A Syntax-based Learning Approach to Geo-locating Abnormal Traffic Events using Social Sensing

Yang Zhang, Xiangyu Dong, Daniel Zhang, Dong Wang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA
{yzhang42, xdong2, yzhang40, dwang5}@nd.edu

***Abstract*—Social sensing has emerged as a new sensing paradigm to observe the physical world by exploring the "wisdom of crowd" on social media. This paper focuses on the *abnormal traffic event localization* problem using social media sensing. Two critical challenges exist in the state-of-the-arts: i) "content-only inference": the limited and unstructured content of a social media post provides little clue to accurately infer the locations of the reported traffic events; ii) "informal and scarce data": the language of the social media post (e.g., tweet) is informal and the number of the posts that report the abnormal traffic events is often quite small. To address the above challenges, we develop SyntaxLoc, a syntax-based probabilistic learning framework to accurately identify the location entities by exploring the syntax of social media content. We perform extensive experiments to evaluate the SyntaxLoc framework through real world case studies in both New York City and Los Angeles. Evaluation results demonstrate significant performance gains of the SyntaxLoc framework over state-of-the-art baselines in terms of accurately identifying the location entities that can be directly used to locate the abnormal traffic events.**

***Index Terms*—Syntax-based Learning, Abnormal Detection, Localization, Social Sensing**

I. Introduction

Social sensing has emerged as a new sensing paradigm to collect the state of the physical world by exploring the "wisdom of crowd" on social media [1]. Such a new paradigm is driven by the proliferation of portable devices for individuals (e.g., smartphone), ubiquitous wireless communication technology (e.g., 4/5G), and mass information dissemination media (e.g., Twitter, Facebook) [2]. An important application domain for social media sensing is to acquire real-time traffic situation awareness (e.g., congestion, traffic accidents) by exploring the rich information on online social media [3]. A key problem in such applications is to identify accurate locations of the abnormal traffic events so effective precautions and timely responses can be provided to improve the traffic safety and efficiency [4]. We refer to this problem as *abnormal traffic event localization*.

A simple way of addressing the above problem is to identify the geotags of a social media post (e.g., "coordinates" field of a tweet [1]. However, this simple solution suffers from the location data sparsity problem on social media: only a very limited number of social media posts are associated with geotags mainly due to privacy concerns (e.g., fewer than 0.5% tweets has geo-tags [5]). The location estimation problem has been well-studied by previous works in social sensing [5]–[12]. However, those approaches cannot be directly adapted to solve the abnormal traffic event localization problem due to two critical challenges: *content-only inference* and *informal and scarce data*.

Content-only Inference: A possible approach to address the aforementioned location data sparsity problem is to infer the event location by analyzing the content of social media posts [13]. While this method eliminates its dependency on the availability of sparse geotags, the limited and unstructured content in a social media post (e.g., 280 characters in a tweet) makes the location inference problem challenging [14]. A few content based location inference methods can be applied to solve this problem [6], [11]. However, two important limitations exist: i) the current content-only location inference methods are not accurate enough to geo-locate abnormal traffic events (e.g., the average inference error is about 20 miles) [6]; ii) some existing solutions require prior or external knowledge (e.g., private user social network relations and activities) for location inference [11], which is not always available due to the privacy or legal concerns.

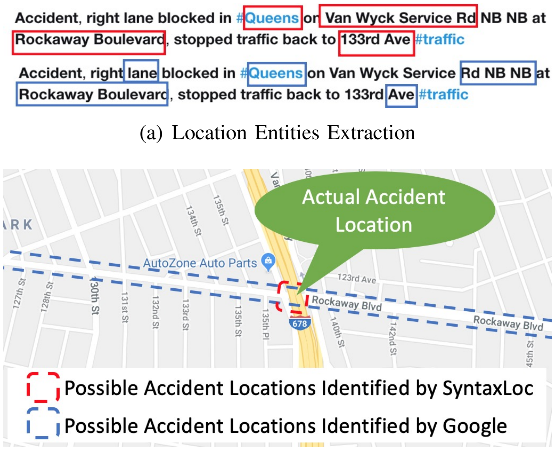
Informal and Scarce Data: Unlike formal documents (e.g., news reports) with well-established semantic structure and rich content, social media posts are known to be informal (i.e., short and incomplete sentences with special symbols (e.g.,#) inserted [15], [16]. In addition, the average number of social media posts that report an abnormal traffic event is often quite small. For example, our case study in New York City shows that most traffic incidents are only reported by a single tweet. In general, it is a challenging task to infer locations of the abnormal traffic events from such informal and scarce social

ASONAM '19, August 27-30, 2019, Vancouver, Canada
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6868-1/19/08...$15.00
https://doi.org/10.1145/3341161.3343708

[1] https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location.html

media sensing data [17].

To address the above challenges, we develop SyntaxLoc, a syntax-based probabilistic learning framework to accurately identify the location entities [2] by exploring the syntax of social media content. The identified location entities then can be used to accurately geo-locate the abnormal traffic events reported on social media. For example, a traffic accident reported on Twitter is shown in Figure 1(a). The red boxes on the first instance indicate all location entities in the tweet that can be used to accurately geo-locate the traffic accident. The blue boxes on the second instance indicate the location entities identified by *Google Named Entity Detection service* [3]. We can observe that Google's tool fails to identify a few location entities that are essential to accurately geo-locate the traffic accident reported in the tweet. As a result, Google's tool can only infer the traffic accident at the street level while our scheme can accurately locate the accident at the intersection where the accident actually happened as shown in Figure 1(b).



Accident, right lane blocked in #Queens on Van Wyck Service Rd NB NB at Rockaway Boulevard, stopped traffic back to 133rd Ave #traffic

Accident, right lane blocked in #Queens on Van Wyck Service Rd NB NB at Rockaway Boulevard, stopped traffic back to 133rd Ave #traffic

(a) Location Entities Extraction

(b) Accident Location Inference

Figure 1. Example of Abnormal Traffic Event Localization

To the best of our knowledge, SyntaxLoc is the first syntax-based learning approach to address the abnormal traffic event localization problem by purely analyzing the social media content. In particular, to address the *content-only inference* challenge, we propose a syntax-based pattern learning module in SyntaxLoc to explicitly learn the syntax-based patterns and represent the learned patterns using novel probability representations. To address the *informal and scarce data* challenge, we develop a probabilistic-based entity extraction module to accurately identify the location entities through a principle probabilistic learning framework. We perform extensive experiments to evaluate the SyntaxLoc framework through two real-world Twitter datasets collected from New York City and Los Angeles. The evaluation results demonstrate significant performance gains of our SyntaxLoc framework over the state-of-the-art baselines (e.g., *Google Named Entity Detection* and *Stanford CoreNLP* platforms) in terms of accurately identifying the location entities from social sensing data.

The main contributions of this paper are summarized as follows:

- We develop SyntaxLoc to address the abnormal traffic event localization problem (i.e., accurately identifying the location entities) by exploring the syntax of social media content.
- We address two important challenges (i.e., content-only inference, informal and scarce data) using a principled syntax-based probabilistic learning framework.
- We perform extensive experiments to evaluate our solution through two real world case studies. The results demonstrate significant performance gains of our scheme compared to state-of-the-art baselines.

II. Related Work

A. *Social Sensing*

Social sensing has emerged as a new sensing paradigm by using humans as sensors [18], [19]. Social sensing has been widely applied in various application domains [20]–[25] including damage assessment in disaster response [22], multi-modal data fusion [23], crowd video sharing [24], and environment and urban infrastructure monitoring [25]. Abnormal traffic event localization using social sensing remains to be an important challenge that has not been well-addressed in intelligent transportation systems. The goal is to identify accurate locations of the abnormal traffic events so effective precautions and timely response can be provided to improve the traffic safety and efficiency. In this paper, we develop a SyntaxLoc framework to address this problem by accurately identifying the location entities from social media posts that can be directly used to locate the abnormal traffic events.

B. *Location Inference in Social Media Sensing*

The location estimation problem has been well-studied by previous works in social sensing [5]–[12]. For example, Chen *et al.* proposed a probabilistic learning framework to geo-locate Twitter users by analyzing the content of collected tweets [5]. Schulz *et al.* developed a multi-indicator approach to determine the location of a tweet by examining the user's profiles [6]. Li *et al.* developed a unified discriminative influence model to infer users' home locations based on their social network activities [7]. Kinsella *et al.* proposed a location inference scheme to estimate the location of each individual tweet at a city-level using a probabilistic based language model trained on geotagged tweets [8]. However, those approaches cannot be directly adapted to solve the abnormal traffic event localization problem because they either require prior or external knowledge (e.g., complete gazetteer database, private user activities) or are unable to provide fine-grained location inference results (e.g., average location error of 20-100 miles). Different from the previous solutions, we develop a novel SyntaxLoc framework that purely relies on the syntax of post content to provide accurate location inference

[2] We refer to *location entities* as the named-entities in a social media post that indicate the location of the abnormal traffic event.

[3] https://cloud.google.com/natural-language/

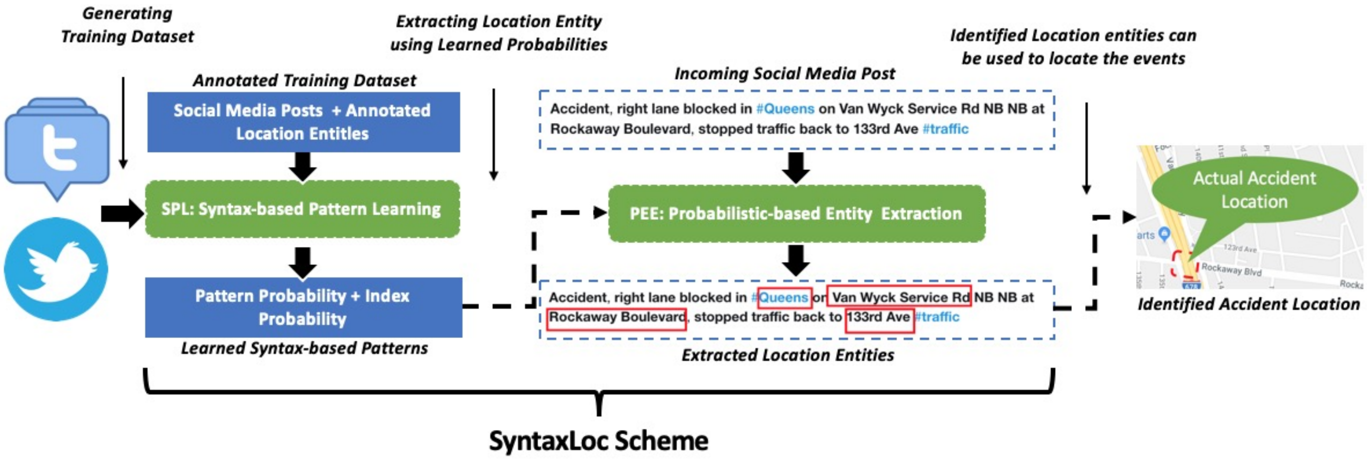

Generating Training Dataset
Annotated Training Dataset
Social Media Posts + Annotated Location Entities
SPL: Syntax-based Pattern Learning
Pattern Probability + Index Probability
Learned Syntax-based Patterns
Extracting Location Entity using Learned Probabilities
Incoming Social Media Post
Accident, right lane blocked in #Queens on Van Wyck Service Rd NB NB at Rockaway Boulevard, stopped traffic back to 133rd Ave #traffic
PEE: Probabilistic-based Entity Extraction
Accident, right lane blocked in #Queens on Van Wyck Service Rd NB NB at Rockaway Boulevard, stopped traffic back to 133rd Ave #traffic
Extracted Location Entities
Identified Location entities can be used to locate the events
Actual Accident Location
Identified Accident Location
SyntaxLoc Scheme

Figure 2. Overview of SyntaxLoc Framework

results for abnormal traffic events reported on online social media.

C. Probabilistic Learning Technique

Our SyntaxLoc framework is also closely connected to the probabilistic learning technique in machine learning. Such a learning technique has been widely used in application domains like computer version, natural language processing, and information retrieval [26]–[28]. For example, Li *et al.* proposed an image annotation framework for image-word correlation estimation using multi-correlation probabilistic matrix factorization [26]. Zettlemoyer *et al.* developed a structured classification algorithm to map sentences to logical forms using probabilistic categorical grammars [27]. Huang *et al.* proposed a transductive learning framework to improve the image retrieval accuracy using the probabilistic hypergraph ranking [28]. To the best of our knowledge, SyntaxLoc is the first syntax-based probabilistic learning approach to address the abnormal traffic event localization problem by exploring the syntax of social media content in social sensing applications.

III. Preliminary

In this section, we formally introduce the problem of abnormal traffic event localization using social sensing in intelligent transportation systems.

Definition 1: **Social Media Posts** (S): We define (S) to be the social media posts about abnormal traffic events reported by social media users (e.g., a tweet). In particular, we define $S = \{S^1, S^2, ..., S^B\}$, where S^b represents a collected social media post and b is its index.

Definition 2: **Location Entity** (L): We define the Location Entity (L) to be a set of named entities in a social media post that indicate the location of the abnormal traffic event (see Figure 1(a))). In particular, we define $L^b = \{L^b_1, L^b_2, ..., L^b_A\}$ to be the set of A location entities from the post S^b.

The goal of our problem is to accurately identify all location entities of an abnormal traffic event from the social media post. Using the definitions above, we formally define our problem as follows:

$$\underset{\widehat{L^b}}{\arg\max} \Pr(\widehat{L^b} = L^b \mid S^b), \forall 1 \leq b \leq B \tag{1}$$

where $\widehat{L^b}$ is the set of *estimated* location entities in social media post S^b. We present the key modules of our SyntaxLoc framework in the following section.

IV. SyntaxLoc Framework

A. Overview of SyntaxLoc Framework

The overall architecture of the SyntaxLoc is shown in Figure 2. In general, SyntaxLoc consists of two major modules:

- *Syntax-based Pattern Learning (SPL) Module:* it explicitly learns the syntax-based patterns and represents the learned patterns using two novel probability representations (i.e., *pattern probability* and *index probability*) given a training set of social media posts. The learned probability representations then can be used to effectively identify location entities for the unlabeled social media posts.
- *Probabilistic-based Entity Extraction (PEE) Module:* it is designed to accurately identify the location entities of an incoming social media post by leveraging the learned syntax-based patterns captured in the SPL module through a principled probabilistic learning framework. The extracted location entities can then be used to locate the abnormal traffic events.

B. Syntax-based Pattern Learning (SPL)

We first present a *syntax-based pattern learning (SPL)* module to capture the syntax-based patterns, which can be used to effectively identify location entities from the social media posts. We first define a few key terms.

Definition 3: **Entity** (e): we define a continuous set of words that belong to the same *part-of-speech* as a single entity e (e.g.,

Table I
EXAMPLE OF SYNTAX MODEL OF A TWEET

Social Media Post S	"Accident on Chestnut St approaching Main St"
Entity e (Syntax)	"Accident (*NOUN*)", "on (*ADP*)", "Chestnut St (*NOUN*)","approaching (*VERB*)", "Main St (*NOUN*)"
2-Syntax Model $M^{(2)}$	$p_1^{(2)}$: *NOUN_ADP* ["Accident (*NOUN*)","on (*ADP*)"] $p_2^{(2)}$: *ADP_NOUN* ["on (*ADP*)", "Chestnut St (*NOUN*)"] $p_3^{(2)}$: *NOUN_VERB* ["Chestnut St (*NOUN*)", "approaching (*VERB*)"] $p_4^{(2)}$: *VERB_NOUN* ["approaching (*VERB*)", "Main St (*NOUN*)"]
3-Syntax Model $M^{(3)}$	$p_1^{(3)}$: *NOUN_ADP_NOUN*: ["Accident (*NOUN*)", "on (*ADP*)", "Chestnut St (*NOUN*)"] $p_2^{(3)}$: *ADP_NOUN_VERB* ["on (*ADP*)", "Chestnut St (*NOUN*)", "approaching (*VERB*)"] $p_3^{(3)}$: *NOUN_VERB_NOUN* ["Chestnut St (*NOUN*)", "approaching (*VERB*)", "Main st (*NOUN*)"]
4-Syntax Model $M^{(4)}$	$p_1^{(4)}$: *NOUN_ADP_NOUN_VERB* ["Accident (*NOUN*)", "on (*ADP*)", "Chestnut St (*NOUN*)", "approaching (*VERB*)"] $p_2^{(4)}$: *ADP_NOUN_VERB_NOUN* ["on (*ADP*)", "Chestnut St (*NOUN*)", "approaching (*VERB*)", "Main St (*NOUN*)"]

Chestnut St). This can be done by applying existing language tools [4] to tag the *part-of-speech* for a given word in the post.

Definition 4: **n-Syntax Pattern** (**$p^{(n)}$**): we define an *n-Syntax Pattern* to be a contiguous syntax sequence of n entities from a given social media post. For instance, "*NOUN_ADP_NOUN*" is *3-syntax pattern* for the 3 entities "*Accident_on_Chestnut St*".

Definition 5: **n-Syntax Model** (**$M^{(n)}$**): we consider the collection of all possible n-syntax patterns $p^{(n)}$ as *n-Syntax Model* $M^{(n)}$.

Table I shows a simplified example of a tweet and its relevant entities, syntax patterns, and syntax models we define above.

We further define two sets of probabilities, namely *pattern probability* and *index probability*, to identify location entities from the social media post.

Pattern Probability: we first define the probability of an n-syntax pattern $p^{(n)}$ in an n-syntax model $M^{(n)}$ as follows:

$$\Pr(p^{(n)}|M^{(n)}) = \frac{|p^{(n)}|}{|M^{(n)}|} \tag{2}$$

where $|p^{(n)}|$ indicates the number of occurrences of the n-syntax pattern $p^{(n)}$ in a given set of social media posts. $|M^{(n)}|$ indicates the total number of all n-syntax patterns.

Index Probability: We define the probability of a location entity index i^n in an n-syntax pattern $p^{(n)}$ as follows:

$$\Pr(i^{(n)}|p^{(n)}) = \frac{|i^{(n)}|}{|p^{(n)}|} \tag{3}$$

where $|i^{(n)}|$ indicates the number of the location entities in the i^{th} entity given the n-syntax pattern $p^{(n)}$.

The pseudocode of the *syntax-based pattern learning (SPL)* module is shown in Algorithm 1. The input to the module is a training set of social media posts with annotated location entities for each social media post. The outputs are the learned *pattern probability* $\Pr(p^{(n)}|M^{(n)})$ and *index probability* $\Pr(i^{(n)}|p^{(n)})$ that can be used to effectively identify location entities for the unlabeled social media posts.

Algorithm 1 Syntax-based Pattern Learning (SPL)

```
collect a training set of $S$ as $\bar{S}$
annotate $\bar{L}^b$ for each $\bar{S}^b$ in $\bar{S}$
compute $|M^{(n)}|$ in $\bar{S}$
compute $|p^{(n)}|$ for each $M^{(n)}$ in $\bar{S}$
generate $\Pr(p^{(n)}|M^{(n)})$ using Equation 2
compute $|i^{(n)}|$ for each $p^{(n)}$ in $\bar{S}$
generate $\Pr(i^{(n)}|p^{(n)})$ using Equation 3
output $\Pr(p^{(n)}|M^{(n)})$ and $\Pr(i^{(n)}|p^{(n)})$ as the inputs for PEE module
```

C. Probabilistic-based Entity Extraction (PEE)

The PEE module is designed to accurately identify the location entities from the social media posts by leveraging the learned *pattern probability* and *index probability* from SPL.

In particular, we can derive the likelihood of entity e to be a location entity L as follows:

$$\begin{aligned}\Pr(e \in L|i^{(n)}, p^{(n)}, M^{(n)}) &= \Pr(i^{(n)}|p^{(n)}) \\ &\times \Pr(p^{(n)}|M^{(n)}) \times \Pr(M^{(n)})\end{aligned} \tag{4}$$

where $\Pr(i^{(n)}|p^{(n)})$ is the index probability and $\Pr(p^{(n)}|M^{(n)})$ is the pattern probability. $\Pr(M^{(n)})$ is the weight of n-syntax model that indicates the importance of each n-syntax model in identifying the location entities. It is usually set to be a small value if no prior knowledge is given.

We observe that an entity e can occur in multiple n-syntax pattern $p^{(n)}$ with different index $i^{(n)}$ on different n-syntax model $M^{(n)}$ (e.g., "Chestnut St (*NOUN*)" occurs in different 2-,3-,4-syntax patterns shown in Table I). Therefore, we combine the derived likelihood over different syntax patterns as follows:

$$\Pr(e \in L) = \sum_{M^{(n)}} \sum_{(i^{(n)}, p^{(n)})} \Pr(e \in L|i^{(n)}, p^{(n)}, M^{(n)}) \tag{5}$$

Finally, if the value of $\Pr(e \in L)$ exceeds a predefined threshold Δ [5], we classify the entity e to be a location entity

[4] https://cloud.google.com/natural-language/docs/analyzing-syntax

[5] Δ is an application specific parameter. In the evaluation section, we show the robustness of our scheme in terms of various Δ values.

as follows:

$$\begin{cases} \mathbf{1} : \{\Pr(e \in L) > \Delta\} \\ \mathbf{0} : \{\Pr(e \in L) \leq \Delta\} \end{cases} \quad (6)$$

where “1” (i.e., true) indicates entity e is a location entity and “0” (i.e., false) otherwise. The identified location entities can then be used to accurately geo-locate the abnormal traffic events reported on social media.

The pseudocode of the *probabilistic-based entity extraction (PEE)* is shown in Algorithm 2. The inputs to the module are the *pattern probability* $\Pr(p^{(n)}|M^{(n)})$ and *index probability* $\Pr(i^{(n)}|p^{(n)})$ learned in SPL module. For an incoming social post S^b, the output is the set of location entities L^b in S^b.

Algorithm 2 Probabilistic-based Entity Extraction (PEE)

```
obtain $\Pr(p^{(n)}|M^{(n)})$ and $\Pr(i^{(n)}|p^{(n)})$ using SPL module
for each $e$ in $S^b$ do
    for $e$ in each $(i^{(n)}, p^{(n)})$ and $M^{(n)}$ do
        compute $\Pr(e \in L|i^{(n)}, p^{(n)}, M^{(n)})$ using Equation 4
    end for
    compute $\Pr(e \in L)$ using Equation 5
end for
if $\Pr(e \in L) > \Delta$ then
    $e$ is in $L^b$
else
    $e$ is not in $L^b$
end if
```

D. Summary of SyntaxLoc Framework

Finally, we summarize SyntaxLoc using the pseudocode in Algorithm 3. The input of our SyntaxLoc framework is the set of social media posts S about abnormal traffic events reported by social media users. The output is the set of location entities L^b in each social media post S^b that can be used to locate the abnormal traffic event reported in S^b.

Algorithm 3 Summary of the SyntaxLoc Framework

```
generate $\Pr(p^{(n)}|M^{(n)})$ and $Pr(i^{(n)}|p^{(n)})$ using SPL module
for each $S^b$ in $S$ do
    for each $e$ in $S^b$ do
        examine $e$ using PEE module
        if $e$ is a location entity then
            add $e$ to $L^b$
        end if
    end for
    output $L^b$ for $S^b$
end for
```

V. Evaluation

In this section, we perform extensive experiments to evaluate the proposed SyntaxLoc framework using two real-world Twitter datasets collected from New York City and Los Angeles. In addition, we compare the performance of SyntaxLoc with state-of-the-art baselines including *Google Named Entity Detection* and *Stanford CoreNLP* platforms. The evaluation results demonstrate significant performance gains of our scheme over the state-of-the-art baselines in terms of accurately identifying the location entities from social sensing data.

Table II
Evaluation Metrics Definitions

Metrics	Definition
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$\frac{2\times Precision\times Recall}{Precision+Recall}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$

A. Dataset

We collected two datasets from Twitter using *Get Old Tweets* [6] on abnormal traffic events (e.g., accidents, congestion) from New York City and Los Angeles, respectively. In particular, we randomly sample 200 tweets [7] in each dataset for our evaluation. We manually annotate the location entities in each tweet to collect the ground-truth labels. In particular, the New York City dataset contains 2,412 entities and 20.4% of them are location entities. The Los Angeles dataset contains 2,851 entities and 16.7% of them are location entities.

B. Baselines

We compare the performance of the SyntaxLoc framework with several state-of-the-art baselines in identifying the location entities from social media data as follows:

- **Google Named Entity Detection** [8] **(Google):** The state-of-the art commercial entity recognition platform that identifies entities and associated types (e.g., location) through powerful pre-trained machine learning models.
- **Stanford CoreNLP** [9] **(Stanford):** An integrated NLP toolkit that can identify the location entities from various types of documents [10].
- **Spacy** [10]**:** An industrial-strength natural language tool that provides the named entity recognition service through well-trained entity detection models.

C. Metrics

In the evaluation, we use the standard metrics for binary classification (i.e., location entity v.s. non-location entity): *Precision*, *Recall*, *F1-score*, *Accuracy* to evaluate the performance of all compared schemes. Their definition are given in Table II (*TP*: True Positives, *FP*: False Positives, *TN*: True Negatives, *FN*: False Negatives). In particular, TP and TN

[6] https://github.com/Jefferson-Henrique/GetOldTweets-python
[7] The number of tweets is mainly limited by the human power of annotation.
[8] https://cloud.google.com/natural-language/
[9] https://stanfordnlp.github.io/CoreNLP/
[10] https://spacy.io/

Table III
EVALUATION RESULTS (NEW YORK CITY)

	Precision	Recall	F1-Score	Accuracy
SyntaxLoc$_{0.4}$	**0.6147**	**0.8152**	**0.7009**	**0.8568**
SyntaxLoc$_{0.5}$	**0.6545**	**0.7826**	**0.7128**	**0.8702**
SyntaxLoc$_{0.6}$	**0.6881**	**0.6956**	**0.6918**	**0.8724**
Google	0.4659	0.4456	0.4555	0.7807
Stanford	0.1333	0.0652	0.0875	0.7203
Spacy	0.4838	0.3260	0.3896	0.7897

Table IV
EVALUATION RESULTS (LOS ANGELES)

	Precision	Recall	F1-Score	Accuracy
SyntaxLoc$_{0.4}$	**0.5447**	**0.7790**	**0.6411**	**0.8648**
SyntaxLoc$_{0.5}$	**0.6122**	**0.6976**	**0.6521**	**0.8846**
SyntaxLoc$_{0.6}$	**0.6444**	**0.6744**	**0.6590**	**0.8918**
Google	0.4680	0.5116	0.4889	0.8342
Stanford	0.1190	0.0581	0.0781	0.7873
Spacy	0.3571	0.2325	0.2816	0.8162

indicate the number of *ground-truth* location entities and non-location entities that are correctly identified by corresponding schemes, respectively. FN and FP indicate the number of *ground-truth* location entities and non-location entities that are falsely identified by the corresponding schemes, respectively. Please note that we do not evaluate the accuracy in terms of estimating the actual geo-locations (i.e., coordinates) of the reported abnormal traffic events due to the difficulty of obtaining the ground-truth locations of all such events reported on social media [29].

D. Evaluation Results

1) Identification Performance: In our experiments, we randomly split each dataset (i.e., New York City dataset, Los Angeles dataset) into 80% training subset and 20% testing subset. We also vary the threshold Δ (defined in Equation 6) from 0.4 to 0.6 for the SyntaxLoc framework (e.g., SyntaxLoc$_{0.6}$ represents SyntaxLoc framework with $\Delta = 0.6$). The results are shown in Table III and Table IV. We observe that our SyntaxLoc framework outperforms all baselines in all evaluation metrics for both datasets. The performance gains achieved by SyntaxLoc compared to the best-performing baseline in New York City dataset on precision, recall, F1-score, and accuracy are 22.2%, 36.9%, 25.7%, and 8.2%, respectively. Similar performance gains are also observed in the Los Angeles dataset. Such performance gains of SyntaxLoc are achieved by judiciously learning the syntax patterns and accurately capturing the location entities from social media posts under a principled probabilistic learning framework. We also observe that our SyntaxLoc framework consistently outperforms all baselines as the Δ value changes in both New York City and Los Angeles datasets. Such consistent performance gains demonstrate the robustness of our scheme with respect to the Δ parameter of our model.

We also evaluate our SyntaxLoc framework by changing the ratio of training subset from 60% to 80% for both New York City and Los Angeles datasets. The performance results of SyntaxLoc are presented in Figure 3 and Figure 4. We observe a stable performance of SyntaxLoc framework over different sizes of training subset.

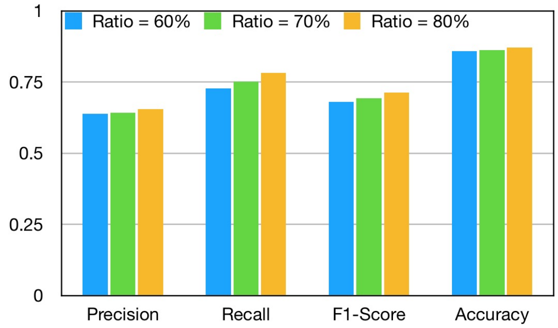


Figure 3. Performance of SyntaxLoc with Different Training Ratio (New York City)

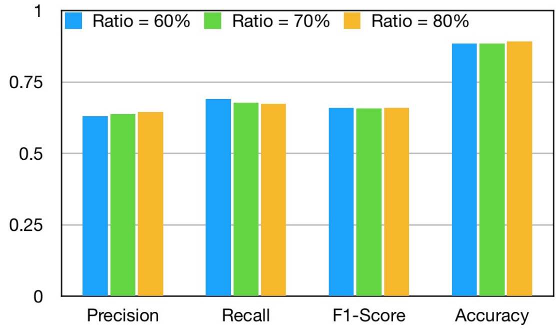


Figure 4. Performance of SyntaxLoc with Different Training Ratio (Los Angeles)

2) Most Frequent n-Syntax Patterns: Finally, we show the top 10 most frequent n-syntax patterns (defined in Definition 4) and associated location entity indexes (defined in Equation 3) learned by our SyntaxLoc framework in both New York City and Los Angeles datasets. The results are shown in Table V and Table VI. We observe that our scheme is able to capture the most important syntax patterns that can be used to effectively identify the location entities from social media posts. For example, the top two most frequent 2-syntax patterns learned by our SyntaxLoc framework in New York City (i.e.,"*ADP_NOUN*" with location entity index of 2, "*NOUN_ADP*" with location entity index of 1) match well with the example shown in Figure 1(a). In particular, "*ADP_NOUN*" matches "at_Rockaway Boulevard" and "*NOUN_ADP*" matches "Queens_on" in the example. In addition, we also observe that the two datasets share several similar

Table V
MOST FREQUENT N-SYNTAX PATTERNS AND LOCATION ENTITY INDEXES (NEW YORK CITY)

Rank	2-Syntax Pattern	Index	3-Syntax Pattern	Index	4-Syntax Pattern	Index
1	*ADP_NOUN*	2	*NOUN_ADP_NOUN*	3	*ADP_NOUN_ADP_NOUN*	4
2	*NOUN_ADP*	1	*NOUN_ADP_NOUN*	1	*ADP_NOUN_ADP_NOUN*	2
3	*NOUN_VERB*	1	*ADP_NOUN_ADP*	2	*NOUN_ADP_NOUN_ADP*	3
4	*NOUN_NOUN*	1	*ADP_NOUN_CONJ*	2	*ADP_NOUN_CONJ_NOUN*	2
5	*NOUN_CONJ*	1	*NOUN_CONJ_NOUN*	1	*NOUN_ADP_NOUN_CONJ*	3
6	*CONJ_NOUN*	2	*ADP_NOUN_VERB*	2	*ADP_NOUN_CONJ_NOUN*	4
7	*NOUN_NOUN*	2	*NOUN_CONJ_NOUN*	3	*NOUN_ADP_NOUN_CONJ*	1
8	*VERB_NOUN*	2	*ADP_NOUN_NOUN*	2	*NOUN_ADP_NOUN_VERB*	3
9	*NOUN_ADV*	1	*NOUN_VERB_CONJ*	1	*NOUN_VERB_CONJ_VERB*	1
10	*ADJ_NOUN*	2	*ADV_ADP_NOUN*	3	*NOUN_ADV_ADP_NOUN*	4

Table VI
MOST FREQUENT N-SYNTAX PATTERNS AND LOCATION ENTITY INDEXES (LOS ANGELES)

Rank	2-Syntax Pattern	Index	3-Syntax Pattern	Index	4-Syntax Pattern	Index
1	*ADP_NOUN*	2	*NOUN_ADP_NOUN*	3	*ADP_NOUN_ADP_NOUN*	4
2	*NOUN_ADP*	1	*NOUN_ADP_NOUN*	1	*ADP_NOUN_ADP_NOUN*	2
3	*NOUN_VERB*	1	*ADP_NOUN_ADP*	2	*NOUN_ADP_NOUN_ADP*	3
4	*NOUN_NOUN*	1	*ADV_ADP_NOUN*	3	*NOUN_ADV_ADP_NOUN*	4
5	*NOUN_NUM*	2	*ADP_NOUN_VERB*	2	*NOUN_ADP_NOUN_VERB*	3
6	*NOUN_CONJ*	1	*NOUN_VERB_NOUN*	1	*NOUN_ADP_NOUN_VERB*	1
7	*ADP_NUM*	2	*ADP_NOUN_NOUN*	2	*ADP_NOUN_VERB_NOUN*	2
8	*NOUN_NUM*	1	*NOUN_NOUN_ADP*	1	*NOUN_VERB_NOUN_ADV*	1
9	*NUM_VERB*	1	*ADP_NOUN_CONJ*	2	*ADP_NOUN_NOUN_ADP*	2
10	*CONJ_NOUN*	2	*ADP_NOUN_NUM*	3	*NOUN_NOUN_ADP_NOUN*	1

top-ranked frequent n-Syntax patterns (e.g., "*ADP_NOUN*" for 2-Syntax Pattern, and "*NOUN_ADP_NOUN*" for 3-Syntax Pattern). Such a similarity demonstrates the potential capability of SyntaxLoc framework in a transfer learning setting [30], e.g., we can first train a model using the dataset collected in New York City and then transfer the learned model to identify the location entities from the social media posts collected in Los Angeles.

VI. CONCLUSION

This paper develops a novel SyntaxLoc framework to solve the abnormal traffic event localization problem using social sensing. The SyntaxLoc framework addresses two critical challenges: *content-only inference* and *informal and scarce data*. In particular, we develop a syntax-based pattern learning module and a probabilistic-based entity extraction module to accurately identify the location entities in a social media post. The evaluation results demonstrate significant performance gains of our SyntaxLoc framework compared to state-of-the-art baselines by identifying the location entities from social sensing data more accurately.

ACKNOWLEDGEMENT

This research is supported in part by the National Science Foundation under Grant No. CNS-1831669, CBET-1637251, Army Research Office under Grant W911NF-17-1-0409. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] D. Wang, T. Abdelzaher, and L. Kaplan, *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann, 2015.

[2] D. Y. Zhang, C. Zheng, D. Wang, D. Thain, X. Mu, G. Madey, and C. Huang, "Towards scalable and dynamic social sensing using a distributed computing framework," in *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE, 2017, pp. 966–976.

[3] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2013, pp. 344–353.

[4] J. Lingad, S. Karimi, and J. Yin, "Location extraction from disaster-related microblogs," in *Proceedings of the 22nd international conference on world wide web*. ACM, 2013, pp. 1017–1020.

[5] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 759–768.

[6] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser, "A multi-indicator approach for geolocalization of tweets," in *Seventh international AAAI conference on weblogs and social media*, 2013.

[7] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012.

[8] S. Kinsella, V. Murdock, and N. O'Hare, "I'm eating a sandwich in glasgow: modeling locations with tweets," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM, 2011, pp. 61–68.

[9] M. T. Rashid, D. Y. Zhang, Z. Liu, H. Lin, and D. Wang, "Collab-drone: A collaborative spatiotemporal-aware drone sensing system driven by socialsensing signals," to appear in 2019 28th International Conference on Computer Communication and Networks (ICCCN). IEEE, 2019, accepted.

[10] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[11] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 61–70.

[12] D. Y. Zhang, D. Wang, H. Zheng, X. Mu, Q. Li, and Y. Zhang, "Large-scale point-of-interest category prediction using natural language processing models," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 1027–1032.

[13] O. Ajao, J. Hong, and W. Liu, "A survey of location inference techniques on twitter," *Journal of Information Science*, vol. 41, no. 6, 2015.

[14] T. B. N. Hoang and J. Mothe, "Location extraction from tweets," *Information Processing & Management*, vol. 54, no. 2, 2018.

[15] Y. Zhang, N. Vance, D. Zhang, and D. Wang, "On opinion characterization in social sensing: A multi-view subspace learning approach," in *2018 14th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2018, pp. 155–162.

[16] D. Y. Zhang, R. Han, D. Wang, and C. Huang, "On robust truth discovery in sparse social media sensing," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1076–1081.

[17] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, 2007.

[18] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *Computer*, vol. 52, no. 1, pp. 36–45, 2019.

[19] D. Y. Zhang, N. Vance, and D. Wang, "When social sensing meets edge computing: Vision and challenges," to appear in 2019 28th International Conference on Computer Communication and Networks (ICCCN). IEEE, 2019, accepted.

[20] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Information Processing in Sensor Networks (IPSN), 2012 ACM/IEEE 11th International Conference on*. IEEE, 2012, pp. 233–244.

[21] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications," in *2013 IEEE 33rd International Conference on Distributed Computing Systems*. IEEE, 2013, pp. 530–539.

[22] D. Y. Zhang, Y. Zhang, Q. Li, T. Plummer, and D. Wang, "Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications," in *Distributed Computing Systems (ICDCS), 2019 IEEE 39th International Conference on*. IEEE, 2019.

[23] Y. Zhang, Y. Lu, D. Zhang, L. Shang, and D. Wang, "Risksens: A multi-view learning approach to identifying risky traffic locations in intelligent transportation systems using social and remote sensing," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1544–1553.

[24] N. Vance, T. Rashid, , D. Y. Zhang, and D. Wang, "Towards reliability in online high-churn edge computing: A deviceless pipelining approach," in *The 5th IEEE International Conference on Smart Computing (SMARTCOMP 2019)*. IEEE, 2019.

[25] Y. Zhang, D. Zhang, N. Vance, and D. Wang, "Optimizing online task allocation for multi-attribute social sensing," in *2018 27th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 2018, pp. 1–9.

[26] Z. Li, J. Liu, X. Zhu, T. Liu, and H. Lu, "Image annotation using multi-correlation probabilistic matrix factorization," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1187–1190.

[27] L. S. Zettlemoyer and M. Collins, "Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars," *arXiv preprint arXiv:1207.1420*, 2012.

[28] Y. Huang, Q. Liu, S. Zhang, and D. N. Metaxas, "Image retrieval via probabilistic hypergraph ranking," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3376–3383.

[29] Y. Gu, Z. S. Qian, and F. Chen, "From twitter to detector: Real-time traffic incident detection using social media data," *Transportation research part C: emerging technologies*, vol. 67, pp. 321–342, 2016.

[30] Y. Zhang, H. Wang, D. Zhang, and D. Wang, "Deeprisk: A deep transfer learning approach to migratable traffic risk estimation in intelligent transportation using social sensing," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2019.