Characterizing Collective Attention via Descriptor Context in Public Discussions of Crisis Events

Ian Stewart  Diyi Yang  Jacob Eisenstein
Georgia Institute of Technology
School of Interactive Computing
Atlanta, GA
istewart6@gatech.edu, diyi.yang@cc.gatech.edu, jacobe@gatech.edu

Abstract
Collective attention is central to the spread of real world news and the key to understanding how public discussions report emerging topics and breaking news. Most research measures collective attention via activity metrics such as post volume. While useful, this kind of metric obscures the nuanced content side of collective attention, which may reflect how breaking events are perceived by the public. In this work, we conduct a large-scale language analysis of public online discussions of breaking crisis events on Facebook and Twitter. Specifically, we examine how people refer to locations of hurricanes in their discussion with or without contextual information (e.g. writing “San Juan” vs. “San Juan, Puerto Rico”) and how such descriptor expressions are added or omitted in correlation with social factors including relative time, audience and additional information requirements. We find that authors’ references to locations are influenced by both macro-level factors such as the location’s global importance and micro-level social factors like audience characteristics, and there is a decrease in descriptor context use over time at a collective level as well as at an individual-author level. Our results provide insight that can help crisis event analysts to better predict the public’s understanding of news events and to determine how to share information during such events.

1 Introduction
Today, millions of people experience and discuss news and events happening around the world through online media. Breaking news events, especially crisis events, often attract significant collective attention from the general public (Lin et al. 2014), resulting in bursts of discussion on social media (Leavitt and Clark 2014) [Lehmann et al. 2012]. During such events, public observers often focus on important locations, people or organizations (hereafter “named entities”) depending on their relevance to the unfolding crisis (Wakamiya et al. 2015). A spike in attention directed toward a particular location may signal an important update, such as the need for aid for the location (Varga et al. 2013). While collective attention is often measured with activity metrics such as post volume (Mitra, Wright, and Gilbert 2016), such metrics often focus on an aggregate quantity summary of attention without considering the nuanced content side of attention dynamics.

One way to model the content of collective attention is to examine how people talk about breaking news events, especially their descriptions of locations, which are a major component of crisis events. For instance, after Hurricane Maria struck Puerto Rico in 2017, more Americans became familiar with the locations mentioned in news coverage about the island (DiJulio, Muñana, and Brodie 2017). In the immediate aftermath of Hurricane Maria, many news headlines referred to “San Juan” without extra context such as “the capital of Puerto Rico”, largely because they expected their audience had already become familiar with the city due to the recent crisis. To better understand the nuanced dynamics of collective attention, we take a closer look at how people refer to locations of hurricanes during breaking crisis events via their usage of descriptor context phrases with respect to location mentions. Such descriptor phrases provide additional contextual information for named entities (people, organizations and locations) (Staliunaitė et al. 2018), helping to locate unfamiliar entities and disambiguate names that could have multiple referents. This is especially important when the writers assume their audiences have limited knowledge about the entities (Prince 1992).

In crisis events, we are particularly interested in the factors that influence descriptor phrase usage, which can be seen as a content-based reflection of collective attention over the course of the event. These factors include how a writer anticipates their audience’s understanding of the location being discussed, and whether a writer includes extra information outside of a descriptor phrase to help disambiguate the location. Studying how and when online discussions use or omit descriptor context when referring to locations can help crisis event participants more effectively track public awareness of an uncertain situation, better infer the public’s understanding of news events, and more strategically determine how to share information during such events.

Figure 1 shows an example of a shift in descriptor use during a crisis event. In public Twitter discussion of
Hurricane Maria in 2017, the location “San Juan” was less likely to receive a descriptor (e.g., “San Juan, PR”) following the peak in collective attention volume. While this shift appears to be due to time (Stalnäkte et al. 2018), the shift in descriptor use may also stem from non-temporal factors as well, such as an author’s expectations of their audience (audience design) and additional information such as links to external news articles, included in the same sentence with the location (a micro-level aspect of the discussion). Jointly modeling such macro-level factors, like post volume, and micro-level factors, from authors and information expectations, and their influence on a writer’s use of descriptor context can help reveal a more comprehensive picture of the dynamics in collective attention.

Concretely, this work examines the public discussion of five recent devastating natural disasters on Facebook and Twitter. We investigate how people refer to locations of hurricanes with or without descriptor phrases in their discussion and how such descriptor context use changes in response to factors related to audience, writer attributes and temporal trends. Our research questions are:

- **RQ1**: What factors influence people’s use of descriptor context when referring to locations of hurricane events?
- **RQ2a**: How does the use of descriptor context for locations change over time at a collective level?
- **RQ2b**: How does the use of descriptor context for locations change at an individual author level?

To address these research questions, we first analyzed posts written on Facebook in public groups concerning Hurricane Maria relief, and found that location mentions receive descriptors more often when the locations are not local to the group of discussion, suggesting that descriptors may be used to help explain new information to audiences. By looking at public posts written on Twitter concerning natural disasters, we found that the aggregate rate of descriptor phrases decreased following the peaks in these locations’ collective attention, supporting prior findings in the change in named entity use (Stalnäkte et al. 2018).

To assess potential individual-level causes of such content dynamics, we examined a set of characteristics related to audiences and authors, and we found that authors tend to use fewer descriptors if they had mentioned a location before, and to use more descriptors if the author received more audience engagement (e.g., more retweets and likes).

To sum up, our work demonstrates intuitive patterns in the use of descriptor phrases as a means of expressing shared knowledge expectations, which is an under-explored aspect of the content side of collective attention. Studying the use of descriptor phrases as well as other writing conventions in public discussions can provide insight into a writer’s expectations of their audience, and therefore a more fine-grained view into information sharing dynamics.

2 Related Work

The term collective attention refers to the attention that a public group of people pays to a particular event or topic (Sasahara et al. 2013), often as a result of a shared interest among the people. Collective attention is an important component in the spread of information (Wu and Huberman 2007), and it can shift either vary rapidly or gradually in response to particular events such as sports games (Lehmann et al. 2012), natural disasters (Varga et al. 2013), and political controversy (Garimella et al. 2017). With the wealth of digital data available to researchers today, studies have often quantified collective attention using the volume of posting and sharing activity in social media sites such as Reddit and Twitter (Leavitt and Clark 2014; Mitra, Wright, and Gilbert 2016). While these kinds of activity metrics provide an aggregate summary of attention dynamics, they largely obscure the nuanced content of collective attention such as how people refer to such particular events via language and how such referring language evolves over time. As an initial effort to understand this under-explored content aspect of collective attention, our research focuses on how people refer to named entities (e.g., locations, organizations) of breaking crisis events in their discussion, which are essential information for these events, and how such referring changes among large groups of people over the course of those crisis events.

When describing a named entity, a writer may add descriptive information in the form of a dependent clause (Kang et al. 2019) (e.g., “San Juan, in Puerto Rico”), to provide additional, contextual information for the audience to be familiar with the entity. The dependent clause may describe attributes of the entity that are relevant to a specific topic, such as “San Juan, epicenter of Hurricane Maria relief effort,” or attributes that are generally relevant, such as “San Juan, Puerto Rico.” From a collective perspective, prior work that examined the use of descriptor phrases in news media found that writers tend to drop such phrases as the entities gradually become more and more familiar (i.e., shared knowledge) among discussion participants over time (Stalnäkte et al. 2018). In addition to relative time, Siddharthan, Nenkova, and McKeown (2011) found that salience or the importance of the named entity, i.e., whether an entity plays a major role in the story or narrative being told, determines the need for a descriptor phrase, since a perceived salient or important entity is likely...
Table 1: Summary statistics for Twitter data.

| Event     | Hashtags                   | Date range     | Tweets | LOCATION NEs | LOCATION examples                  |
|-----------|----------------------------|----------------|--------|--------------|-----------------------------------|
| Florence  | #florence, #hurricaneflorence | [30-08-18, 26-09-18] | 66595  | 28670        | Wilmington, New Bern, Myrtle Beach |
| Harvey    | #harvey, #hurricaneharvey   | [17-08-17, 10-09-17] | 679400 | 181636       | Houston, Corpus Christi, Rockport  |
| Irma      | #irma, #hurricaneirma       | [29-08-17, 20-09-17] | 809423 | 229315       | Miami, Tampa, Naples               |
| Maria     | #maria, #hurricanemaria, #huracanamaria | [15-09-17, 09-10-17] | 313088 | 57237        | San Juan, Vieques, Ponce           |
| Michael   | #michael, #hurricanemichael | [06-10-18, 23-10-18] | 52506  | 22007        | Panama City, Mexico Beach, Tallahassee |

Table 2: Summary statistics for active authors on Twitter.

| Event | Authors | Tweets | LOCATION NEs |
|-------|---------|--------|--------------|
| Florence | 186     | 17624  | 29066        |
| Harvey  | 164     | 31563  | 50050        |
| Irma    | 178     | 45913  | 77114        |
| Maria   | 139     | 11332  | 18204        |
| Michael | 146     | 8828   | 14655        |

3 Data

Crisis events such as hurricanes present a useful case study for the development of collective attention, due to the large volume of online participation and large uncertainty among event observers towards the situation during the crisis events [Varga et al. 2013]. We chose to study the collective attention changes in public discourse related to hurricanes, due to hurricanes’ lasting economic impact, their broad coverage in the news, and their relevance to specific geographic regions. We collected social media data related to five recent devastating hurricanes, and we describe the data collection (§3.1), location detection (§3.2), and descriptor detection (§3.3) for the following datasets:

1. Twitter data: 2 million public tweets related to 5 major hurricanes, collected in 2017 and 2018.
2. Facebook data: around 30,000 posts from 60 public groups related to disaster relief in Hurricane Maria, collected in 2017.

3.1 Collection

Twitter Dataset The Twitter posts were collected using hashtags from five major disasters that recently struck the United States: Hurricane Florence (2018), Hurricane Harvey (2017), Hurricane Irma (2017), Hurricane Maria (2017), and Hurricane Michael (2018). We used hashtags that contained the name of the event in full and shortened form, e.g. #Harvey and #HurricaneHarvey for Hurricane Harvey. During 2017 and 2018, we streamed tweets that contained hashtags related to the natural disasters at the start of each disaster for up to one week after the dissolution of the hurricane. We augmented this data with additional tweets available in a 1% Twitter sample that contains the related hashtags, restricting our time frame to one day before the formation of the hurricane and one week after the dissipation of the hurricane. Manual inspection revealed minimal noise generated by the inclusion of the name-only hashtags. Summary statistics about the Twitter data are presented in Table 1. In addition to these tweets, we also collected additional event-related tweets from the most frequently-posting authors in each dataset (“active authors”), which were needed to evaluate per-author descriptor use change (see §4.3). Table 2 summarizes the detailed statistics about the active author data.

1According to NOAA estimates, e.g. Harvey’s estimates available here: [https://www.nhc.noaa.gov/data/tcr/AL092017_Harvey.pdf](https://www.nhc.noaa.gov/data/tcr/AL092017_Harvey.pdf)
3.2 Extracting and Filtering Locations

We extracted locations mentioned using, for English tweets, a distantly supervised named entity recognizer adapted to Twitter data (Ritter et al. 2011) and for Spanish tweets, a general purpose named entity recognizer (Manning et al. 2014). These NER systems are highly accessible, widely-used, and well-performing across multiple domains. We further evaluated the performance of these NER systems on a sample of tweets (100 tagged LOCATIONs per dataset, 500 total) and found reasonable precision for the LOCATION tag (81-96% across all datasets). For this work, we are interested in named entities that may require descriptors, which include cities and counties. We therefore restrict our analysis only to named entities (NEs) that (1) are descriptors, which include cities and counties. We therefore operationalized this as the occurrence of a well-known entity in a dependent clause relative to the location, which is straightforward to detect. Here, we used the population of a location as a proxy to determine how “well-known” that location is. The underlying assumption is that a more well-populated location may be more likely to be known or heard by more people and can therefore help describe the preceding location. In this work, we used the frequency of such descriptor phrases as the dependent variable: higher frequency of descriptor phrases uses indicates that the location may be new knowledge, while a lower frequency indicates that a location is more likely to be shared knowledge.

To extract sentence structure from text, we used dependency parsing, which decomposes a sentence into a directed acyclic graph connecting words and phrases. Following Staliūnaitė et al. (2018), we used a small set of dependencies to capture the “MODIFIER” phrase type in a subclause (adjectival clause, appositional modifier, prepositional modifier, numeric modifier) and another set of dependencies to capture the “COMPOUND” type in a super-clause (nominal modifier, compound, appositional modifier). A summary of our phrase patterns to capture descriptor phrases is provided in Table 3.

Facebook Dataset

The Facebook data was collected in the aftermath of Hurricane Maria by searching for public discussion groups that included at least one of Puerto Rico’s municipalities in the title (e.g., “Guayama: Huracán Maria” refers to Guayama municipality). Relatives and friends of Puerto Ricans often posted in these groups to seek additional information about those still on Puerto Rico, who could not be reached by telephone due to infrastructure damage. We restricted our analysis to Facebook groups related to Hurricane Maria because of the limited information causing more discussion of specific locations (as compared to the other hurricane events that had more up-to-date information available online). In total, we collected 31,414 public posts from 61 groups, from the time of their creation to one month afterward (Sept 20 to Oct 20 2017). Only posts in Spanish were retained (determined using langid.py and Baldwin 2012) because it was the majority language in the posts. Note that, due to Facebook data restrictions and API changes, we were not able to collect posts in Facebook groups for the other four hurricanes events, which we acknowledge as a limitation.

Table 3: Phrase patterns to capture descriptor phrases in location mentions. Head location marked with underline, context location marked with double underline.

| Phrase patterns | Dependency types | Example |
|-----------------|-------------------|---------|
| LOCATION + LOCATION_STATE | n/a | San Juan, PR |
| LOCATION + [LOCATIONCONTEXT] MODIFIER | adjective, apposition, preposition, numeric modifier | San Juan, [capital of Puerto Rico] |
| LOCATION + LOCATION_CONTEXT NOUN_COMPOUND | nominal, compound, apposition, conjunction | the [Vega Alta neighborhood of San Juan] |
| LOCATION + LOCATION_STATE CONJUNCTION |  | San Juan, Guayama [and Vieques, Puerto Rico] |

3.3 Extracting Descriptor Phrases

One way in which a writer mentions or helps introduce a new entity (e.g., “San Juan”) in their discussion is by linking it to a more well-known entity (e.g., “Puerto Rico”) in an descriptor phrase. We operationalized this as the occurrence of a well-known entity in a dependent clause relative to the location, which is straightforward to detect. Here, we used the population of a location as a proxy to determine how “well-known” that location is. The underlying assumption is that a more well-populated location may be more likely to be known or heard by more people and can therefore help describe the preceding location. In this work, we used the frequency of such descriptor phrases as the dependent variable: higher frequency of descriptor phrases uses indicates that the location may be new knowledge, while a lower frequency indicates that a location is more likely to be shared knowledge.

To extract sentence structure from text, we used dependency parsing, which decomposes a sentence into a directed acyclic graph connecting words and phrases. Following Staliūnaitė et al. (2018), we used a small set of dependencies to capture the “MODIFIER” phrase type in a subclause (adjectival clause, appositional modifier, prepositional modifier, numeric modifier) and another set of dependencies to capture the “COMPOUND” type in a super-clause (nominal modifier, compound, appositional modifier). A summary of our phrase patterns to capture descriptor phrases is provided in Table 3.

Taking into account the characteristics of text from two different domains, for the Twitter data we used the spacy shift-reduce parser (Honnibal and Johnson 2015) to extract the dependencies; for the Facebook data, the dependencies were extracted using the SyntaxNet transition-based parser (Andor et al. 2016) following initial tests that showed higher accuracy on SyntaxNet versus other comparable alternatives.

Accessed 1/2019: https://github.com/saffsd/langid.py
Accessed 10/2019: https://github.com/saffsd/langid.py
Accessed 1/2019: https://github.com/aritter/twitter_nlp
Accessed 9/2017: http://download.geonames.org/export/dump/allCountries.zip
Accessed 10/2018: https://nlp.stanford.edu/software/stanford-ner-2018-10-16.zip
Accessed 1/2019: https://cloud.google.com/natural-language/docs/analyzing-syntax
Accessed 1/2019: https://nlp.stanford.edu/software/stanford-language/docs/analyzing-syntax

---

4 Accessed 10/2017: https://github.com/aritter/twitter_nlp
5 Accessed 1/2019: https://github.com/aritter/twitter_nlp
6 Accessed 1/2019: https://spacy.io/usage
7 Accessed 1/2019: https://cloud.google.com/natural-language/docs/analyzing-syntax
We address our research questions in three analyses as follows.

4.1 What Affects the Use of Descriptor Context?

This section investigates RQ1 about what factors influence people’s use of descriptor context when referring to locations of hurricane events. We are particularly interested in the correlations between descriptor uses and a set of indicators of whether locations may be considered as old information. Here, a descriptor phrase may be omitted for locations that are geographically local to a group of people, i.e. knowledge that already shared among the group and are therefore assumed to be old information (e.g., if someone mentions the location “San Juan” in a group based in a region containing San Juan). To examine this research question, we compared the rate of descriptor uses for location mentions using both Facebook and Twitter data.

For the Facebook data, we determined whether the group’s region contains the location mentioned based on whether the most likely match for the location in the gazetteer is contained in that region. We then operationalized a set of explanatory variables mentioned above as follows: location mention frequency (importance), author in-group posting frequency (author status), and group size (audience), as summarized in Table 4. For the Twitter data, we operationalized a similar set of explanatory variables using the following: location mention frequency (importance), whether the author is an organization (author status), whether the author is a local (commitment), post length (information), URL presence (information), and image/video presence (information).

We built two logistic regression models to predict descriptor phrase use from the location containment variable with fixed effects on the categorical variables (location strings, authors, and groups) on the Facebook data and the Twitter data separately (N=18432 and N=49020, respectively). In detail, we used an elastic net regression (Zou and Hastie 2005) in order to reduce the risk of overfitting to fixed effects variables. For this analysis, rare categorical values (N < 20) for the fixed effects are replaced with RARE values to avoid overfitting to uncommon categories. The columns “RQ1 (Facebook)” and “RQ1 (Twitter)” in Table 5 report the results of our logistic regression models.

On Facebook, we observed that local locations are

| Factor      | Variable                          | Description                                      |
|-------------|-----------------------------------|--------------------------------------------------|
| Importance  | Prior location mentions           | Frequency of location within the group or event   |
| Author      | In-group posts                    | Posts that an author made within a group          |
|             | In-event posts                    | Posts that an author made about an event (log-transformed) |
|             | In-event posts about location     | Posts that an author made about an event that mention the location (log) |
|             | Organization                      | Whether the author is predicted to be an organization (based on metadata) |
|             | Local                             | Whether the author is predicted to be local to the event (based on self-reported location) |
| Audience    | Location is local to group        | Whether the location exists within the group’s associated region |
|             | Group size                        | Number of unique members who have posted in the group |
|             | Prior engagement                  | Mean normalized log-count of retweets and likes received by an author (in t-1) |
|             | Change in prior engagement        | Change in prior engagement received by an author (between t-2 and t-1) |
| Information | Has URL                           | Whether the post contains a URL                  |
|             | Has image/video                   | Whether the post contains a URL with an associated image/video |
| Time        | Time since start                  | Days since first post about event                |
|             | During peak                       | Whether post was written during peak of collective attention toward location |
|             | Post peak                         | Whether post was written at least 1 day after the peak of collective attention toward location |

Table 4: Summary of explanatory variables and corresponding metrics, used for descriptor phrase prediction.

Validation of Extraction Performance To assess the accuracy of our phrase patterns in capturing descriptor phrases, we asked two annotators (computer science graduate students) who had not seen the data to annotate a random sample of 50 tweets containing at least one location from each data set (250 tweets total). The annotators received instructions on how to determine if a location was marked by a descriptor phrase, including examples that were not drawn from the data, and the annotators marked each location mention as either (1) a “LOCATION + LOCATION_STATE” pattern, (2) one of the other descriptor patterns in Table 3 or (3) no descriptor phrase. The annotators achieved high agreement on each separate descriptor type (Cohen’s κ = 0.96 for the state pattern, κ = 0.91 for the other patterns). We then extracted posts with perfect agreement and detected descriptor phrases using the phrase patterns proposed. We found that our phrase patterns achieve reasonable precision and recall (96.6% and 87.5% respectively) in identifying descriptor phrases compared to raters’ annotations. This validation check demonstrated that our proposed syntactic patterns can capture descriptor phrases reasonably well.

4 Results

We assume that a location string matches a given location candidate if the candidate has the highest population in the gazetteer.

See Appendix A for details on determining whether an author is an organization or local.

L2 normalization weight of 0.01 chosen through grid search to maximize log-likelihood on held-out data (90-10 train/test split).
Table 5: Logistic regression results for all analysis, predicting the presence of a descriptor phrase. All regressions include fixed effects for location, (for Facebook) group, and (for Twitter) event. * indicates \( p < 0.05 \), otherwise \( p > 0.05 \) after multiple hypothesis correction. All models have significantly higher log-likelihood as compared to the null model. Note that the author population for RQ2b is restricted to active authors and therefore different from the author population in RQ1 (Twitter) and RQ2a.

We found that our operationalized factors of importance, author, audience and information correlate differently with writers’ use of descriptor phrases. Consistently, local locations or authors being a local are associated with lower rates of descriptor use, suggesting that the lack of descriptor context indicates shared knowledge among a large group of discussion participants.

4.2 Collective Change in Descriptor Context Use

This section investigates RQ2a on how the use of descriptor context for location mentions changes over time. Specifically, we used longitudinal data to examine the collective tendency to use more or fewer descriptor phrases over time. The intuition is that over the course of crisis events more collective attention to a particular location may result in more awareness of the location among discussion participants, therefore reducing the need for context.

In addition to the aforementioned explanatory factors used, we incorporated an additional set of variables to capture this temporal dynamics: relative peak time, i.e. whether the location is mentioned during or after the peak in post volume; and time since start, i.e. days since the beginning of the hurricane. Here, the definition of peak in collective attention is critical, because it determines the point at which an entity is expected to become shared knowledge (Staliūnaitė et al. 2018). Following Mitra, Wright, and Gilbert (2016), we defined the time of peak collective attention \( t_i \) for each location \( i \) as the (24-hour) period during which it is mentioned the most frequently:

\[
\hat{t}_i = \arg \max_{t \in T} f_i(t)
\]

where \( f_i(t) \) is the raw frequency of location \( i \) at time \( t \) (see Figure 1 for peak in “San Juan”

\[\text{peak in “San Juan”}\]
posts). We defined pre-peak as the period that ends \( t_{buffer} \) days before the frequency peak, during-peak as the period at most \( t_{buffer} \) days before and at most \( t_{buffer} \) days after, and post-peak as the period that begins \( t_{buffer} \) days after the frequency peak (we set \( t_{buffer} = 1 \)). To improve the stability of the fixed effects estimates, we removed all locations that are mentioned on fewer than \( N = 5 \) separate dates, and combined all authors with 1 post into a RARE bin.

As shown in the “RQ2a (Twitter)” column of Table 5, the main variable of interest, i.e. the post-peak time period, had less descriptor use than the earlier time periods (\( \beta = 0.127 \)). This suggests that a location may become shared knowledge after receiving a burst of collective attention and further validated our previous example study (see Figure 1). Furthermore, we found that descriptor phrase use decreased with the amount of time since the start of the event (\( \beta = 0.120 \)). This answers our RQ2a that the use of descriptor context for location mentions decreases over time, indicating that authors’ expectations of those locations being shared knowledge change gradually over the course of the event as well as in a burst following the attention peak.

One potential cause for the decrease in descriptor context may be the change in population after the peak in collective attention (e.g. influx of locals). To this end, we re-ran the regression above and replaced the author variables (“local” and “organization”) with a fixed effect for all authors. We found that the post-peak effect was still significant and negative (\( \beta = 0.253, p < 0.05 \)), which suggests that a change in author population is unlikely correlated with the decrease in descriptor use.

To summarize, we found consistently less descriptor use over the course of crisis events even after controlling for other explanatory factors, supporting prior work in long-term descriptor context change (Staliūnaitė et al. 2018).

4.3 Individual Change in Descriptor Context Use

The previous section showed that the collective descriptor use decreases after the peak in post volume even after controlling for other explanatory factors. This section further examines such changes at an individual author level (RQ2b), to determine whether authors modulate their use of descriptors over the course of the event in response to perceived changes in shared knowledge of those locations of hurricanes events. For example, does an author’s prior participation in discussion of the event lead to less descriptor use for the same event? Are authors who have a larger audience more likely to use descriptor phrases?

To better model the author-level changes in descriptor use, we introduced the following new factors into our regression models: number of prior posts by author during event (author-level), number of prior posts by author about the location during event (author-level), engagement received by author at \( t - 1 \) (audience), and change in engagement received by author between \( t - 2 \) and \( t - 1 \) (audience). These factors required a longitudinal sample of frequently-posting authors, i.e. active authors. Thus a set of active authors was identified in each data set, based on their relative post volume.

We scraped all publicly available tweets posted by these active authors that mention one of the event’s hashtags during the event time period (e.g., all posts for a Harvey-related active author from [17-08-17,10-09-17] that use #Harvey or #HurricaneHarvey). The locations and descriptor phrases were processed in the same way as before (see 3.2), and we report the relevant statistics for these active authors in Table 2. We built similar regularized regression models only using data from the active authors who posted at least once during each of the time periods, in order to isolate authors who may have changed over time.

The results are described in the “RQ2b (Twitter)” column of Table 5. We found several significant trends among those author and audience factors. (1) Authors’ prior mentions of a location are associated with less descriptor use (\( \beta = -0.237 \)). This negative correlation indicates that a location entity may be gradually understood as shared knowledge as the author repeats the location to the same audience (Galati and Brennan 2010). (2) More prior engagement from audiences correlates with more descriptor uses (\( \beta = 0.292 \)), which suggests that the authors need to plan their messages in response to cues from a larger and potentially more diverse audience. (3) Surprisingly, we found that this active author population showed no significant temporal tendency toward more or less descriptor use over the course of an event, during the peak or after the peak in collective attention. This null result held even when we performed the regression without the additional author and audience variables (i.e. same setup as RQ2a but including only the active author population).

We hypothesize that the active authors may be different from the overall population, i.e. active authors may have their different ways of responding to trends in collective attention and thus are less likely be influenced by such temporal trends, whereas less active authors are more likely to be influenced by them. We tested this by identifying less active authors (“regulars”) as those with lower post volumes and conducting the regression analysis with only those regulars. We found that these less active authors do show a significant decrease in descriptor use following the peak in collective attention (\( \beta = -0.127, p < 0.05 \)) and a decrease in descriptor use over time (i.e., with more time since start, \( \beta = -0.098, p < 0.05 \)). This suggests that less active authors may be more likely to accommodate to such collective temporal trends in descriptor context use, while the active authors are less responsive.

Addressing RQ2b, highly active authors do not change their descriptor context use over time, while relatively less active users show a decrease in such context use over the course of crisis events. However, the active authors do show significant modulation of their context use in response.
to more audience engagement and more mentions of the location in prior posts.

5 Discussion

By examining how people refer to crisis-impacted locations over the course of those crisis events, we found several trends in the use of contextualization related to audience expectations: (1) When authors are local to a place, or are writing for an audience who is expected to be local, they are less likely to use descriptor phrases to contextualize references to locations, reflecting shared knowledge among the author and audience. (2) At a collective level, there is a decreased descriptor use over the course of crisis events even after controlling for a set of explanatory variables. (3) At an individual level, highly active authors change their descriptor use in response to prior audience engagement but not after the peak in collective attention, whereas relatively less active users show a significant decrease in such context use over time.

5.1 Implications

This study demonstrates the benefits of studying the content of collective attention rather than merely quantity: studying how location entities are mentioned can provide more insight into writer intentions and expectations. For example, the initial example of “San Juan” losing descriptor use over time reveals a different narrative than its frequency alone would reveal, i.e. that the entity became shared knowledge in discussion during the hurricane. Furthermore, unlike other linguistic analysis techniques such as topic models, the method proposed for capturing descriptor phrases does not require extra interpretation and can be directly applied to large-scale social media data without the need for post-hoc interpretation, which could be beneficial for event monitors.

In addition to methods, this study provides insight into the role of audience among writers, which is relevant even during the extreme case of crisis events. The finding about local locations receiving fewer descriptors, along with the finding about active authors’ audience response, suggests that authors may accommodate to group norms in order to improve their odds of receiving a response. Furthermore, active authors’ descriptor use may correlate with audience engagement but not with overall collective attention peaks because these authors focus more on their own social behavior rather than responding to global trends.

Overall, the study highlights a set of practical and theoretical implications by looking at the content side of collective attention. First, we provide alternative ways to examine collective attention by looking at the content side of collective attention. For example, Mitra, Wright, and Gilbert (2016) found that more sustained attention (lower variance in post volume) toward an event on social media correlates with lower perceived credibility of the event. Such analyses can be further enriched via our content-based operationalization of collective attention: for instance, future work might analyze how certain shared knowledge towards crisis events inferred from location mentions and descriptor context uses correlates with the perceived credibility of those events. Our work could also help distinguish nuances in hashtag uses of collective attention (Lehmann et al. 2012) to better understand different forms of the manifestation of collective attention. Furthermore, our work sheds light on how and when online discussions use or omit descriptor context when referring to locations during crisis events. This can help crisis event participants more effectively track public awareness of an uncertain situation, better infer the public’s understanding of news events, and more strategically determine how to share information during such events. It also has theoretical implications for understanding linguistic structures (e.g. descriptor phrases) and social change.

5.2 Limitations

Our work is also subject to several limitations. First, the analysis of what factors influence the use of descriptive context are mainly correlational without casual validations. Our formulation of descriptor phrases is not exhaustive and may have missed other syntactic construction that indicate that an entity is considered new information (i.e. false negatives). A speaker may use a preceding descriptor phrase, instead of a subordinate descriptor phrase, to indicate that the entity is not shared knowledge (e.g. “a city called San Juan”). In addition, we focused on only a set of specific crisis events due to their representative usages of location mentions and large volume of online discussions. Future work can build upon our work and generalize it to other different types of crisis events. In addition, we are unable to rule out the possibility that another event attracted attention to the locations under discussion before the crises began (e.g. a political news story relevant to the event’s region). Lastly, the study focuses exclusively on location names because of their geographic relevance to events, but other types of named entities (people, organizations) are also likely to undergo changes in descriptor use in response to increased attention (Stalinaitė et al. 2018).

To conclude, this study adds a new content-based perspective to the measurement of collective attention, by analyzing how people discuss breaking news events. By examining five recent hurricane events, our research demonstrated how referring expressions are shaped by author and audience expectations of collective attention over time and across communities.

Acknowledgments

The authors thank members of Georgia Tech’s SocWeb group for their valuable feedback. This research was supported by NSF award IIS-1452443 and NIH award R01-GM112697-03.

References

[Andor et al. 2016] Andor, D.; Alberti, C.; Weiss, D.; Severyn, A.; Presta, A.; Ganchev, K.; Petrov, S.; and Collins, M. 2016. Globally normalized transition-based neural networks. In ACL, 2442–2452.
committed) to the event’s region will already be aware of the locations under discussion (Kogan, Palen, and Anderson 2015) and will be less likely to use context than an author who is unfamiliar with the region’s locations. Next, organizations such as government agencies are often responsible for disseminating official information to help crisis responders and effectively organize aid (Houston et al. 2015). An author who represents an official organization may want to minimize uncertainty in their messages and use more context than an author who does not represent an organization, i.e. citizen observer.

We determine author local status and organization status using a sample of metadata available from archived tweets corresponding to the time periods of interest (covering ∼20% of all authors in the data). Following prior work in geolocation (e.g. Kariryaa et al. 2018), we approximate the local status of an author posting about an event based on whether the author’s self-reported profile location mentions a relevant city or state in the event’s affected region (e.g. for Hurricane Maria, a local author would mention “Puerto Rico” or “PR” in their location field).

Organizations are difficult to identify automatically, because there is no single indicator of organization status in a Twitter user’s profile information. To determine organization status, we apply a pretrained classifier (Wood-Doughty, Mahajan, and Dredze 2018) to the author’s metadata, including name, description, and social attributes.

For both local and organization status, we find reasonable precision with respect to a small subset of hand-labeled authors from our data.\footnote{17 One of the authors annotated 500 accounts as organizations and locals, based on available metadata, and compared these labels to those produced by the local proxy and organization classifier. The local proxy achieved precision of 87% and recall of 58%, and the organization classifier achieved precision of 87% and recall of 54%.

\footnote{16 Accessed 7/2019: https://bitbucket.org/mdredze/demographer/src/peoples2018/}