Social Event Extraction: Task, Challenges and Techniques

Hao Li
Computer Science
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: lih13@rpi.edu

Heng Ji
Computer Science
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: jih@rpi.edu

Lin Zhao
Research and Technology Center
Robert Bosch LLC
Palo Alto, California 94304
Email: lin.zhao@us.bosch.com

***Abstract*—Social media (e.g., Facebook and Twitter) serves as a popular platform for online communication and information dissemination, where users can effectively share information such as their recent activities and plans. This kind of information is extremely valuable for building recommendation systems. For example, a user might wish to receive alerts whenever a concert takes place near his current location or when a party will be held in his neighborhood. However, such events may not be widespread across social networks, they have not received sufficient attention. Moreover, traditional event discovery and event extraction techniques trained from formal genres cannot be effectively adapted to this domain. In this paper, we present the first formal definition of** *social events*, **discuss the annotation challenges and release a benchmark for the research community. Further more, we propose two novel solutions for extracting elements from social events: (1) an unsupervised content segmentation framework to extract event phrases (2) utilize external knowledge bases to detect fine-grained event locations and unveil their background information. Experimental results convincingly demonstrate that our approach can accurately extract social events from social media.**

Keywords—Social Event; Event Extraction; Social Media

I. Introduction

Event extraction is a task of identifying events and their participants (arguments) from documents. Traditional event extraction work mainly focused on formal genres such as newswire (e.g. [1], [2], [3], [4]). For example, the sentence *"the US-led coalition troops are reportedly thrusting into the second Iraqi city of Basra."* includes a *"Movement-Transport"* event indicated by a trigger word *"thrusting"*, and a set of event arguments: the Person entity *"troops"* as the Artifact and the Geo-political entity *"Basra"* as the Destination.

With the rapid development of social media and social networks, there has been an increasing interest in detecting novel or popular events from tweet stream (e.g. [5], [6], [7], [8], [9]). Meanwhile, some researchers focus on location based event detection where the events are associated with fine-grained locations (e.g. [10], [11]) from social media. However, most previous work suffered from one or more of the following problems: (1) lack of an unified representation; (2) heavily relied on information redundancy or geo-tagging data; (3) do not contain fine-grained event types.

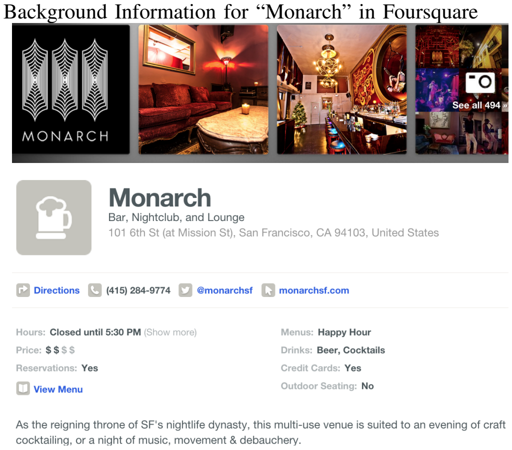
In this paper, we aim to address the above problems for the task of social event extraction. A social event is an event that is well organized and targets at a group of participants to gather at a certain time and location but may not be trending in the social media stream. For example, the following Facebook post *"Check out one of our favorite songs by Revenant. They'll be performing live in San Francisco at our next event Thursday May 22nd @ Monarch."* is about an *"Art & Entertainment"* event triggered by *"performing live"* in *"Monarch Bar"* on *"Thursday May 22nd"*.

The merits of social event extraction are profound: (1) In contrast to traditional social networks which are mainly based on acquaintance ("I know you"), social event extraction plays an important role in constructing interests/preferences ("I like it") based social networks. Thus it generates a more open, richer and efficient social network structure. (2) Unlike the location arguments in coarse-grained events which mainly focus on countries and cities such as *"U.S"* and *"San Francisco"*, social event extraction can not only identify Point Of Interests (POIs) such as *"Monarch Bar"* but can also unveil its background information (type, address, etc. in Figure 1). It further enables the location-based services (LBS) such as location-based social activity recommendation. It will be extremely helpful if the localized activity recommendation system can push notification to a music fan when getting close to the location where his favorite band performs.

However, the existing event extraction methods are inappropriate for this task because of the following unique characteristics:

1) The performance of existing name tagger trained from formal genres inevitably degrades when they are adapted to identify important event arguments such as event titles (e.g. *"Tech in Motion"*) in informal genres. Even the Twitter name tagger [12] with tailored POS tagger and shallow parsing can only achieve 51% F-socre for named entity segmentation and 66% F-score for named entity classification in tweets [12]. The performance is much worse than state-of-the-art name tagging for formal genres. [13].
2) Fine-grained locations are more important than coarse-grain locations in social events. In the previous example, it is extremely useful to know the event location is *Monarch Bar* and its detailed information, instead of only knowing it is in *San Francisco*. However, both Stanford name tagger [14] and Twitter name tagger fail to identify *"Monarch"* as an event location, therefore it is impossible to further pinpoint *"Monarch"* as a bar.

Fig. 1. Background Information for “Monarch” in Foursquare



Monarch
Bar, Nightclub, and Lounge
101 6th St (at Mission St), San Francisco, CA 94103, United States

Directions (415) 284-9774 @monarchsf monarchsf.com

Hours: Closed until 5:30 PM (Show more)
Price: $ $ $ $
Reservations: Yes
View Menu

Menus: Happy Hour
Drinks: Beer, Cocktails
Credit Cards: Yes
Outdoor Seating: No

As the reigning throne of SF's nightlife dynasty, this multi-use venue is suited to an evening of craft cocktailing, or a night of music, movement & debauchery.

To solve these problems, we first discuss the annotation challenges. Furthermore, we propose an language model based message segmentation framework to extract event phrases and utilize Foursquare[1] and Yelp[2] services to detect fine-grained event locations.

The major novel contributions of this paper are as follows:

- To the best of our knowledge, this is the first work on automatic event extraction for social events. We discuss the annotation challenges and will release our annotated corpus which could serve as a benchmark for the research community.
- We propose a general yet effective framework to extract event phrases.
- We resolve fine-grained event locations and their background information by detecting location candidates and linking them to external knowledge bases (e.g. Foursquare and Yelp).

II. Annotation Challenges for Social Events

A. *Definition and Representation*

Before conducting human annotation for social events, we first need to decide the representation for social events. To the date there are three most popular event representations: Coarse-grained predicate-argument representations, Fine-grained event trigger and argument representations and Fine-grained event tuple representation. Table I lists the comparison among these three event representations.

All of these existing representations do not well fit social events: Definition 1 does not provide the cluster type for each predicate; Definition 2 is inadequate as social events include open argument types such as performers, sports team, etc.; Definition 3 focuses on general domain and utilizes a tweet cluster to represent each event type. However, fine-grained social event types are needed in our scenario. Also it is not able to identify fine-grained event locations which are critical for local social events. In this paper we propose a new tuple representation <event type, event phrase, time, location> to represent each social event. For the previous sentence, the corresponding tuple representation will be <Art & Entertainment, performing live, May 22nd, Monarch>. This new representation has the following advantages:

1) It is able to provide brief but sufficient information of a particular social event. We could also incorporate with other NLP techniques such as Named Entity Recognition (NER) and Key Word Mining to unveil more details of a social event.
2) It can well distinguish two social events. By intuition two different social events cannot happen on the same day at the same location. Some instances such as multiple bands playing in a bar in the same evening are not distinct events, but a lineup of multiple artists. Some events (e.g. a festival) can last several days with a lineup of artists and will be considered as a single event.

B. *Annotation Challenges*

The annotation process for time and location is straightforward. We tag the time expression that refers to the time of a social event. We annotate both coarse-grained and fine-grained locations for each social event. Coarse-grained event locations are areas such as “San Francisco” and “Oakland” while fine-grained event locations are POIs such as “War Memorial Opera House”and “Monarch Bar”. We confirm a coarse-grained location with respect to a gazetteer[4] while confirm a fine-grained location by linking it to Foursquare or Yelp.

The main annotation challenges lie in the following two event elements: social event type and event phrases.

1) Social Event Type: We aim to not only provide a fine-grained event type definition similar to ACE and DEFT, but also cover as many as the common local social event types.

In order to define the types of social events, we make use of the information from Foursquare and Yelp which are the two most influential location-based social networking (LBSN) platforms that provide professional local search and discovery service. The advantages of using Foursquare and Yelp are as follows: as each social event type has a latent connection with its associated location type (e.g., people attend concerts at music halls whereas watch a football games in stadiums), we can easily imagine the event type through its location type. We can take advantage of the high quality location type hierarchies Foursquare and Yelp provided. The second and third column in Table III present the top level category-mapping between Foursquare and Yelp that introduces the general view of social event locations. Accordingly, we can conclude the corresponding social event types in the first column as one of the following types: Art & Entertainment, Active Life, Professional, Party and Other. Table II shows the examples of each social event type.

The social event type annotation includes two steps. We first evaluate if the post includes a social event; then we categorize social events into the following subtypes: Art & Entertainment, Active Life, Professional, Party and Other. The

[1] https://foursquare.com/

[2] https://yelp.com/

[4] www.census.gov/geo/maps−data/data/gazetteer2014.html

TABLE I. EXISTING EVENT REPRESENTATIONS

Representation	Example	Advantage	Disadvantage
Coarse-grained predicate-argument representations	Propbank [15], [16] and FrameNet [17]	each single predicate as an event type thus covers a wide range of event types	does not attempt to cluster the predicates; assigns coarse-grained roles to arguments
Fine-grained event trigger and argument representations	NIST Automatic Content Extraction (ACE) program and the DARPA Deep Exploration and Filtering of Texts (DEFT) program's Event Mention Detection and Event Argument Extraction task [3]	clusters predicates and defines fine-grained argument roles; event arguments include both participants and some context properties such as time, place and logistics (e.g., instrument, methods)	limited pre-defined event types and event arguments
Fine-grained event tuple representation	TwiCal-Event [18]	high coverage of event types; differentiate event phrases and named entities	coarse-grained social event types and event locations

TABLE II. SOCIAL EVENT EXAMPLES

Social Event Type	Example
Art & Entertainment	Last night's Madame Butterfly was one of the best SF Opera productions I've ever seen.
Active Life	LADIES: Tomorrow is the Day! - LADIES' STYLING with Shahla Fisher, SAT. MAY 31, 12:30 - 2:30. At Allegro Dance.
Professional	Have you RSVP'd for Tech in Motion: OC's Spring into Networking and Tech Mixer at Spireon? This event will feature ping-pong, basketball and amazing OC techies on April 24th!
Party	Ticket prices for our Streets of SF NYE party at Fort Mason with MOBY will increase at the end of the day this Wednesday
Other	Don't miss our Costume Shop Sale on March 22 & 23–your opportunity to pick up the items that will wow ALL of your friends the next time you want to dress up in style!

TABLE III. FOURSQUARE/YELP ROOT CATEGORY AND LOCAL SOCIAL EVENT TYPE MAPPING

Local Social Event Type	Foursquare Categories	Yelp Categories
Art & Entertainment, Party	Nightlife Spot	Nightlife
Art & Entertainment	Arts & Entertainment	Arts & Entertainment
Active Life	Outdoors & Recreation	Active Life
Professional	College & University	Educations
Party	Food	Food Restaurant
Active Life, Other	Travel & Transport	Bicyles Hotels & Travel
Mixed of the five types	Event	Local Flavor
Professional, Other	Professional & Other Services	Automotive Mass Media Public Service Religious Org. Health & Medical Professional Service
Professional, Other	Shop & Service	Pets Shopping Beauty & Spas Home Services Local Services Financial Services Event Planning & Sevc. Public Services & Gov.

TABLE IV. EVENT PHRASE ANNOTATION EXAMPLE

Post	Event Phrases	Social Event Type
Are you going to Phono Del Sol this weekend?	{going, Phono Del Sol}	Art & Entertainment
Great to have Mobile First Entertainment at Tech in Motion Events OC last night!	{Tech in Motion}	Professional

Art & Entertainment type includes music venues, operas, shows, etc; the Active Life type covers all events related to sports or outdoor activities; the Professional type includes conferences, workshops, technical events, educational event, etc; Party type is about gathering together for some special purpose.

2) Local Social Event Phrase: An event phrase is a sequences of words which are most representative of a particular event. It is closely related to the event triggers defined in Automatic Content Extraction (ACE)[5]. According to the ACE annotation guideline, an event trigger is defined as the word that precisely describes the occurrence of an event. In most cases, an event trigger will be the main verb in the sentence that describes the event directly. For example, *"died"* triggers an *DIE* event in *"He died yesterday of renal failure."*. An event trigger can also be a participle and an adjective (*"rioting"* triggers an *DEMONSTRATE* event in *"The rioting crowd approached the Capitol."*) or a noun and pronoun. (*"meeting"* triggers an *MEET* event in *"Protestors interrupted their meeting."*).

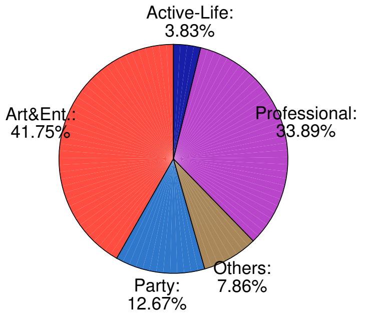
However, the similar scheme for the event phrases in local social events are problematic because of two major difference between social event phrases and event triggers:

- some triggers of social events are so generic that they may appear as more than one event type. In such cases, the event argument sometimes plays a crucial role in determining the event type. For example, in the post *"Are you going to Phono Del Sol this weekend?"*, *"going"* is the event trigger but without knowing its argument *"Phono Del Sol"*, the name of a music festival, we are not able to tell that the post discusses an Art & Entertainment social event. In another post *"Great to have Mobile First Entertainment at Tech in Motion Events OC last night!"*, without knowing two event arguments: a tech company *"Mobile First Entertainment"* and the tech event title *"Tech in Motion"*, the event trigger *"have"* is meaningless.
- Social event phrases are more likely to include multiple words. For example, in ACE training data, more than 95% of the event triggers in the training data consist of a single word. However, 53% of the event phrases have more than one word in social event phrases.

Because of these unique aspects of social events, we decide to annotate event phrases as the minimal set of phrases (could be both event triggers and arguments) that can specifically determine the social event type. Table IV shows the event phrases of the previous two Facebook posts.

[5] http://www.itl.nist.gov/iad/mig/tests/ace/

Fig. 2. Event Type Distribution



III. Data

There exist some tweets datasets but they are not appropriate for our task. One reason is that the percentage of “social events” in tweets is low. For example, in the Twi-Cal event dataset [18], less than 5% of the tweets include social events and the streaming tweets will contain even fewer social events. The other reason is that some tweet datasets covered a limited number of social event types. For example, the tweet data in [19] contains only Art & Entertainment social events. As a result, we decided to gather the data from Facebook public accounts as they are setup by a group of people sharing similar interests thus have high probability to include social event information.

To retrieve Facebook data through its API[6], we followed five Facebook public pages and joined five Facebook groups. Among them, there are two for Art & Entertainment events, one for Active Life events, two for Professional events, one for Party events, and others include mixture type of social events. We collected all posts from users and the public account page, together with all comments. After filtering out the trivial posts that include only URLs or have fewer than five words, we randomly selected 1600 posts for annotation and evaluation.

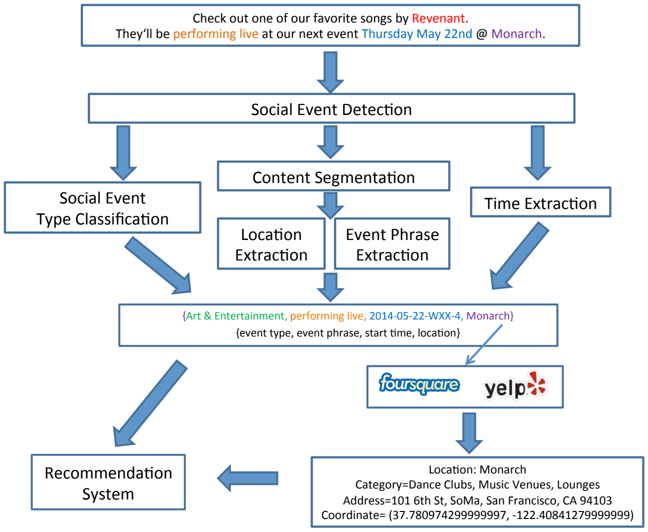
In our annotated corpus, the percentage of social event related posts is 63.6%. Figure 2 presents the social event distributions in our corpus. Around 76.2% and 28.3% of the social events are associated with explicit time expressions and fine-grained locations respectively.

During the human annotation process, there are seven annotators for each event element annotation step. The average inter-agreement for social event detection and social event categorization are 90% and 85%, respectively. The inter-agreement for event phrase annotation is 30%. It is extremely low because of two reasons: it is a challenging task and we allow annotating multiple event phrases per Facebook post. The average inter-agreement for both time and location annotations are approximately 90%.

IV. Approach

In this section, we will introduce our approach to tackle the challenges of each event extraction component. Figure 3 is an example that illustrates the roadmap: two-stage event type classification and time expression extraction; Unsupervised Event Phrase Extraction; and fine-grained event location extraction.

Fig. 3. Social Event Extraction Pipeline


Check out one of our favorite songs by Revenant.
They'll be performing live at our next event Thursday May 22nd @ Monarch.
Social Event Detection
Social Event Type Classification
Content Segmentation
Time Extraction
Location Extraction
Event Phrase Extraction
{Art & Entertainment, performing live, 2014-05-22-WXX-4, Monarch}
{event type, event phrase, start time, location}
foursquare
yelp
Recommendation System
Location: Monarch
Category=Dance Clubs, Music Venues, Lounges
Address=101 6th St, SoMa, San Francisco, CA 94103
Coordinate= (37.780974299999997, -122.40841279999999)

A. Event Type Classification and Time Extraction

Given a Facebook post, the first step is to first identify whether it mentions a social event. After that we classify the identified social events into one of the five types: Art & Entertainment, Active Life, Professional, Party and Other. To tackle the problem, we have developed a two stage supervised learning model based on Support Vector Machines (SVMs) to classify social event types for each post. The first stage is a binary classifier to predict social/non-social events and the second stage is a multi-class classifier for five social event types. We incorporate the following features into two clssifiers:

- Content: Bag-of-Words and bigrams
- Wordnet: unigrams that are under three WordNet hierarchies: social group, social event and group action
- Meta data: binary feature indicating the appearance of URL links; type of the Facebook post
- Name: the total number and the string representation of each type of the named entities output extracted by Stanford name tagger and Twitter name tagger
- Other: tense label (past, present or future) and sentiment label (positive, neutral or negative)

To resolve temporal expressions we make use of SUTime [20], which takes a reference date, some text as input, output marked temporal expressions with unambiguous calendar references. Although this mostly rule-based temporal tagger is built on regular expression patterns and was designed for use on newswire text, its precision on Facebook posts achieved 90% accuracy which is a reasonable performance for our purpose. SUTime's high precision on Facebook posts can be explained by the fact that some temporal expressions are relatively unambiguous [18].

B. Unsupervised Event Phrase Extraction

As we mentioned in Section II-B2, some event arguments are crucial and should be included as event phrases. However, most local social events are less well-known thus their event arguments (e.g. event title *“Tech In Motion”*, band *“Turn Me On Dead”*) are tail entities which have limited web presence

[6] https://developers.facebook.com/

TABLE V. Possible Content Segmentation Example

<table>
<tr><td>Original</td><td>Tech in Motion Silicon Valley Women In Tech panel</td></tr>
<tr><td>Segmentation 1</td><td>Tech in Motion|Silicon Valley|Women In Tech|panel</td></tr>
<tr><td>Segmentation 2</td><td>Tech in Motion|Silicon Valley Women|In|Tech panel</td></tr>
<tr><td>Segmentation 3</td><td>Tech|in|Motion Silicon Valley Women|In|Tech|panel</td></tr>
</table>

and popularity. We manually selected fifty named entities that serve as event arguments in social events and none of them has an entry in the Wikipedia database. For these cases, dictionary-based features in supervised systems may not work well. Moreover, supervised POS taggers are very likely to fail on these cases. For example, Stanford POS tagger assign the tag sequence "NN IN NN" to "Tech In Motion" and "VB PRP IN JJ" to "Turn Me On Dead". It makes the existing supervised name taggers difficult to detect them correctly. For example, both Stanford name tagger and Twitter name tagger failed to identify *"Tech In Motion"* and *"Turn Me On Dead"* correctly, not even be able to detect their boundaries.

Considering the challenges and the drawbacks of supervised methods, we proposed to adapt unsupervised content segmentation to extract important event phrases using web scale data [21], [22]. The intuition is to have a good "prior" distribution $p(w_1...w_n)$ over which phrases $w_1, ...w_n$ are or aren't probable in a language. Given a Facebook post $p \in P_i$, the problem is to split it into n consecutive segments, $p = s_1 s_2 ... s_n$. The optimal split should follow the objective function:

$$\arg\max_{s_1,s_2...s_n} f(p) = \sum_{i=1}^{m} f(s_i) \qquad (1)$$

where $f(x)$ measures the stickness of a segment based on word collocation. In the paper, we use PMI based stickness function for a segment $s = w_1...w_n$ as follows:

$$PMI(s) = log \frac{Pr(w_1...w_n)}{\frac{1}{n-1}\sum_{i=1}^{n-1} Pr(w_1...w_i)Pr(w_{i+1}...w_n)} \qquad (2)$$

Each Facebook post of T terms can have as many as 2^{T-1} segmentations. The brief algorithm is to generate all possible segment boundaries in a Facebook message and rank their likelihoods using the N-gram service. Table V presents three possible segmentation results for *"Tech in Motion Silicon Valley Women In Tech panel"*. Segmentation 1 is the desired output with the highest likelihood. Here *"Tech in Motion"* is the title of a professional social event, *"Silicon Valley"* is a coarse-grained event location and *"Women In Tech"* is the panel name.

To precisely estimate the prior probabilities $Pr(s)$, we need a sufficiently large corpus as the global context of each segment. In this paper, we utilize Microsoft Web N-gram corpus [23] to approximate such prior probabilities. Microsoft Web N-gram corpus is constructed from the web documents indexed by the Microsoft Bing service in the EN-US market. After filtering out spam and low quality web pages, the URLs visited by Bing are at the order of hundreds of billion. Different type of the web documents are then downloaded, parsed and tokenized by Bing. It provides a sufficient English corpus to estimate prior probabilities of segments.

Note that an alternative content segmentation method is based on chunking as it is able to identify the constituents (noun groups, verbs, verb groups, etc.) within the sentence. However, it requires large amount of training data and acurate POS tags thus is not able to produce satisfying results for our task (e.g. Segmentation 3 is the actual OpenNLP[7] chunking output).

We group the same event phrases (e.g. "SF opera" and "San Francisco Opera") generated by Section IV-B using a set of heuristic rules. Then we construct an event phrase graph $G(V, E)$ where each node is a segment after noise filtering, and the edge $e_a b$ between two nodes (segments) s_a and s_b is weighed by the Jaccard Index:

$$w(e_a b) = \frac{|M(s_a) \cap M(s_b)|}{|M(s_a) \cup M(s_b)|} \qquad (3)$$

where M(s) is the set of posts containing segment s.

Then we follow Li et al. [22] to apply the random walk model on the graph to achieve the final ranking of the event phrases. The intuition is that event phrases tend to co-occurs more often than those are not. In each post, we select three phrases with highest ranking scores as event phrases.

C. Event Location Extraction

LBSN platforms such as Foursquare and Yelp provide APIs to query their fine-grained location databases. Given a pair of query of <keyword, radius location>, the API is able to retrieve the top-K fine-grained location entries that are most relevant to the query near the radius location. We utilize the segments extracted by Section IV-B together with Foursquare API and Yelp API to identify the fine-grained event locations. The detailed steps for extracting the fine-grained location for each Facebook post is as follows:

(1) Extract the coarse-grained location by Stanford name tagger. It will be used as the radius location in Step 3.

(2) Identify fine-grained location candidates by filtering the segments with the following heuristic rules: (1). the segments should be noun phrases (2) the segments should have at least one word beginning with a upper-case letter (3). the segments should be unlikely to be tagged as person names by any NER tools.

(3) Pair each candidate segment from Step 2 with the coarse-grained location from Step 1 to query the Foursquare and Yelp database and retrieve the top-K entries.

(4) For each returned entry, judge if it is the same as the segment based on their name similarity. If their name similarity is higher than a threshold then the segment is the fine-grained event location.

V. Experiments

Table VI shows the accuracy for the two-stage social event type classification. All experiments are conducted using libSVM [24] with RBF kernel in the 5-fold cross validation setting. The main reason of the errors is that some posts require background knowledge to make the correct prediction. For

[7] https://opennlp.apache.org/documentation.html

TABLE VI. PERFORMANCE OF EVENT TYPE CLASSIFICATION

Features	1st Stage	2nd Stage
(1): BoW	82.13%	76.42%
(2): (1)+Bigram	84.25%	78.51%
(3): (2)+WordNet	84.81%	**78.88%**
(4): (3)+Meta	**85.00%**	78.64%

example, without the knowing *Phono Del Sol* is the name of a music festival, it is difficult to correctly classify the post *Are you going to Phono Del Sol this weekend?* as an Art & Entertainment social event.

For the segmentation settings, we chose "phrase-bing-query/2013-12/1" as the Microsoft Web N-gram model and set the maximum length of a possible segment up to five. We randomly picked up and evaluated segmentation output of 100 Facebook messages, Web N-gram approach is able to surpass the chunking based method by a large margin.

To evaluate the performance of event phrase extraction, for each post we use the union set of event phrase from all annotators as the answer pool, if the system generated event phrase is able to find a match in the pool, it is treated as a correct one. As there is no existing baseline system for this task, we adapt the supervised system TWI-CAL [18] to conduct a fair comparison: we extract all the named entities and event mentions as the event phrases and the performance is only 28%. Our unsupervised event phrase extraction component is able to achieve a 34.0% performance. The reason of the low performance is the limit amount of data thus the concept graph is sparse.

For Event Location Extraction, our approach is able to achieve 38.8% accuracy, 78.3% recall and 51.89% F-1 score. There are several reasons for the errors:

(1) The segmentation errors result in incorrect boundaries thus we miss some location candidates.
(2) The location mentioned in the post is not included in Foursquare and Yelp database.
(3) The location mentioned in the post is ambiguous. For example, both *San Francisco Opera* and *AT&T park* are location entries in Foursquare/Yelp in the post *"on July 5th join the San Francisco Opera as we fill AT&T Park with drama, passion and glorious music..."*. However, the correct event location is *AT&T park* while *San Francisco Opera* represents the event series name.
(4) Many of the mentions of fine-grained locations are partial location names. For example, the music venue *Mezzanine* is mentioned as *MEZZ* in the post *"Boys Noize this Thursday at MEZZ is almost sold out."*. The short, ambiguous, and partial names make the problem of linking location candidates to Foursquare and Yelp extremely challenging.

VI. RELATED WORK

Event Detection in Social Media

Both [26] and [27] aim at grouping content available in social media applications such as Flickr, Youtube, Panoramino etc. into clusters of documents describing the same event. However, in our paper, we focus on predicting the social event type of each individual facebook message, more important, to extract the event tuple from the message content.

[5] analyzed signals in the frequency domain. They applied Discrete Fourier Transformation (DFT) to convert the signals from the time domain into the frequency domain. A spike in the frequency domain corresponded to a trending event. [7] tackled event discovery task for Twitter by detecting important word tokens and clustering them to represent novel events then analyzed word-specific signals in the time domain. [25] ranked clustered tweet segments to discover novel events. However, these work aimed to identify frequently discussed events based on information redundancy, which is not appropriate for local social events characteristics. In addition, compared with their approaches of representing an event as a set of words, we proposed to ouput an event in a fine-grained structure. [18] is similar to our approach in representing each event in a tuple, however, they focused on more general domain and is not capable to handle fine-grained location extraction problem.

Location Extraction in Social Media

Early work [28] tried to extract locations from disaster tweets by retaining existing NER tools including Standford-NER, OpenNLP and TwitterNLP. The experiments showed that extracting fine-grained locations remains inferior. Some recent work [29] utilized checkin tweets from Foursquare as external knowledge to boost the trained linear chain CRF model to extract time-aware locations. Compare to their work, we focus on local social event related locations instead of the time dimension.

VII. CONCLUSION

To summarize, in this paper we present the task of social events extraction. We discuss the annotation challenges and two main technical challenges of extracting elements from social events. We then propose two novel solutions to tackle the problems: an unsupervised message segmentation framework to extract event phrases and external resources to detect fine-grained event locations and unveil their background information. Experimental results show that our approach achieved promising performance. In the future, we would like to improve the content segmentation performance through collective inference. Furthermore, we plan to conduct research on clustering Facebook messages and tweets that mention the same social events and construct a joint social event extraction model for both genres.

ACKNOWLEDGMENT

This work was supported by the U.S. DARPA Award No. FA8750-13-2-0045 in the Deep Exploration and Filtering of Text (DEFT) Program,the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), U.S. NSF CAREER Award under Grant IIS-0953149, U.S. AFRL DREAM project, U.S. Department of Homeland Security CCICADA, IBM Faculty Award, Google Research Award, Disney Research Award, Bosch Research Award, and RPI faculty start-up grant. 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 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] H. Ji and R. Grishman, "Refining event extraction through cross-document inference," in *ACL*, 2008, pp. 254–262.

[2] Q. Li, H. Ji, and L. Huang, "Joint event extraction via structured prediction with global features," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, 2013, pp. 73–82. [Online]. Available: http://aclweb.org/anthology/P/P13/P13-1008.pdf

[3] S. Liao and R. Grishman, "Using document level cross-event inference to improve event extraction," in *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, 2010, pp. 789–797. [Online]. Available: http://www.aclweb.org/anthology/P10-1081

[4] Y. Hong, J. Zhang, B. Ma, J. Yao, G. Zhou, and Q. Zhu, "Using cross-entity inference to improve event extraction," in *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, 2011, pp. 1127–1136. [Online]. Available: http://www.aclweb.org/anthology/P11-1113

[5] Q. He, K. Chang, and E.-P. Lim, "Analyzing feature trajectories for event detection." in *SIGIR*, W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, Eds. ACM, 2007, pp. 207–214.

[6] S. Petrovic, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *HLT-NAACL*, 2010, pp. 181–189.

[7] J. Weng and B.-S. Lee, "Event detection in twitter," in *ICWSM*, 2011.

[8] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," in *ICWSM*, 2011.

[9] R. McCreadie, C. Macdonald, I. Ounis, M. Osborne, and S. Petrovic, "Scalable distributed event detection for twitter," in *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, 2013, pp. 543–549. [Online]. Available: http://dx.doi.org/10.1109/BigData.2013.6691620

[10] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, "Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs," in *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, 2011, pp. 2541–2544. [Online]. Available: http://doi.acm.org/10.1145/2063576.2064014

[11] H. Abdelhaq, C. Sengstock, and M. Gertz, "Eventweet: Online localized event detection from twitter," *PVLDB*, vol. 6, no. 12, pp. 1326–1329, 2013. [Online]. Available: http://www.vldb.org/pvldb/vol6/p1326-abdelhaq.pdf

[12] A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named entity recognition in tweets: An experimental study," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2011, pp. 1524–1534. [Online]. Available: http://www.aclweb.org/anthology/D11-1141

[13] Q. Li and H. Ji, "Incremental joint extraction of entity mentions and relations," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, 2014, pp. 402–412. [Online]. Available: http://aclweb.org/anthology/P/P14/P14-1038.pdf

[14] J. R. Finkel, T. Grenager, and C. D. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, 2005. [Online]. Available: http://acl.ldc.upenn.edu/P/P05/P05-1045.pdf

[15] M. Palmer, P. Kingsbury, and D. Gildea, "The proposition bank: An annotated corpus of semantic roles," *Computational Linguistics*, vol. 31, no. 1, pp. 71–106, 2005.

[16] N. Xue and M. Palmer, "Adding semantic roles to the chinese treebank," *Natural Language Engineering*, vol. 15, no. 1, pp. 143–172, 2009.

[17] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The berkeley framenet project," in *COLING-ACL*, 1998, pp. 86–90.

[18] A. Ritter, Mausam, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, 2012, pp. 1104–1112. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339704

[19] E. Benson, A. Haghighi, and R. Barzilay, "Event discovery in social media feeds," in *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, 2011, pp. 389–398. [Online]. Available: http://www.aclweb.org/anthology/P11-1040

[20] A. X. Chang and C. D. Manning, "Sutime: A library for recognizing and normalizing time expressions," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, 2012, pp. 3735–3740. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/summaries/284.html

[21] D. Downey, M. Broadhead, and O. Etzioni, "Locating complex named entities in web text," in *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, 2007, pp. 2733–2739. [Online]. Available: http://dli.iiit.ac.in/ijcai/IJCAI-2007/PDF/IJCAI07-439.pdf

[22] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B. Lee, "Twiner: named entity recognition in targeted twitter stream," in *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, 2012, pp. 721–730. [Online]. Available: http://doi.acm.org/10.1145/2348283.2348380

[23] K. Wang, C. Thrasher, E. Viegas, X. Li, and B. P. Hsu, "An overview of microsoft web n-gram corpus and applications," in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2, 2010, Los Angeles, California, USA - Demonstration Session*, 2010, pp. 45–48. [Online]. Available: http://www.aclweb.org/anthology/N10-2012

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[25] C. Li, A. Sun, and A. Datta, "Twevent: segment-based event detection from tweets," in *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, 2012, pp. 155–164. [Online]. Available: http://doi.acm.org/10.1145/2396761.2396785

[26] T. Reuter and P. Cimiano, "Event-based classification of social media streams," in *International Conference on Multimedia Retrieval, ICMR '12, Hong Kong, China, June 5-8, 2012*, 2012, p. 22. [Online]. Available: http://doi.acm.org/10.1145/2324796.2324824

[27] T. Reuter, P. Cimiano, L. Drumond, K. Buza, and L. Schmidt-Thieme, "Scalable event-based clustering of social media via record linkage techniques," in *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2760

[28] J. Lingad, S. Karimi, and J. Yin, "Location extraction from disaster-related microblogs." in *WWW (Companion Volume)*. International World Wide Web Conferences Steering Committee / ACM, 2013, pp. 1017–1020.

[29] C. Li and A. Sun, "Fine-grained location extraction from tweets with temporal awareness," in *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, 2014, pp. 43–52. [Online]. Available: http://doi.acm.org/10.1145/2600428.2609582