75	**50**	**56**	51

Table 3: System performances on typing, spatial grounding, and temporal grounding. **EM-city/-state**: exact match scores up to the city-/state-level. **Binary**: judging if a quantity is an overall-quantity ending on DCT. **S-N/E-N**: EM scores when the start/end time is non-trivial (see text for more details). **End-to-end**: quantities receiving correct predictions on all steps based on "EM-city" (spatial) and "Binary" (temporal). **T5 (all domains)** uses the same typing systems trained in-domain, but combine the spatiotemporal grounding data from all domains in training.

city level (58-74% EM-city). This is expected because a city mentioned close to the quantity does not necessarily mean that the quantity is for the city.[8] This phenomenon also varies across domains: BLM protests were in a few major cities, the EM-city score of the naive method is thus relatively high (74%), while for Calfornia wildfires, there were more cities to choose from, leading to a low EM-city of 58%. In contrast, T5 can produce more granular information at the city level, and maintain a relatively stable score across domains (71-80% EM-city). As for temporal grounding, due to the nature of news articles, the naive baseline that treats all quantities as an overall quantity ending on DCT yields reasonably good performances in all domains; but for quantities with a non-trivial start time or end time, the naive baseline largely fails.

T5 (all domains) We also combine the training data for spatiotemporal grounding from all domains and train a single T5 system (but keep T5 in-domain systems for typing), which achieves the best scores for almost all metrics in Table 3. One outlier is the Fire domain, where spatial grounding scores and the Binary score for temporal grounding drop, probably due to skewness in Fire data (e.g., most of the Fire quantities are cities in California and overall quantities). This suggests that spatiotemporal phenomena can be generally transferred across different domains.

Finally, the **end-to-end** column in Table 3 shows how many of these quantities have received correct predictions on typing, spatial grounding (based EM-city), and temporal grounding (based on "Binary"). The reported performance does not count for quantities that are not recognized, so we view this as the precision of the system. We see that the naive baseline has very low performance due to errors propagated at each step, while with this framework, T5 is trained to produce significantly better results. Note that depending on the use case, one can simply collect more training data, or focus on only a few important event types, to further improve the end-to-end performance.

6 Related works

Existing NLP works on events have focused on detection (e.g., detecting LIFE and BUSINESS events; ACE (2005)), common sense (e.g., Rashkin et al. (2018); Sap et al. (2019); Zhang et al. (2020a)), and relationships (e.g., coreferential Chen and Ji (2009), temporal UzZaman et al. (2013), causal Do et al. (2011), and parent-child relations Glavaš et al. (2014)). There is also a line of recent works specifically on temporal semantics: time expression extraction and normalization (Laparra et al., 2018), temporal relation extraction (Ning et al., 2018a, 2019b, 2020b), temporal common sense (Zhou et al., 2019, 2020), temoral slot filling (Surdeanu, 2013), and timeline construction (Do et al., 2012; Ning et al., 2018b; Li et al., 2019). Prior works on *quantities* either focus on math calculations (Roy et al., 2015; Roy and Roth, 2018) or common sense reasoning (e.g., mass distribution of animals; Elazar et al. (2019)), and not on *quantity events* and the associated spatiotemporal extent studied in this work.

[8] *"The State Department of Public Health in Springfield reports a total case of [268]."* is a quantity for the state.

Existing works on spatial semantics have focused on natural language navigation (Chen et al., 2019; Kim et al., 2020), human-machine interaction (Landsiedel et al., 2017; Roman Roman et al., 2020), dialogue systems (Udagawa et al., 2020), and clinical analysis (Kordjamshidi et al., 2015; Datta and Roberts, 2020). Works on geocoding (Gritta et al., 2018; Kulkarni et al., 2020) map spatial mentions to coordinates, which can be applied to our work for finer geolocation mapping.

A recent work from Zong et al. (2020), which extracts COVID-19 related events from tweets, is closely related to our work. Besides that they worked on tweets instead of news articles, the key differences are: (1) instead of span selection used in Zong et al. (2020), we propose formalisms deeper into the spatiotemporal extent of quantity events and capture more nuances in spatiotemporal semantics; (2) we show that our STEQE framework generally applies to multiple domains and not only for the COVID-19 pandemic; (3) we release our entire data collection pipeline on CROWDAQ for public use and extension.

7 Discussion

As §5 shows, the performance bottleneck of STEQE is mainly at temporal grounding: with almost perfect quantity recognition and very good typing and spatial grounding results, temporal grounding performance is typically much lower than the other tasks. While typing and spatial grounding are ready for practical research into few- and zero-shot settings along the lines of what is done in entity typing (Zhou et al., 2018; Obeidat et al., 2019; Zhang et al., 2020b), temporal grounding still requires more investigation even in in-domain settings.

Why is temporal grounding so challenging? First, news articles tend to mention many overall quantities ending on publication time, leading to imbalanced datasets. For instance, 86% in Fire fall into this category, leaving little training data for other quantities; in contrast, this number is only 32% in BLM, and the S-N and E-N scores are much higher in BLM than those in Fire. Second, temporal grounding often requires reasoning, an effect known to be difficult in many works on temporal semantics (Ning et al., 2020b; Zhou et al., 2021). For instance in Fig. 4, to figure out the time span of *"80,"* we need to understand that (1) it happened on *"Sunday"* (2) the *"Sunday"* is a Sunday in the past instead of in the future, and (3) it is most likely the most recent Sunday instead of earlier ones.

DCT: Tuesday, 06/16/2020
Text: Black Lives Matter demonstrators in a tiny Ohio town...Sunday. The small demonstration has about [80] people, organized by local Bethel residents.

Figure 4: The start time of *"80"* needs reasoning.

Another direction to improve on STEQE is to aggregate from multiple articles, given that the same quantity or similar quantities are typically covered by multiple sources. Cross-document event coreference has many unique difficulties (e.g., see Upadhyay et al. (2016); Bugert et al. (2020)), but knowing the quantity event type, location, and time span may make it relatively easy to find coreference to strengthen one's belief in its prediction, or demote outliers that are likely wrong predictions.

The proposed STEQE framework may also be used to detect misinformation and perhaps in social science studies too. For instance, we found that one website mistakenly reported Virginia's COVID-19 case number on Apr 2, 2020 to be 17k, while it should be 1.7k; we also found signs that news agencies might have mentioned case numbers in New York city less frequently after a sharp increase, but turned to report case numbers in New Jersey in April 2020. These social science analyses are beyond the scope of this work, but the examples above point to interesting potential uses of these information extraction systems.

8 Conclusion

Many important news events are associated with quantities. With practicality in mind, we dive deep into the semantics of quantity events and proposes a meta-framework for spatiotemporal quantity extraction: we formulate the problem as four information extraction tasks which lead to quick and reliable data annotation via crowdsourcing; we also build a T5 baseline to show the difficulties of the task and discuss transfer learning opportunities. We use this meta-framework to build datasets on three separate sociopolitical events: the COVID-19 pandemic, BLM protests, and California fires. Our meta-framework is shown to be readily extensible to different domains of quantity events, an appealing feature for quick response to future events. The new datasets we collect as examples of this framework can also directly contribute to future studies on spatiotemporal quantity extraction.

References

2005. The ACE 2005 (ACE 05) Evaluation Plan. Technical report.

Michael Bugert, N. Reimers, and Iryna Gurevych. 2020. Cross-document event coreference resolution beyond corpus-tailored systems. *ArXiv*, abs/2011.12249.

Taylor Cassidy, Bill McDowell, Nathanel Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 501–506.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. TOUCHDOWN: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Muhao Chen, Hongming Zhang, Qiang Ning, Manling Li, Heng Ji, Kathleen McKeown, and Dan Roth. 2021. Event-centric natural language processing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6–14.

Zheng Chen and Heng Ji. 2009. Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*.

Surabhi Datta and Kirk Roberts. 2020. A hybrid deep learning approach for spatial trigger extraction from radiology reports. In *Proceedings of the Third International Workshop on Spatial Language Understanding*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Goran Glavaš, Jan Šnajder, Marie-Francine Moens, and Parisa Kordjamshidi. 2014. HiEve: A corpus for extracting event hierarchies from news stories. In *LREC*.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2018. What's missing in geographical parsing? *Language Resources and Evaluation*, 52:603 – 623.

Hyounghun Kim, Abhaysinh Zala, Graham Burri, Hao Tan, and Mohit Bansal. 2020. ArraMon: A joint navigation-assembly instruction interpretation task in dynamic environments. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Parisa Kordjamshidi, Dan Roth, and Marie-Francine Moens. 2015. Structured learning for spatial information extraction from biomedical text: Bacteria biotopes. 16(126).

Sayali Kulkarni, Shailee Jain, Mohammad Javad Hosseini, Jason Baldridge, E. Ie, and L. Zhang. 2020. Spatial language representation with multi-level geocoding. *ArXiv*, abs/2008.09236.

Christian Landsiedel, Verena Rieser, Matthew Walter, and Dirk Wollherr. 2017. A review of spatial reasoning and interaction for real-world robotics. *Advanced Robotics*, 31(5):222–242.

Egoitz Laparra, Dongfang Xu, and Steven Bethard. 2018. From characters to time intervals: New paradigms for evaluation and neural parsing of time normalizations. *Transactions of the Association for Computational Linguistics (TACL)*, 6:343–356.

Manling Li, Ying Lin, Joseph Hoover, Spencer Whitehead, Clare Voss, Morteza Dehghani, and Heng Ji. 2019. Multilingual entity, relation, event and human value extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.

Qiang Ning, Hangfeng He, Chuchu Fan, and Dan Roth. 2019a. Partial or complete, that's the question. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019b. An Improved Neural Baseline for Temporal Relation Extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Qiang Ning, Hao Wu, Pradeep Dasigi, Dheeru Dua, Matt Gardner, Robert L. Logan IV, Ana Marasović, and Zhen Nie. 2020a. Easy, reproducible and quality-controlled data collection with CROWDAQ. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020b. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Qiang Ning, Hao Wu, and Dan Roth. 2018a. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018b. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Demo Track)*.

Rasha Obeidat, Xiaoli Z. Fern, Hamed Shahbazi, and P. Tadepalli. 2019. Description-based zero-shot fine-grained entity typing. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The TIMEBANK corpus. In *Corpus Linguistics*, page 40.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, M. Matena, Yanqi Zhou, W. Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 463–473.

Homero Roman Roman, Yonatan Bisk, Jesse Thomason, Asli Celikyilmaz, and Jianfeng Gao. 2020. RMM: A recursive mental model for dialogue navigation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Subhro Roy and Dan Roth. 2018. Mapping to declarative knowledge for word problem solving. *Transactions of the Association for Computational Linguistics*, 6:159–172.

Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics (TACL)*, 3.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

M. Surdeanu. 2013. Overview of the tac2013 knowledge base population evaluation: English slot filling and temporal slot filling. *Theory and Applications of Categories*.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Takuma Udagawa, Takato Yamazaki, and Akiko Aizawa. 2020. A linguistic analysis of visually grounded dialogues based on spatial expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Shyam Upadhyay, Nitish Gupta, Christos Christodoulopoulos, and Dan Roth. 2016. Revisiting the evaluation for cross document event coreference. In *Proc. of the International Conference on Computational Linguistics (COLING)*.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TEMPEVAL-3: Evaluating time expressions, events, and temporal relations. *Proceedings of the Joint Conference on Lexical and Computational Semantics (*SEM)*, 2:1–9.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Hongming Zhang, Xin Liu, Haojie Pan, Y. Song, and Cane Wing ki Leung. 2020a. Aser: A large-scale eventuality knowledge graph. In *Proceedings of the International World Wide Web Conferences (WWW)*.

Tao Zhang, Congying Xia, Chun-Ta Lu, and Philip Yu. 2020b. Mzet: Memory augmented zero-shot fine-grained named entity typing. *ArXiv*, abs/2004.01267.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ben Zhou, Daniel Khashabi, Chen-Tse Tsai, and Dan Roth. 2018. Zero-shot open entity typing as type-compatible grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal Common Sense Acquisition with Minimal Supervision. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, A. Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. 2020. Extracting COVID-19 events from Twitter. *ArXiv*, abs/2006.02567.

A Qualification setups

Note that exams for quantity recognition, spatial & temporal grounding are domain-agnostic, and exams for quantity typing are domain-specific. The way exams work on CROWDAQ is that we provide a pool of questions and CROWDAQ will randomly select a specified number of them. We also do not allow a crowd worker to make too many attempts. Table 4 shows the setup and statistics of those exams.

CROWDAQ provides diagnostic information on each question too. In Table 4, we also show the number of questions where less than 70% examinees were correct (i.e., "Hard"). The total number of attempts in each exam and how many of them got scores higher than the passing score are also reported.

Qual ID	Qual Name	Question Pool		CROWDAQ Configuration			Workers' Performance	
		#Total	#Hard	#Questions	#Attempts	Passing Grade	#Attempts	#Succeeded
$\mathcal{Q}$	Recognition	11	2	10	3	90	952	895
$\mathcal{SG}$	Spatial Grounding	17	8	12	3	90	1454	897
$\mathcal{TG}$	Temporal Grounding	12	6	10	3	90	1180	674
$\mathcal{T}$-C	Typing (COVID)	11	3	10	3	90	1156	698
$\mathcal{T}$-B	Typing (BLM)	11	4	8	3	85	760	457
$\mathcal{T}$-F	Typing (Fire)	14	7	12	3	90	905	476

Table 4: The qualification exam setups in this study. **Question Pool**: All the questions we provided to CROWDAQ; hard questions are those where less than 70% attempts were correct. **CROWDAQ Configuration**: #questions to display each time, #attempts allowed, and the required passing grade. **Workers' Performance**: the total number of attempts and succeeded.

B Corpus statistics

Table 5 shows a more complete version of our earlier Table 2. The extra columnes are the total number of qualified workers for each task, the Gini index, and the total number of sentences/documents annotated here. Gini is a metric proposed by TORQUE (Ning et al., 2020b) to measure the skewness of crowdsourcing data collection. Our Gini is significantly higher and we think the reason is that many crowd workers only attempted a couple our HITs. Regarding the definition of WAWA, we realize that Ning et al. (2020b) has provided a very good explanation about it; please refer to the appendix E of Ning et al. (2020b) about it.

Task	Worker Pool				Size			Quality	
	Req. Qual ID(s)	#Qualified	#Actual	Gini	#Quant.	#Sent.	#Doc.	WAWA	Expert
Typing (COVID)	$\mathcal{Q}$, $\mathcal{T}$-C	299	52	0.74	1.5k	1.5k	1.3k	95%	100%
Typing (BLM)	$\mathcal{Q}$, $\mathcal{T}$-B	291	74	0.53	4k	3.9k	3k	87%	94%
Typing (Fire)	$\mathcal{Q}$, $\mathcal{T}$-F	231	68	0.62	2k	2k	1.4k	91%	96%
Sp. Grd. (COVID)	$\mathcal{T}$-C, $\mathcal{SG}$	258	91	0.74	3.4k	3.3k	2.9k	91%	98%
Sp. Grd. (BLM)	$\mathcal{T}$-B, $\mathcal{SG}$	141	50	0.68	1.5k	1.5k	1.2k	80%	96%
Sp. Grd. (Fire)	$\mathcal{T}$-F, $\mathcal{SG}$	160	63	0.71	2k	2k	1.3k	92%	90%
Temp. Grd. (COVID)	$\mathcal{T}$-C, $\mathcal{TG}$	399	132	0.81	4.3k	4.2k	3.5k	86%	100%
Temp. Grd. (BLM)	$\mathcal{T}$-B, $\mathcal{TG}$	190	57	0.71	1.6k	1.6k	1.2k	77%	96%
Temp. Grd. (Fire)	$\mathcal{T}$-F, $\mathcal{TG}$	215	63	0.74	1.6k	1.6k	1.1k	82%	96%

Table 5: Corpus statistics. The required qualifications (QID), numbers of actual annotators (#W) and annotated quantities (#Q), worker agreement with aggregate (WAWA), and expert evaluation on 50 random samples after worker aggregation. The WAWA metric is for the "state" choice in spatial grounding, and the "overall number" judgment in temporal grounding (reported by CROWDAQ directly). The expert evaluation scores are all accuracy, except for F_1 for quantity recognition.

C Example annotations

Figure 5 shows two examples in each of the three domains in this study.

Domain, DCT	Quantity	Type	Spatial Grd.	Temporal Grd.
COVID-19 Sat, 2020-08-15	*Tennessee has conducted 1,757,690 tests with 1,631,297 negative results .*	*Test performed for COVID-19: result is negative*	US, Tennessee	Overall number ends at DCT
COVID-19 Wed, 2020-08-12	*Wyandotte County is reporting 4,895 confirmed cases...The county said on Tuesday that 99 people have died from the coronavirus since the start of the outbreak*	*Deaths: definitely caused by COVID-19*	US, Kansas, Wyandotte County	Overall number ends on 2020-08-11
Wildfires Mon, 2020-09-14	*...large fires across 10 states...At least 35 people have died in California , Oregon and Washington.*	*People impacted*	US	Overall number ends at DCT
Wildfires Tue, 2020-09-22	*The blaze had more than doubled in size over the past week to 170 square miles (440 square kilometers), ... from Los Angeles.*	*Physical measurements*	US, California, Los Angeles	2020-09-15 to 2020-09-22
Protests Tue, 2020-06-16	*Black Lives Matter demonstrators in a tiny Ohio town...Sunday. The small demonstration has about 80 people, organized by local Bethel residents.*	*Number of participants in protests or relevant activities*	US, Ohio, Bethel	2020-06-14 to 2020-06-14
Protests Sun, 2020-05-31	*A CNN analysis found about 80% of the 51 people booked into a Minneapolis jail during two days of protests are actually from Minnesota .*	*Number of arrests due to the protests or following skirmishes*	US, Minnesota, Minneapolis	unknown

Figure 5: Example annotations of quantity typing, spatial grounding, and temporal grounding across three domains. Quantity span is highlighted. Text snippets are cut short to only keep the sentence with the quantity and other relevant information.

D Reproducibility

For T5-based experiments related to model performances in Table 3, we choose the learning rate from [5e-2, 5e-3, 5e-4] and select 5e-3 for final experiments. We use a batch size of 32 and run 20 epochs for each setting. All parameters are tuned on the development set as described in §5. Experiments on average finish in 3 hours on a single Nvidia RTX 8000 GPU.