Proceedings of the Ninth International AAAI Conference on Web and Social Media

SEEFT: Planned Social Event Discovery and Attribute Extraction by Fusing Twitter and Web Content

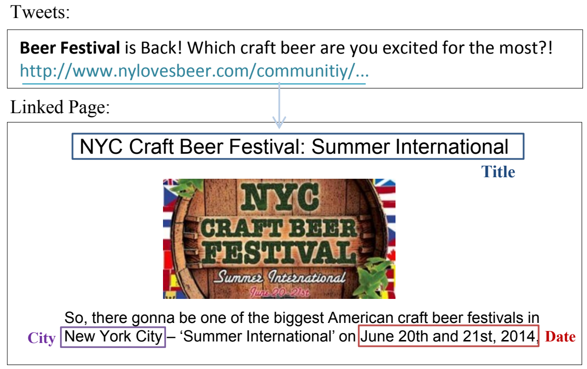
Yu Wang, David Fink, Eugene Agichtein
Emory University, Atlanta, GA, USA
yu.wang@emory.edu, djfink@emory.edu, eugene@mathcs.emory.edu

Abstract

Social events comprise some of the most popular topics in social media. Automatically identifying planned social events and extracting structured information, such as event title, date, and location, would enable more effective index, display and search for social events. However, the informal and noisy nature of language used in social media can degrade the quality of event extraction, resulting in broken titles, incorrect or absent attributes - making the resulting event databases not suitable for realistic applications. Previous work mostly focused on event identification and categorization in Twitter. Yet, event title extraction, arguably one of the most useful and difficult tasks in this domain, has never been investigated. In this paper, we address the task of identifying and extracting structured information (titles, dates, locations) for planned social events, and introduce SEEFT[1], a social event extraction system, which uses social media content to discover events. To extract the event title and other attributes, SEEFT fuses the original social media content and the content of other Tweets and webpages. Experiments over multiple popular event types and more than a thousand of event instances show that SEEFT significantly outperforms the previous state-of-the-art system in event identification. Moreover, by fusing information from multiple sources, SEEFT is able to extract event titles with high accuracy, providing the foundation for practical applications such as event discovery, search, and recommendation.

Introduction

Social events, such as concerts, sport games, and academic conferences, constitute major activities in our professional and personal life. As a subset of the general events traditionally addressed in Natural Language Processing research (e.g., some of the tasks defined in the Automatic Content Extraction (ACE) competition (Doddington et al. 2004)), *planned social events* are usually scheduled in advance, and encourage people to attend. Identifying and extracting social event information, and providing event search platforms has emerged as an active research and business area. Many websites, such as WikiCFP, EventBrite, MeetUp.com, and many local "fun things to do" websites, have been developed to collect social events, and help users search events of their interests. However, most of these websites require manual entry of the events in the system, which greatly limits the scale and scope of the event search engines.


Tweets:

Beer Festival is Back! Which craft beer are you excited for the most?! http://www.nylovesbeer.com/communitiy/...

Linked Page:

NYC Craft Beer Festival: Summer International

So, there gonna be one of the biggest American craft beer festivals in New York City – 'Summer International' on June 20th and 21st, 2014.

Figure 1: Example event tweet and linked page: The linked page contains formal title, date and city information.

The proliferation of social media, particularly the explosive popularity of Twitter, naturally attracts attention of social event organizers, attendees and commenters. Increasingly, event announcements, updates, and notices appear as microblog posts. However, due to the sheer volume of posts in popular microblogging services, it has become difficult for users to find relevant events or have their event-related postings noticed. To tackle this problem, automatic event identification and extraction of key event attributes, the "What", "Where", and "When" of the event, is an attractive solution. Yet, the typically informal and terse language of microblog posts in general, and event-related posts in particular, makes this task challenging. As illustrated in the example in Figure 1, the major problems with event extraction directly from microblog posts are: (1) the title of the event tends to be abbreviated or incomplete, make it often inadequate for users to search and comprehend the event; and (2) the typically short posts (e.g., Tweets) tend to omit crucial event-related information, such as dates and locations.

Fortunately, information about social events tends to also be available on "traditional" webpages. According to our

[1] SEEFT stands for "Social Event Attribute Extraction by Fusing Twitter and web content"

analysis on event-related Tweets, a substantial amount (35% - 60%) of them have embedded URLs, which link the brief microblog content with the more comprehensive information on a webpage. Unlike Tweets, these pages tend to be more readable, have rich contextual information, and contain unique structural meta data such as html tags, making the extraction of event information more accurate and more feasible than from the original microblog posts. In this light, we propose to extract event attributes, especially event titles, by fusing information from both microblog posts and linked webpages.

Previous work in this domain focused mostly on automatic event identification and categorization (Ritter et al. 2012) (Sakaki, Okazaki, and Matsuo 2010), which only tells if a Tweet contains an event, and possibly what kind. Although some work tried to recognize entities, time or venue of the events (Benson, Haghighi, and Barzilay 2011) (Popescu, Pennacchiotti, and Paranjpe 2011), none of these attributes can represent the event as well as the event title, which we attempt to extract. Event title extraction, arguably one of the most useful and difficult tasks in this domain, has never been investigated prior to this work, mainly because it is such a challenging problem.

In this paper, we develop SEEFT, a social event discovery and extraction system that Fuses information from microblog posts (Twitter) and external event-related webpages, to identify relevant events, and to automatically extract structured event attributes (titles, dates, locations). SEEFT takes potential event-related Tweets as input, and retrieves the webpages by following the embedded links in the Tweets. By fusing content from Tweets and linked webpages, SEEFT will produce event titles if one is identified. SEEFT extends the scope of information sources to larger relevant Tweet and Web collections by querying the microblog and web search engines with initial event titles. Finally, the system examines and fuses evidence from all of these sources, and outputs the structured event information (i.e., the "What", "When", and "Where" information critical for social events). Thus, events can be indexed in topical, geospatial, and temporal dimensions, which in turn could help build a content-aware search engine (Derczynski, Yang, and Jensen 2013). Compared to social event search websites which collect manually created event information, such as EventBrite, MeetUp and WikiCFP etc., the potential search platform powered by SEEFT would cover events with more diversity and on a larger scale.

We evaluate SEEFT on three major event types in Twitter, namely, concerts, conferences, and festivals. Experiments show that our fusion approach significantly outperforms state-of-the-art system in event identification, and improve the quality of event title extraction over the system that uses microblog content only. Note that the three event types used in our evaluation are very popular in Twitter which comprise thousands of events on a daily basis. Already in existing commercial event search applications, each of these event types has multiple dedicated websites, such as WikiCFP, Reverbnation, etc, where our system can have direct impact. To demonstrate the flexibility of our approach, we investigate cross-domain event title extraction, that is, applying an extractor trained on one type of events to another type. The results imply that certain event types are similar so that cross-domain extractor works well, e.g., conference and festival. Others are distinct in terms of the language style in event titles, making it difficult to transfer the extractor learned from one type to another.

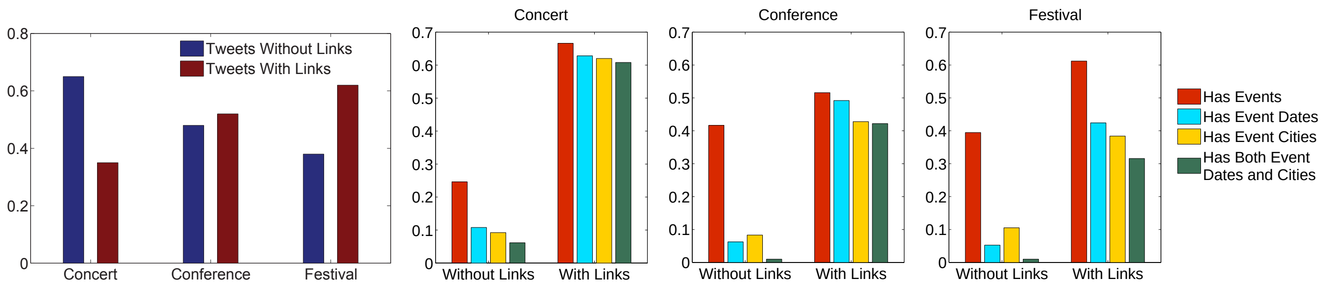
Related Work

The focus of this work is social event attribute extraction from Twitter and relevant web content. Related work falls into the following areas: (1) event identification and extraction in social media; (2) event extraction from the web and other domains; (3) cross-document information fusion.

There has been an growing interest in event identification in social media. Sakaki et al. (Sakaki, Okazaki, and Matsuo 2010) built a classifier to identify Tweets about earthquakes in Japan. Becker et al. (Becker, Naaman, and Gravano 2011) proposed a classifier to distinguish between event-related and non-event Tweet clusters. Ritter et al. (Ritter et al. 2012) implemented an event-phrase tagger to detect open domain events in Tweets. Parikh et al. (Parikh and Karlapalem 2013) built a popularity-based event detection system which relies on the bursty temporal pattern of event keywords in Tweets. Many other popularity based models (Weng and Lee 2011) (Li, Sun, and Datta 2012) were also proposed to identify events on Twitter. In contrast, SEEFT does not require the frequent mention of event phrases. Instead, it looks into the event webpages to gain additional signals to identify events. In addition, event identification has been explored for various event types, including music events (Benson, Haghighi, and Barzilay 2011), game events (van Oorschot, van Erp, and Dijkshoorn 2012), activist events (Ploeger et al. 2013), controversial events (Popescu and Pennacchiotti 2010), etc. Our work proposes a user-driven event extraction system which does not limit to a specific type of event.

Event attribute extraction in social media has been investigated to obtain event attribute information and facilitate other applications. Benson et al. (Benson, Haghighi, and Barzilay 2011) developed an extractor for music events to tag artists and venues. The open domain event extractor built by Ritter et al. (Ritter et al. 2012) extracts entities, dates and event phrases which could in turn help render events in a calendar fashion. Popescu et al. (Popescu, Pennacchiotti, and Paranjpe 2011) proposed a method to extract entities, actions and public opinion about the events from Twitter. Our work has a special focus on social events, which requires the ability of extracting event titles, dates and locations in order to make the result suitable for realistic applications, such as event search and recommendations. Event titles are usually considered as the identities of social events, whereas entities are generally insufficient to represent the events. Unlike entities extracted in previous work, SEEFT is designed to extract complete, accurate and human readable event titles.

General knowledge and information extraction (IE) from web content has been extensively investigated (Agichtein and Gravano 2000) (Etzioni et al. 2008). Sekine (Sekine 2006) proposed a user-driven system to extract information based on user-specified queries. Event detection and extraction, as a special task of IE, was actively explored in the



(a) Fraction of Tweets with links and without links.

(b) Fraction of events and event attributes extracted from Tweets and linked content by human labelers.

Figure 2: Statistics of embedded links in Tweets and Event Information Identified by Human Labelers: Fraction of Tweets with linked content (a); Fraction of event attributes in linked content (b).

web and other domains. Arrskog et al. (Arrskog et al. 2012) built a generic local event attribute extractor for web content. Chen and Roy (Chen and Roy 2009) built a spatial analysis model to detect events on Flickr and group event-related photos together. Many other information sources, such as emails and web search logs, were also proposed to identify events. Although our work relies on web content to extract event attributes, social media data plays a key role in collecting and analyzing event information.

Analyzing information from multiple sources and documents, such as social media, embedded webpages and search engine results, can provide redundant or additional evidence which could potentially boost the confidence and comprehensiveness of extraction. Mann and Yarowsky (Mann and Yarowsky 2005) fused the information extracted from multiple documents together by majority vote and produce more accurate results. The music event extractor developed by Benson et al. (Benson, Haghighi, and Barzilay 2011) also has the consensus-based idea which aggregates information from multiple Tweets in a graphical model. In this paper, we adapt the method of probabilistic voting to fuse the event attributes extracted from multiple sources.

Defining Social Events

Traditional event extraction tasks consider events to involve entities, actions, and objects in time and space. In this section, we first define the important category of events of interest – namely social events. We then characterize how social event information is disseminated through microblogging posts and the related external web content.

Problem Definition

As a subset of general events, planned social events have more explicit temporal and geospatial properties. For example, organizers of academic conferences and concerts usually announce and highlight where and when the event will be held on the event websites. Besides the time and location, social events also have a noun phrase title. Thus, we define social events of our interests as follows:

In the scope of this paper, we focus on the social events which (1) appear in social media, and (2) have webpages containing their formal title, date, and location.

(1) Social media presence: Social media has already become the expected channel for event organizers, attendees and commenters to communicate and disseminate information about social events. Most popular social event websites, such as WikiCFP and Reverbnation, have official Twitter accounts which automatically generate Tweets about their events, making Twitter cover almost all of the events on those websites. Although it is difficult to measure the coverage of social events on social media in general, we assume that the events without any presence on social media are very unpopular and less important.

(2) General web presence: It is very common for a social event to have its own webpage. Besides the websites hosted by the organizers themselves, many platforms, such as MeetUp, EventBrite and Facebook, allow users to create customized event pages.

The intersection of these two requirements yields a substantial amount of social events that can be feasibly extracted. Embedded links in Tweets naturally connect social media content with the external web domain. A pilot study conducted in this paper indicates that if a Tweet mentions real social events (manually annotated), more than 60% of concert Tweets have embedded external links. The percentages are even higher for conference Tweets (67%) and festival Tweets (78%). On the other hand, Tweets containing "event terms" (e.g., concert) are likely to have embedded links (as shown in Figure 2a).

The problem we address in this paper is: ***Given*** the Tweets potentially containing information about social events, with associated external content, ***Extract*** the structured event information including event *title*, the (starting) *date*, and *location*, if there is any.

Note that we attempt to build an event attribute extractor which produce results only if an event is identified. In other words, our extractor can be used for ***event identification*** by default.

Advantages of Using Twitter for Event Discovery

- Microblogging platforms offer a great opportunity to collect ***fresh*** and ***diverse*** event content.
- Social media posts (e.g. Tweets) are relatively focused due to its restricted length, which may help ***interpret*** and

disambiguate event web content.

- Most microblog posts come with rich meta information (author's social network and self-reported profile and location etc.), which can be used to determine the ***popularity*** and ***audience*** of events.

Furthermore, it is challenging to collect social event information (e.g., event webpages) without using social media. As far as we know, there is no dedicated repository or index for general event webpages. Existing indices have either low recall and a very restricted domain of events (WikiCFP, MeetUp), or include many non-event pages (i.e., general-purpose search engines). While issuing a generic event query (e.g., "Conference") to a general Web search engine may retrieve some event pages, however, the ranking of the search results is usually stable and favors very popular events, making it difficult to discover newly created or local events through general search engines.

Advantages of Fusing External Web Content

Compared to event-related Tweets, an event webpage usually contains more ***complete***, ***well-formed*** and ***reliable*** event information. To illustrate how much additional information the event webpages can provide, we collect potential event-related Tweets by issuing three queries to Twitter: *Concert*, *Conference* and *Festival*. After obtaining 100 Tweets for each query, we ask labelers to identify if there are any events in the Tweets. If yes, the labelers are then asked to extract titles, dates and cities for the identified events. During the event identification and attribute extraction, the labelers can look at the Tweet and any linked pages in the Tweet. Finally, we split the Tweets into two groups: Tweets with no embedded links and Tweets with links. The results are visualized in Figure 2.

Figure 2 shows several interesting phenomena: (1) Figure 2a shows a substantial amount (35% - 60%) of Tweets contain links to external webpages. (2) Figure 2b indicates that it is more likely for the Tweets with links to contain an event. (3) Finally, linked pages in the Tweets provide much more event information (event dates and cities) than the Tweets themselves.

Event Title Extraction and Fusion

Our system starts by extracting event titles for two reasons: (1) Event titles, rather than dates and cities, are more likely to appear in the Tweets. (2) knowing where the title is on the webpage could help more accurately identify event dates and cities.

Sentence-Level CRF-Based Event Title Extractor

The task of identifying event titles from Tweets and webpages content is considered as a sequence tagging problem, and a Linear Chain Conditional Random Field (CRF) model is employed to learn the inter-dependencies between words in the title. The inputs to the tagging process are sentences, where a Tweet is considered as a sentence, and webpages are broken down into sentences according to HTML DOM structure and punctuation. Each token in a sentence is tagged with either *Title* or *Non-title*. Traditional Begin-Inside-Outside tags performed at slightly improved precision, but worse recall as compared to the binary tags used in this paper, and therefore not utilized.

Feature Name	*Description*
Features for Title Extraction	
Query Dependent	i.e., Matches query issued to Twitter
Tweet Dependent	i.e., Token is contained in the Tweet
Token	Token itself (as is & lowercase)
Part-of-Speech	POS tags from Stanford Parser[2]
Capitalization	All-caps/Some-caps/First-letter-cap
Number	Token Contains a Number
Stopword	Is a stopword
Punctuation	Is a punctuation mark
Parenthesized	Is contained between parenthesis
Entity	Entity from Stanford NER[3]
Position	Token Wise Position in Sentence
Phrase-Structure-Tree	i.e., Beginning of noun phrase
Additional Features for City and Date Extraction	
Type	City/Date
Position	Sentence wise position relative to title occurrences
Temporal	Past/Today/Future (Date only)

Table 1: Features for Event Attribute Extraction.

The extractor takes a sentence at a time as input, and output the most appropriate event titles ranked by their likelihood scores (results of CRF), if there is any. Annotated event titles are required to train the CRF extractor. Table 1 lists the features used by the CRF extractor during training and runtime.

Extracting Event Title by Fusing Tweets and Embedded Links

Figure 3 illustrates the process of extracting event titles from Tweets and webpages.

Given a sentence, the CRF-based event title extractor provides proper titles with corresponding likelihood scores. However, not all sentences should be treated equally. First, the Tweet usually contains highly abstracted event information, which makes it more valuable than a random sentence on the webpage. Second, many sentences on the webpage could be irrelevant to the event (e.g., ads). In practice, if we dump all sentences on the webpage to CRF extractor, it produces random strings (written in the form of event titles) with high likelihood scores, which could easily mess up the final output. Thus, we use the Tweet to filter sentences on the webpage. Specifically, only top-K closest (with regard to string similarity) sentences to the Tweet are examined by the CRF extractor. The similarity algorithm implemented is based on the work by Turpin et al. (Turpin et al. 2007) on locating relevant information or snippets on a webpage based on a query. The benefits of such filtering are two-folds: (1) it filters out most of the irrelevant content, making the title extraction more robust. (2) it drastically improves the efficiency of title extraction since the CRF extractor only deals with the smaller number of sentences. For experiments in

[2] http://nlp.stanford.edu/software/lex-parser.shtml

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

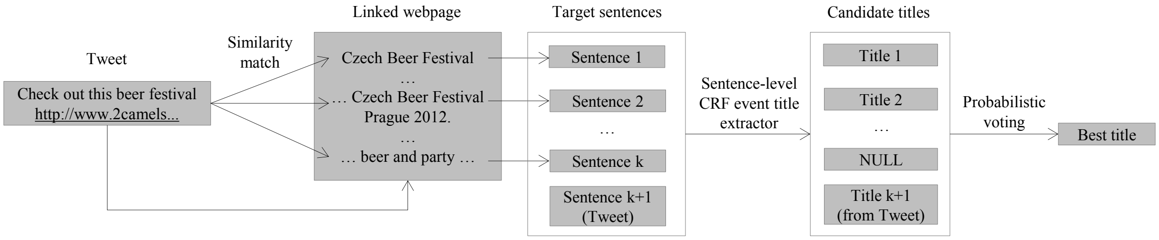


Figure 3: Event title extraction and fusion from Tweets and linked pages.

which all sentences were processed by the extractor, F-Score for title extraction dropped by 4.2 %.

Event title could appear multiple times on the webpage and also in the Tweet. Thus, the CRF extractor usually produces more than one candidate event titles. We then use probabilistic voting to select the "best" title. The weights of candidate titles are from results of CRF extractor. The voting formula is as follows:

$$\hat{T} = \max_{T} \sum_{C} P_{C,T}$$

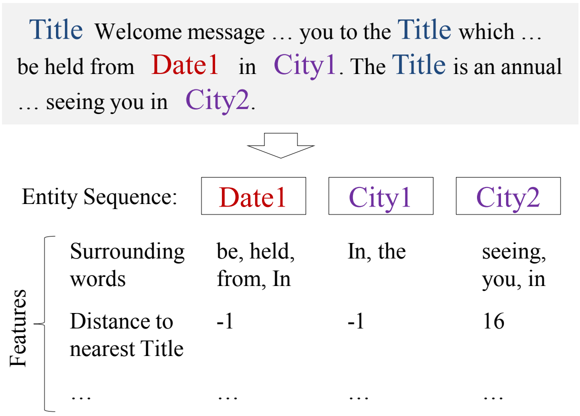
T stands for candidate titles and $\hat{T}$ is the "best" title elected. C is a candidate sentence and $P_{C,T}$ is the likelihood score of title T in sentence C. When all candidate titles are unique strings, this formula simply picks the one with the highest CRF likelihood score. When a candidate title appears more than once, this formula favors titles that appear more frequently, which improves the robustness of the result. This approach is very effective to avoid "long" titles (which includes more words than a proper title, sometimes the whole sentence is tagged as the title) extracted by CRF. On the other hand, when CRF produces very "short" titles (which is not enough to represent the event), they are mostly stopwords. Our system requires at least one word in the extracted titles being non-stopwords.

Event Date and City Extraction and Fusion

After the "best" event title is identified, we move on to extract event attributes. Usually, important event information is often highlighted on the event webpage, and they tend to be close to each other. In general, dates and cities on webpages are relatively easy to be recognized. However, most webpages contain a multitude of dates and cities, the majority of which are not useful for our purposes. For example, a news article confirming a speaker for a conference may contain the publication date of the article, an address for the publisher, the home city of the speaker, a comment section containing noisy information, as well as numerous links to other articles that may contain dates and cities in their headlines. We address this issue by fusing additional information from multiple sources.

Page-Level CRF-Based Event Attribute Extractor

Given a single webpage, we first tag a set of all possible candidate dates and cities by running a temporal resolution tool and a dictionary-based city identifier on every sentence. The candidate dates and cities, as well as each occurrence of the extracted titles on the page, are then grouped into a list ordered by their position of occurrence in the webpage HTML.


Title Welcome message … you to the Title which … be held from Date1 in City1. The Title is an annual … seeing you in City2.
Entity Sequence:
Features

Entity Sequence:	Date1	City1	City2
Surrounding words	be, held, from, In	In, the	seeing, you, in
Distance to nearest Title	-1	-1	16
…	…	…	…

Figure 4: Constructing Entity Sequence for Date and City Extraction.

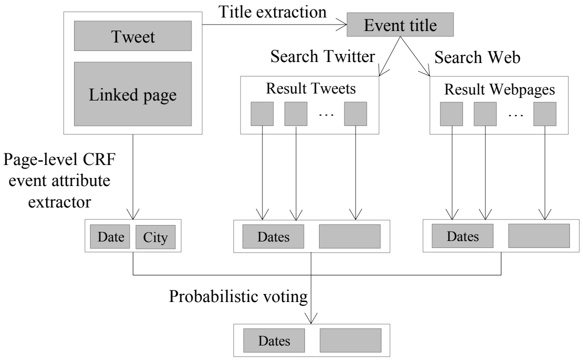


Figure 5: Fusing Search Results for Date and City.

The problem is then reduced again to sequence tagging and a Linear Chain CRF model is again employed. The tags used are *Correct City*, *Correct Date* and *Incorrect Attribute*. We expect this model to work because of similarities found in webpage structures. Intuitively, we expect title, date and city to appear together on a webpage, likely with date and city appearing after the title. An example of this ideal page structure is shown in Figure 4. A list of these features for date and city extraction is shown in Table 1 in the *Features for City and Date Extraction* section.

Fusing from Additional Search Results

Applying page-level CRF event attribute extractor on the linked page may tag no dates and cities at all. One reason is that linked pages sometimes do not contain those information. To obtain more comprehensive and robust event attributes, SEEFT retrieves additional relevant Tweets and webpages by querying Twitter and general web search engines with the extracted event title (as shown in Figure 5). The system then fuses these information sources via probabilistic voting. The voting process is implemented as follows. For each resource obtained from a search result (a Tweet or a webpage), the same extraction process is performed and the proposed event attributes are saved. SEEFT chooses the best estimated attribute value $\hat{A}$ according to the following formula:

$$\hat{A} = \max_{A} \sum_{S} b_s S_A$$

where A is the candidate attribute value. If the attribute is an event city, then A can be "New York" or "Los Angeles". S is the information source. In our case, it can be Twitter or web search results. S_A is the vote for value A from source S. Finally, b_S indicates the weight or the credibility of source S, which can be tuned to fit the data.

The voting score S_A is computed as follows:

$$S_A = \frac{\sum_i w_i S_{Ai}}{\sum_{A'} \sum_i w_i S_{A'i}}$$

where S_{Ai} is a binary value indicating if the i-th result from S proposes attribute value A. w_i is a weight assigned to the i-th result from S. For general web search engine, the position of the results usually implies relevance. Intuitively, the higher ranked results should gain more credibility. In the case of Twitter, results are ranking by timestamp. Thus the weights are uniform in our implementation. Finally, the voting score is normalized by the sum of scores for all proposed values.

As stated above, this voting process is only performed if the first round of extraction does not produce a date or a city. Experiments show that the embedded link in the Tweet is the most reliable information source of the event. This link is provided by a human (the Twitter user) and is therefore more likely to be trustworthy than search results generated using the name found by a machine (the extractor).

Experimental Setup

Two types of experiments are performed: extracting event attributes (1) by fusing Tweets and embedded links, and (2) by fusing additional search results from Twitter and web.

Dataset

In order to evaluate the proposed event extraction system, we collect event Tweets (by querying Twitter API with event type names) of 3 event types, namely concerts, conferences, and festivals. Concerts are very social-oriented events and the language in Tweets and webpages is less formal. "Conference" is an ambiguous query which could represent both academic conferences and sport conferences. Festivals, interestingly, usually have similar title forms to conferences, making them look formal. After collecting Tweets for one day, we sampled 500 potential event-related Tweets (the ones containing event type names) from each event type to label.

	Concert	*Conference*	*Festival*
Tweets with links	10,450	8923	11,805
Sampled Tweets	500	500	500
Events in sampled Tweets	333	322	306
Future events in sampled Tweets	285	177	159
Past or current events in sampled Tweets	48	145	147
Events with dates in sampled Tweets	314	274	212
Events with cities in sampled Tweets	310	247	192
Events with cities and dates in sampled Tweets	304	232	158

Table 2: Dataset: Statistics of Events Identified by Human Labelers.

Labeling

We ask human labelers to identify events and extract event attributes. Each time, the labeler is presented with a event type, a Tweet and the corresponding linked webpage. The following questions are asked: (1) Is there any event identified in the Tweet or the webpage? (2) If yes, extract the most proper title, city and starting date from either the Tweet or the webpage.

The guidelines for picking a proper title are: (1) The title should be accurate and contain no irrelevant words. For instance, if the event is "Social Media Conference", then "Announce Social Media" is not proper. (2) The title should be comprehensive. For example, if an event is "Art and Music Festival", neither "Art Festival" nor "Music Festival" is proper. (3) The quality of the title should be suitable for other applications, such as event search. In the previous example, it is not proper to extract the name as "#ArtandMusic Festival" (which is more likely to occur in Tweets), if there is a better choice.

Table 2 summarizes the data we collected (on 9/12/2013) and labeled. There are 961 events identified by human labelers and 621 of them are future events. Proportionally, there would be about 20,000 events and 13,000 future events in the whole collection of data gathered on a single day. Note that some popular events could appear multiple times in the dataset, which is not addressed in this paper.

Methods Compared

TwiCal (Ritter et al. 2012): the baseline system which is designed to tag event phrases, entities and times from Tweets. It is not fair to compete against TwiCal in the task of title extraction because neither event phrases nor event entities can be considered as event titles. Therefore, we evaluate TwiCal and SEEFT in event identification: to classify Tweets into event-related and non-event-related classes. In

		TwiCal *Tweet only* *(Baseline)*	SEEFT-T *Tweet only*	SEEFT-TW *(Our system)*
Concert	Prec.	0.674	0.738 (+10%)	**0.804** **(+19%)**
	Rec.	**0.987**	0.559 (-43%)	0.850 (-14%)
	F1	0.801	0.636 (-21%)	**0.826** **(+3%)**
Conference	Prec.	0.617	0.730 (+18%)	**0.873** **(+42%)**
	Rec.	0.689	**0.882** **(+28%)**	0.879 (+28%)
	F1	0.651	0.799 (+23%)	**0.876** **(+35%)**
Festival	Prec.	0.617	0.628 (+2%)	**0.700** **(+14%)**
	Rec.	0.683	**0.915** **(+34%)**	0.853 (+25%)
	F1	0.648	0.745 (+15%)	**0.769** **(+19%)**

Table 3: Event identification: TwiCal vs. SEEFT-T vs. SEEFT-TW. Relative improvement is computed against TwiCal.

TwiCal, if any event phrase is tagged in a Tweet, we consider that an event is identified in the Tweet. In contrast, our system, SEEFT, produces an event title if any event is identified.

SEEFT-T (*Tweet only*): the system which identifies and extracts event information from Tweets only. To fit the language style of Tweets and provide a fair comparison, the CRFs of this system are trained with Twitter content.

SEEFT-TW (*Tweet + Linked Pages*): the system which makes use of both Tweets and Linked Pages to identify events and extract event attributes.

SEEFT-TWS (*Tweet + Linked Pages + Twitter Search Results (TSR) and/or Web Search Results (WSR)*): the system which extracts event titles first, and gets additional event pages by searching the extracted title on search engine. Search results are used to obtain more comprehensive attribute information.

Fusing Linked Pages to Extract Event Attributes

We first compare SEEFT-T, SEEFT-TW and TwiCal (Section) in event identification, and then test SEEFT-TW against SEEFT-T on event attribute extraction. The experiments are carried on three different event types independently. All numbers from SEEFT are computed based on a 10-fold cross-validation setting.

Event Identification

Event identification is considered to be a relatively easier task than event attribute extraction. Again, TwiCal identifies events by tagging event phrases; SEEFT produces event titles if any event is identified.

		SEEFT-T	SEEFT-TW
Concert	Prec.	0.615	**0.651 (+6%)**
	Rec.	0.465	**0.688 (+48%)**
	F1	0.530	**0.669 (+26%)**
Conference	Prec.	0.332	**0.796 (+140%)**
	Rec.	0.401	**0.801 (+100%)**
	F1	0.363	**0.799 (+120%)**
Festival	Prec.	0.410	**0.606 (+48%)**
	Rec.	0.598	**0.739 (+24%)**
	F1	0.487	**0.666 (+37%)**

Table 4: Event Title Extraction Results.

Table 3 shows that our system outperforms both TwiCal and SEEFT-T in all three event types by F1 measure. Interestingly, TwiCal obtains good recall in identifying concert event. The reason is that TwiCal was trained on a collection of event-related Tweets, and concerts consist of the a big portion of that. It turns out the term *concert* is an event phrase in TwiCal, so that it recognizes almost every Tweet containing the word *concert* as an event. When the event type is less popular, TwiCal gives lower recall and precision. This observation implicitly indicates that TwiCal has a biased performance in "open domain" events.

SEEFT-TW provides better precision than Tweet only systems (TwiCal and SEEFT-T) in all three types of events. One reason is that Tweets could be ambiguous due to the restricted length. Using linked pages could help better interpret and disambiguate the Tweet content. SEEFT-TW improves the precision by a large margin, especially when the event query is ambiguous, e.g. conference.

Event Attribute Extraction

Event titles, starting dates, and cities are extracted by the baseline system (SEEFT-T) and SEEFT-TW. Note that extracted event titles are judged by human labelers according to whether they are precise and comprehensive enough to represent the event, with regard to index and search purposes.

TwiCal is incapable of extracting event attributes. It tends to tag "event phrases", which cannot be directly used as titles. Without proper titles extracted, it is less meaningful to attempt other attribute extraction.

Event Title Extraction Table 4 reports the performance of event title extraction. By incorporating linked pages, SEEFT outperforms the baseline system on all measures, with recall improvements ranging from 24% to 100% and F1 improvements ranging from 26% to 37%. These results demonstrate that linked pages provide extra title information which could be missing in the Tweets. Our analysis on extracted titles indicates that external event webpages could help when the event title is either missing or very brief in the informally written Tweet (as in the example of Figure 1).

The extracted event titles can be different from the ideal titles in many ways. For example, it can be a superstring (containing the ideal title), a substring (contained in the ideal title), overlap (partial match) with the ideal title, or does not overlap at all. We report the proportion of each string types and the success rate of title extraction accordingly in Table 5.

Table 5 shows that the majority of extracted titles exactly match the ideal titles. In the category of "overlap", the success rate is over 70%. Surprisingly, our system still produces

	Exact Match	*Super-string*	*Sub-string*	*Overlap*	*No Overlap*
Number extracted	481 (58%)	31 (4%)	82 (10%)	155 (19%)	78 (9%)
Percentage labeled correct	100%	81%	81%	70%	40%

Table 5: Title extraction of SEEFT-TW by error type.

		SEEFT-T	SEEFT-TW
Concert	Prec.	**0.917**	0.874 (-5%)
	Rec.	0.355	**0.490 (+38%)**
	F1	0.512	**0.628 (+23%)**
Conference	Prec.	0.758	**0.779 (+3%)**
	Rec.	0.202	**0.470 (+133%)**
	F1	0.319	**0.586 (+84%)**
Festival	Prec.	**0.765**	0.730 (-5%)
	Rec.	0.323	**0.422 (+31%)**
	F1	0.454	**0.535 (+18%)**

Table 6: Event location (city) extraction results.

about 40% acceptable titles even they do not overlap with the ideal titles at all. The reason is that an event could have full titles and brief titles (initial letter of the words in full titles), and sometimes they do not overlap with each other. Also, more than one event can be co-organized together, which happens more often for concerts and festivals.

Event City and Date Extraction Tables 6 and 7 report the extraction results of event cities and starting dates. Many events in our dataset, especially concerts, are one-day events. Ending dates of events are also important for events spanning more than one day. However, our dataset does not contain sufficient multi-day events to train the CRF to pick up ending dates in the cross-validation setting.

As shown in Tables 6 and 7, SEEFT-TW gains a higher F1 score in all three event types and in both city and date extraction. One interesting finding is that SEEFT-T tends to have better precision than SEEFT-TW. The reason is that Tweet content is more focused than webpages, so cities and dates in Tweets are more likely to be the correct ones. On the other hand, webpages usually contain more than one city and date which brings the challenge of selecting the correct ones. However, by using the Tweet content to help interpret the webpages, SEEFT-TW, in some event types, achieves similar or even higher precision with a much higher recall.

Surprisingly, both systems give very low recall in festival starting date extraction. During the labeling, we found it is sometimes difficult to recognize event dates for festivals from either Tweets or webpages because they are on a picture (flyer of events) or written in an implicit way, such as "this weekend" or "next month", which results in fewer dates labeled for festivals. This, in turn, hurts the quality of the CRF due to the lack of training data. In Section , we show how information from additional sources helps improve the recall of attribute extraction.

Cross-Domain Event Title Extraction

To take a step forward towards building an open domain event attribute extraction system, we investigate how the title extractor trained on a certain type of events performs on other types of events. As a comparison, we also show the performance of an in-domain extractor and the extractor trained on the mix of all three types of events, using 10-fold cross-validation.

		SEEFT-T	SEEFT-TW
Concert	Prec.	0.940	**0.963 (+2%)**
	Rec.	0.398	**0.583 (+47%)**
	F1	0.559	**0.726 (+30%)**
Conference	Prec.	**0.939**	0.926 (-1%)
	Rec.	0.168	**0.409 (+144%)**
	F1	0.285	**0.567 (+99%)**
Festival	Prec.	**0.800**	0.655 (-18%)
	Rec.	0.075	**0.090 (+20%)**
	F1	0.138	**0.158 (+15%)**

Table 7: Event (starting) date extraction results.

		Extractor Trained on			
		concert	*conference*	*festival*	*all*
Concert	Prec.	0.651	0.396	0.394	**0.705**
	Rec.	**0.688**	0.114	0.129	0.661
	F1	0.669	0.177	0.195	**0.682**
Conference	Prec.	0.386	**0.796**	0.686	0.740
	Rec.	0.373	**0.801**	0.596	0.786
	F1	0.379	**0.799**	0.638	0.762
Festival	Prec.	0.477	**0.637**	0.606	0.599
	Rec.	0.650	0.690	0.739	**0.771**
	F1	0.550	0.662	0.666	**0.674**

Table 8: Cross-domain title extraction results.

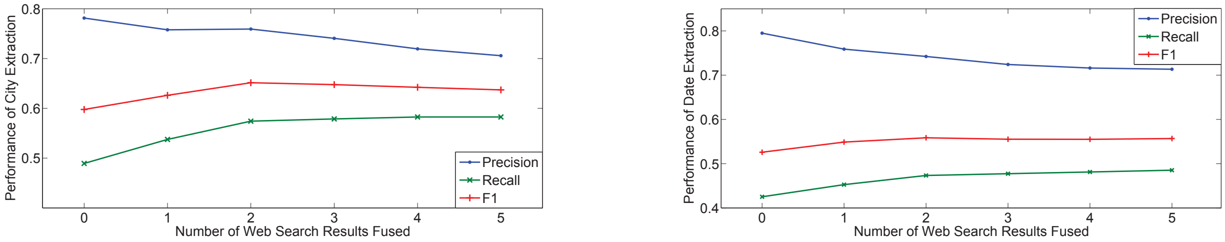
Table 8 shows the results of cross-domain event title extraction. We can see that the extractor trained on *conference* works well on *festival*. During the labeling, we found that the naming convention and the structure of titles are similar between conferences and festivals. However, *conference* and *concert* are not compatible, in part because of the distinct characteristics of the two types of events. Concert titles are more free style, and in many cases the title is be the name of a band. For *concert* and *festival*, the extractor trained on all events performs the best. One reason is that social-oriented events, such as concerts and festivals, have various title forms and learning from diverse event types can help recognize them better.

We also evaluate city and date extraction in the cross-domain fashion. The results look very similar to cross-domain title extraction. That is, the extractor trained on all events produces the best results; festival and conference extractors work well on each other, but the concert event require its own extractor to have reasonable performance.

Fusing Search Results to Extract Event Attributes

When the linked pages contain no date and city information, SEEFT-TW is often unable to identify these attributes. We extend SEEFT to search Twitter and general web with extracted titles and fuse the results.

By using search engines, the scope of information sources is now extended to a larger Tweet and webpage collection. SEEFT-TWS may find and propose cities and dates which are not identified in the previously annotated data because labelers only looked at the original Tweets and embedded



(a) City extraction with varying number of web search results incorporated.

(b) Date extraction with varying number of web search results incorporated.

Figure 6: Event attribute extraction performance by fusing different numbers of web search results.

		SEEFT -TW	SEEFT -TWS *with WSR*	SEEFT -TWS *with TSR*	SEEFT -TWS *with WSR* *and TSR*
City	Prec.	**0.781**	0.755	0.663	0.746 (-4%)
	Rec.	0.489	0.570	**0.630**	0.583 (+19%)
	F1	0.598	0.647	0.646	**0.648** **(+8%)**
Date	Prec.	**0.795**	0.749	0.486	0.749 (-6%)
	Rec.	0.425	**0.477**	0.433	**0.477** **(+12%)**
	F1	0.526	**0.563**	0.457	**0.563** **(+7%)**

Table 9: Comparing system extraction results when fusing search results. TSR is Twitter Search Results, and WSR is Web Search Results (top 2 results).

links. Thus, we randomly selected 300 identified events (100 for each type) and ask labelers to find attributes via search engines to complete the event profiles.

The attribute value candidates extracted from the different sources are fused using the voting mechanism described previously. The credibility scores (b_S) of TSR and WSR are tuned by optimizing F1 measure, and results are computed based on a 5-fold cross validation.

Table 9 shows the improvements over the baseline when additional search results are incorporated. Overall, incorporating search results boosts the recall, but the precision decreases since the extracted event title may retrieve irrelevant search results. Web search results consistently provide better extraction results for cities and dates. However, Twitter search results only help city extraction. This finding implies that when users compose event Tweets, they usually do not include event dates. To see the ceiling of the fusion approach, we also search human labeled titles to get relevant webpages. It seems that search results of extracted titles achieve very similar performance as human labeled titles, which implies the extracted titles are as good as human labeled ones in retrieving relevant event pages.

When tuning the weights or credibility scores for Twitter and web search results, the optimal average weights for city extraction turn out to be 0.58 for web search results and 0.42 for Twitter search results. On the other hand, the optimal weights for date extraction completely go to web search results, giving Twitter search results a 0 credibility score. The weights imply that Twitter is a relatively valuable source for identifying event cities, but not for event dates.

Web search results are ranked by relevance in general. Thus, incorporating more results from web search engine may bring the risk of incorporating irrelevant content. Figure 6a and 6b show the performance of city and date extraction with varying number of top web search results examined. Results show that fusing top 2 web results gives the optimal F1 score. Incorporating more than 2 web results quickly degrades the precision, which in turn hurts the F1 score.

We also experimented with using the fusion model for title extraction. Intuitively, search results, especially web search results, could contain better or at least the same titles as the query title. In practice, however, fusing search results most of the time has no effect, and occasionally degrades quality of extracted titles. The main reason is that results retrieved by the query title usually contain the query title itself, which will be re-extracted by the CRF model.

Discussion and Conclusions

The structured event information, especially event titles, produced by SEEFT could enable and facilitate more sophisticated event indexing and search. Moreover, SEEFT connects events with social media content. The users in social media who post or promote events could potentially provide valuable info to characterize, rank, and estimate the impact of those events. Recommendation systems can also leverage users' social networks to target the audience with proper events. With event titles identified, we could track event-related microblog posts, and organize and display them in conjunction with events.

To conclude, we have introduced the task of event attribute extraction from social media and web content and developed SEEFT, an extraction system which fuses information from microblog posts (Twitter) and external event-related webpages, to identify relevant events, and to automatically extract structured event attributes. Fusing information from multiple sources (especially when the sources, i.e., Tweets and Webpages, are so different) is challenging. We designed the fusion process to leverage the low ambiguity of Tweet content and the comprehensiveness of Webpages, where they naturally complement each other, to reliably identify events and comprehensively extract event attributes. To the best of our knowledge, our work is the first attempt to discover and structure social event information by fusing content from social media and the web. Experiments show that SEEFT outperforms the state-of-the-art event identification system by nearly 20%. By fusing social media and web content, our system improves event title extraction by 60% on average (F1 measure). Moreover, fusing search results from social media and general web search engines gains another 7% on attribute extraction (dates and cities). All these improvements make SEEFT a valuable tool to produce reliable structured event information, which could facilitate search for social event and aid users in exploring and discovering social events on a larger scale.

Acknowledgments

This work was supported by the DARPA grant D11AP00269.

References

Agichtein, E., and Gravano, L. 2000. Snowball: extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, 85–94. New York, NY, USA: ACM.

Arrskog, T.; Exner, P.; Jonsson, H.; Norlander, P.; and Nugues, P. 2012. Hyperlocal event extraction of future events. In *Proceedings of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web*, volume 902.

Becker, H.; Naaman, M.; and Gravano, L. 2011. Beyond trending topics: Real-world event identification on twitter. In *ICWSM*.

Benson, E.; Haghighi, A.; and Barzilay, R. 2011. Event discovery in social media feeds. In *HLT*.

Chen, L., and Roy, A. 2009. Event detection from flickr data through wavelet-based spatial analysis. In *CIKM*, 523–532. New York, NY, USA: ACM.

Derczynski, L. R. A.; Yang, B.; and Jensen, C. S. 2013. Towards context-aware search and analysis on social media data. In *Proceedings of the 16th International Conference on Extending Database Technology*, 137–142.

Doddington, G. R.; Mitchell, A.; Przybocki, M. A.; Ramshaw, L. A.; Strassel, S.; and Weischedel, R. M. 2004. The automatic content extraction (ace) program - tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.

Etzioni, O.; Banko, M.; Soderland, S.; and Weld, D. S. 2008. Open information extraction from the web. *Commun. ACM* 51(12):68–74.

Li, C.; Sun, A.; and Datta, A. 2012. Twevent: segment-based event detection from tweets. In *CIKM*, 155–164.

Mann, G. S., and Yarowsky, D. 2005. Multi-field information extraction and cross-document fusion. In *ACL*, 483–490.

Parikh, R., and Karlapalem, K. 2013. Et: events from tweets. In *WWW Companion*, 613–620.

Ploeger, T.; Kruijt, M.; Aroyo, L.; de Bakker, F.; Hellsten, I.; Fokkens, A.; Hoeksema, J.; and ter Braake, S. 2013. Extractivism: Extracting activist events from news articles using existing nlp tools and services. In *Proceedings of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web*.

Popescu, A.-M., and Pennacchiotti, M. 2010. Detecting controversial events from twitter. In *CIKM*, 1873–1876.

Popescu, A.-M.; Pennacchiotti, M.; and Paranjpe, D. 2011. Extracting events and event descriptions from twitter. In *WWW*, 105–106.

Ritter, A.; Mausam; Etzioni, O.; and Clark, S. 2012. Open domain event extraction from twitter. In *KDD*, 1104–1112.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, 851–860.

Sekine, S. 2006. On-demand information extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*, 731–738. Stroudsburg, PA, USA: Association for Computational Linguistics.

Turpin, A.; Tsegay, Y.; Hawking, D.; and Williams, H. E. 2007. Fast generation of result snippets in web search. In *SIGIR*, 127–134.

van Oorschot, G.; van Erp, M.; and Dijkshoorn, C. 2012. Automatic extraction of soccer game events from twitter. In *Proceedings of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web*, volume 902, 21–30.

Weng, J., and Lee, B.-S. 2011. Event detection in twitter. In *ICWSM*.