Open Domain Event Extraction from Twitter

Alan Ritter
University of Washington
Computer Sci. & Eng.
Seattle, WA
aritter@cs.washington.edu

Mausam
University of Washington
Computer Sci. & Eng.
Seattle, WA
mausam@cs.washington.edu

Oren Etzioni
University of Washington
Computer Sci. & Eng.
Seattle, WA
etzioni@cs.washington.edu

Sam Clark*
Decide, Inc.
Seattle, WA
sclark.uw@gmail.com

ABSTRACT

Tweets are the most up-to-date and inclusive stream of information and commentary on current events, but they are also fragmented and noisy, motivating the need for systems that can extract, aggregate and categorize important events. Previous work on extracting structured representations of events has focused largely on newswire text; Twitter's unique characteristics present new challenges and opportunities for open-domain event extraction. This paper describes TwiCal—the first open-domain event-extraction and categorization system for Twitter. We demonstrate that accurately extracting an open-domain calendar of significant events from Twitter is indeed feasible. In addition, we present a novel approach for discovering important event categories and classifying extracted events based on latent variable models. By leveraging large volumes of unlabeled data, our approach achieves a 14% increase in maximum F1 over a supervised baseline. A continuously updating demonstration of our system can be viewed at `http://statuscalendar.com`; Our NLP tools are available at `http://github.com/aritter/twitter_nlp`.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Language parsing and understanding; H.2.8 [**Database Management**]: Database applications—*data mining*

General Terms

Algorithms, Experimentation

1. INTRODUCTION

Social networking sites such as Facebook and Twitter present the most up-to-date information and buzz about current events. Yet the number of tweets posted daily has recently exceeded two-hundred million, many of which are either redundant [57], or of limited interest, leading to information overload.[1] Clearly, we can benefit from more structured representations of events that are *synthesized* from individual tweets.

Entity	Event Phrase	Date	Type
Steve Jobs	died	10/6/11	Death
iPhone	announcement	10/4/11	ProductLaunch
GOP	debate	9/7/11	PoliticalEvent
Amanda Knox	verdict	10/3/11	Trial

Table 1: Examples of events extracted by TwiCal.

Previous work in event extraction [21, 1, 54, 18, 43, 11, 7] has focused largely on news articles, as historically this genre of text has been the best source of information on current events. In the meantime, social networking sites such as Facebook and Twitter have become an important complementary source of such information. While status messages contain a wealth of useful information, they are very disorganized motivating the need for automatic extraction, aggregation and categorization. Although there has been much interest in tracking trends or memes in social media [26, 29], little work has addressed the challenges arising from extracting structured representations of events from short or informal texts.

Extracting useful structured representations of events from this disorganized corpus of noisy text is a challenging problem. On the other hand, individual tweets are short and self-contained and are therefore not composed of complex discourse structure as is the case for texts containing narratives. In this paper we demonstrate that **open-domain event extraction** from Twitter is indeed feasible, for example our highest-confidence extracted future events are 90% accurate as demonstrated in §8.

Twitter has several characteristics which present unique *challenges* and *opportunities* for the task of open-domain event extraction.

Challenges: Twitter users frequently mention mundane events in their daily lives (such as what they ate for lunch) which are only of interest to their immediate social network. In contrast, if an event is mentioned in newswire text, it

*This work was conducted at the University of Washington

KDD'12, August 12–16, 2012, Beijing, China.

[1] `http://blog.twitter.com/2011/06/200-million-tweets-per-day.html`

is safe to assume it is of general importance. Individual tweets are also very terse, often lacking sufficient context to categorize them into topics of interest (e.g. SPORTS, POLITICS, PRODUCTRELEASE etc...). Further because Twitter users can talk about whatever they choose, it is unclear in advance which set of event types are appropriate. Finally, tweets are written in an informal style causing NLP tools designed for edited texts to perform extremely poorly.

Opportunities: The short and self-contained nature of tweets means they have very simple discourse and pragmatic structure, issues which still challenge state-of-the-art NLP systems. For example in newswire, complex reasoning about relations between events (e.g. *before* and *after*) is often required to accurately relate events to temporal expressions [32, 8]. The volume of Tweets is also much larger than the volume of news articles, so redundancy of information can be exploited more easily.

To address Twitter's noisy style, we follow recent work on NLP in noisy text [46, 31, 19], annotating a corpus of Tweets with events, which is then used as training data for sequence-labeling models to identify event mentions in millions of messages.

Because of the terse, sometimes mundane, but highly redundant nature of tweets, we were motivated to focus on extracting an aggregate representation of events which provides additional context for tasks such as event categorization, and also filters out mundane events by exploiting redundancy of information. We propose identifying important events as those whose mentions are strongly associated with references to a unique date as opposed to dates which are evenly distributed across the calendar.

Twitter users discuss a wide variety of topics, making it unclear in advance what set of event types are appropriate for categorization. To address the diversity of events discussed on Twitter, we introduce a novel approach to discovering important event types and categorizing aggregate events within a new domain.

Supervised or semi-supervised approaches to event categorization would require first designing annotation guidelines (including selecting an appropriate set of types to annotate), then annotating a large corpus of events found in Twitter. This approach has several drawbacks, as it is apriori unclear what set of types should be annotated; a large amount of effort would be required to manually annotate a corpus of events while simultaneously refining annotation standards.

We propose an approach to open-domain event categorization based on latent variable models that uncovers an appropriate set of types which match the data. The automatically discovered types are subsequently inspected to filter out any which are incoherent and the rest are annotated with informative labels;[2] examples of types discovered using our approach are listed in figure 3. The resulting set of types are then applied to categorize hundreds of millions of extracted events without the use of any manually annotated examples. By leveraging large quantities of unlabeled data, our approach results in a 14% improvement in F_1 score over a supervised baseline which uses the same set of types.

[2]This annotation and filtering takes minimal effort. One of the authors spent roughly 30 minutes inspecting and annotating the automatically discovered event types.

	P	R	F_1	F_1 inc.
Stanford NER	0.62	0.35	0.44	-
T-SEG	0.73	0.61	0.67	52%

Table 2: By training on in-domain data, we obtain a 52% improvement in F_1 score over the Stanford Named Entity Recognizer at segmenting entities in Tweets [46].

2. SYSTEM OVERVIEW

TWICAL extracts a 4-tuple representation of events which includes a named entity, event phrase, calendar date, and event type (see Table 1). This representation was chosen to closely match the way important events are typically mentioned in Twitter.

An overview of the various components of our system for extracting events from Twitter is presented in Figure 1. Given a raw stream of tweets, our system extracts named entities in association with event phrases and unambiguous dates which are involved in significant events. First the tweets are POS tagged, then named entities and event phrases are extracted, temporal expressions resolved, and the extracted events are categorized into types. Finally we measure the strength of association between each named entity and date based on the number of tweets they co-occur in, in order to determine whether an event is significant.

NLP tools, such as named entity segmenters and part of speech taggers which were designed to process edited texts (e.g. news articles) perform very poorly when applied to Twitter text due to its noisy and unique style. To address these issues, we utilize a named entity tagger and part of speech tagger trained on in-domain Twitter data presented in previous work [46]. We also develop an event tagger trained on in-domain annotated data as described in §4.

3. NAMED ENTITY SEGMENTATION

NLP tools, such as named entity segmenters and part of speech taggers which were designed to process edited texts (e.g. news articles) perform very poorly when applied to Twitter text due to its noisy and unique style.

For instance, capitalization is a key feature for named entity extraction within news, but this feature is highly unreliable in tweets; words are often capitalized simply for emphasis, and named entities are often left all lowercase. In addition, tweets contain a higher proportion of out-of-vocabulary words, due to Twitter's 140 character limit and the creative spelling of its users.

To address these issues, we utilize a named entity tagger trained on in-domain Twitter data presented in previous work [46].[3]

Training on tweets vastly improves performance at segmenting Named Entities. For example, performance compared against the state-of-the-art news-trained Stanford Named Entity Recognizer [17] is presented in Table 2. Our system obtains a 52% increase in F_1 score over the Stanford Tagger at segmenting named entities.

4. EXTRACTING EVENT MENTIONS

In order to extract event mentions from Twitter's noisy text, we first annotate a corpus of tweets, which is then

[3]Available at `http://github.com/aritter/twitter_nlp`.

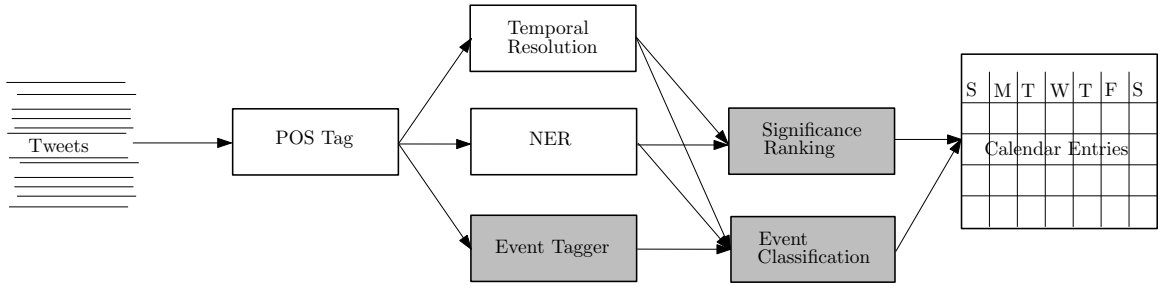

Figure 1: Processing pipeline for extracting events from Twitter. New components developed as part of this work are shaded in grey.

used to train sequence models to extract events. While we apply an established approach to sequence-labeling tasks in noisy text [46, 31, 19], this is the first work to extract event-referring phrases in Twitter.

Event phrases can consist of many different parts of speech as illustrated in the following examples:

- **Verbs:** Apple to *Announce* iPhone 5 on October 4th?! YES!
- **Nouns:** iPhone 5 *announcement* coming Oct 4th
- **Adjectives:** WOOOHOO *NEW* IPHONE TODAY! CAN'T WAIT!

These phrases provide important context, for example extracting the entity, **Steve Jobs** and the event phrase **died** in connection with October 5th, is much more informative than simply extracting **Steve Jobs**. In addition, event mentions are helpful in upstream tasks such as categorizing events into types, as described in §6.

In order to build a tagger for recognizing events, we annotated 1,000 tweets (19,484 tokens) with event phrases, following annotation guidelines similar to those developed for the EVENT tags in Timebank [43]. We treat the problem of recognizing event triggers as a sequence labeling task, using Conditional Random Fields for learning and inference [24]. Linear Chain CRFs model dependencies between the predicted labels of adjacent words, which is beneficial for extracting multi-word event phrases. We use contextual, dictionary, and orthographic features, and also include features based on our Twitter-tuned POS tagger [46], and dictionaries of event terms gathered from WordNet by Sauri et al. [50].

The precision and recall at segmenting event phrases are reported in Table 3. Our classifier, TWICAL-EVENT, obtains an F-score of 0.64. To demonstrate the need for in-domain training data, we compare against a baseline of training our system on the Timebank corpus.

	precision	recall	F1
TWICAL-EVENT	0.56	0.74	0.64
No POS	0.48	0.70	0.57
Timebank	0.24	0.11	0.15

Table 3: Precision and recall at event phrase extraction. All results are reported using 4-fold cross validation over the 1,000 manually annotated tweets (about 19K tokens). We compare against a system which doesn't make use of features generated based on our Twitter trained POS Tagger, in addition to a system trained on the Timebank corpus which uses the same set of features.

5. EXTRACTING AND RESOLVING TEMPORAL EXPRESSIONS

In addition to extracting events and related named entities, we also need to extract when they occur. In general there are many different ways users can refer to the same calendar date, for example "next Friday", "August 12th", "tomorrow" or "yesterday" could all refer to the same day, depending on when the tweet was written. To resolve temporal expressions we make use of TempEx [33], which takes as input a reference date, some text, and parts of speech (from our Twitter-trained POS tagger) and marks temporal expressions with unambiguous calendar references. Although this mostly rule-based system was designed for use on newswire text, we find its precision on Tweets (94% - estimated over as sample of 268 extractions) is sufficiently high to be useful for our purposes. TempEx's high precision on Tweets can be explained by the fact that some temporal expressions are relatively unambiguous. Although there appears to be room for improving the recall of temporal extraction on Twitter by handling noisy temporal expressions (for example see Ritter et. al. [46] for a list of over 50 spelling variations on the word "tomorrow"), we leave adapting temporal extraction to Twitter as potential future work.

6. CLASSIFICATION OF EVENT TYPES

To categorize the extracted events into types we propose an approach based on latent variable models which infers an appropriate set of event types to match our data, and also classifies events into types by leveraging large amounts of unlabeled data.

Supervised or semi-supervised classification of event categories is problematic for a number of reasons. First, it is *a priori* unclear which categories are appropriate for Twitter. Secondly, a large amount of manual effort is required to annotate tweets with event types. Third, the set of important categories (and entities) is likely to shift over time, or within a focused user demographic. Finally many important categories are relatively infrequent, so even a large annotated dataset may contain just a few examples of these categories, making classification difficult.

For these reasons we were motivated to investigate un-

Sports	7.45%	Conflict	0.69%
Party	3.66%	Prize	0.68%
TV	3.04%	Legal	0.67%
Politics	2.92%	Death	0.66%
Celebrity	2.38%	Sale	0.66%
Music	1.96%	VideoGameRelease	0.65%
Movie	1.92%	Graduation	0.63%
Food	1.87%	Racing	0.61%
Concert	1.53%	Fundraiser/Drive	0.60%
Performance	1.42%	Exhibit	0.60%
Fitness	1.11%	Celebration	0.60%
Interview	1.01%	Books	0.58%
ProductRelease	0.95%	Film	0.50%
Meeting	0.88%	Opening/Closing	0.49%
Fashion	0.87%	Wedding	0.46%
Finance	0.85%	Holiday	0.45%
School	0.85%	Medical	0.42%
AlbumRelease	0.78%	Wrestling	0.41%
Religion	0.71%	OTHER	53.45%

Figure 2: Complete list of automatically discovered event types with percentage of data covered. Interpretable types representing significant events cover roughly half of the data.

supervised approaches that will automatically induce event types which match the data. We adopt an approach based on latent variable models inspired by recent work on modeling selectional preferences [47, 39, 22, 52, 48], and unsupervised information extraction [4, 55, 7].

Each event indicator phrase in our data, e, is modeled as a mixture of types. For example the event phrase "cheered" might appear as part of either a PoliticalEvent, or a SportsEvent. Each type corresponds to a distribution over named entities n involved in specific instances of the type, in addition to a distribution over dates d on which events of the type occur. Including calendar dates in our model has the effect of encouraging (though not requiring) events which occur on the same date to be assigned the same type. This is helpful in guiding inference, because distinct references to the same event should also have the same type.

The generative story for our data is based on LinkLDA [15], and is presented as Algorithm 1. This approach has the advantage that information about an event phrase's type distribution is shared across it's mentions, while ambiguity is also naturally preserved. In addition, because the approach is based on generative a probabilistic model, it is straightforward to perform many different probabilistic queries about the data. This is useful for example when categorizing aggregate events.

For inference we use collapsed Gibbs Sampling [20] where each hidden variable, z_i, is sampled in turn, and parameters are integrated out. Example types are displayed in Figure 3. To estimate the distribution over types for a given event, a sample of the corresponding hidden variables is taken from the Gibbs markov chain after sufficient burn in. Prediction for new data is performed using a streaming approach to inference [56].

6.1 Evaluation

To evaluate the ability of our model to classify significant events, we gathered 65 million extracted events of the form

Label	Top 5 Event Phrases	Top 5 Entities
Sports	tailgate - scrimmage - tailgating - homecoming - regular season	espn - ncaa - tigers - eagles - varsity
Concert	concert - presale - performs - concerts - tickets	taylor swift - toronto - britney spears - rihanna - rock
Perform	matinee - musical - priscilla - seeing - wicked	shrek - les mis - lee evans - wicked - broadway
TV	new season - season finale - finished season - episodes - new episode	jersey shore - true blood - glee - dvr - hbo
Movie	watch love - dialogue theme - inception - hall pass - movie	netflix - black swan - insidious - tron - scott pilgrim
Sports	inning - innings - pitched - homered - homer	mlb - red sox - yankees - twins - dl
Politics	presidential debate - osama - presidential candidate - republican debate - debate performance	obama - president obama - gop - cnn - america
TV	network news broadcast - airing - primetime drama - channel - stream	nbc - espn - abc - fox - mtv
Product	unveils - unveiled - announces - launches - wraps off	apple - google - microsoft - uk - sony
Meeting	shows trading - hall - mtg - zoning - briefing	town hall - city hall - club - commerce - white house
Finance	stocks - tumbled - trading report - opened higher - tumbles	reuters - new york - u.s. - china - euro
School	maths - english test - exam - revise - physics	english - maths - german - bio - twitter
Album	in stores - album out - debut album - drops on - hits stores	itunes - ep - uk - amazon - cd
TV	voted off - idol - scotty - idol season - dividend-paying	lady gaga - american idol - america - beyonce - glee
Religion	sermon - preaching - preached - worship - preach	church - jesus - pastor - faith - god
Conflict	declared war - war - shelling - opened fire - wounded	libya - afghanistan - #syria - syria - nato
Politics	senate - legislation - repeal - budget - election	senate - house - congress - obama - gop
Prize	winners - lotto results - enter - winner - contest	ipad - award - facebook - good luck - winners
Legal	bail plea - murder trial - sentenced - plea - convicted	casey anthony - court - india - new delhi - supreme court
Movie	film festival - screening - starring - film - gosling	hollywood - nyc - la - los angeles - new york
Death	live forever - passed away - sad news - condolences - burried	michael jackson - afghanistan - john lennon - young - peace
Sale	add into - 50% off - up - shipping - save up	groupon - early bird - facebook - @etsy - etsy
Drive	donate - tornado relief - disaster relief - donated - raise money	japan - red cross - joplin - june - africa

Figure 3: Example event types discovered by our model. For each type t, we list the top 5 entities which have highest probability given t, and the 5 event phrases which assign highest probability to t.

Algorithm 1 Generative story for our data involving event types as hidden variables. Bayesian Inference techniques are applied to invert the generative process and infer an appropriate set of types to describe the observed events.

for each event type $t = 1 \dots T$ **do**
  Generate β_t^n according to symmetric Dirichlet distribution $\text{Dir}(\eta_n)$.
  Generate β_t^d according to symmetric Dirichlet distribution $\text{Dir}(\eta_d)$.
end for
for each unique event phrase $e = 1 \dots |E|$ **do**
  Generate θ_e according to Dirichlet distribution $\text{Dir}(\alpha)$.
  for each entity which co-occurs with e, $i = 1 \dots N_e$ **do**
    Generate $z_{e,i}^n$ from Multinomial(θ_e).
    Generate the entity $n_{e,i}$ from Multinomial($\beta_{z_{e,i}^n}$).
  end for
  for each date which co-occurs with e, $i = 1 \dots N_d$ **do**
    Generate $z_{e,i}^d$ from Multinomial(θ_e).
    Generate the date $d_{e,i}$ from Multinomial($\beta_{z_{d,i}^n}$).
  end for
end for

listed in Figure 1 (not including the type). We then ran Gibbs Sampling with 100 types for 1,000 iterations of burn-in, keeping the hidden variable assignments found in the last sample.[4]

One of the authors manually inspected the resulting types and assigned them labels such as SPORTS, POLITICS, MUSICRELEASE and so on, based on their distribution over entities, and the event words which assign highest probability to that type. Out of the 100 types, we found 52 to correspond to coherent event types which referred to significant events;[5] the other types were either incoherent, or covered types of events which are not of general interest, for example there was a cluster of phrases such as *applied*, *call*, *contact*, *job interview*, etc... which correspond to users discussing events related to searching for a job. Such event types which do not correspond to significant events of general interest were simply marked as *OTHER*. A complete list of labels used to annotate the automatically discovered event types along with the coverage of each type is listed in figure 2. Note that this assignment of labels to types only needs to be done once and produces a labeling for an arbitrarily large number of event instances. Additionally the same set of types can easily be used to classify new event instances using streaming inference techniques [56]. One interesting direction for future work is automatic labeling and coherence evaluation of automatically discovered event types analogous to recent work on topic models [38, 25].

In order to evaluate the ability of our model to classify aggregate events, we grouped together all (entity,date) pairs which occur 20 or more times the data, then annotated the 500 with highest association (see §7) using the event types discovered by our model.

To help demonstrate the benefits of leveraging large quantities of unlabeled data for event classification, we compare against a supervised Maximum Entropy baseline which makes use of the 500 annotated events using 10-fold cross validation. For features, we treat the set of event phrases that co-occur with each (entity, date) pair as a bag-of-words, and also include the associated entity. Because many event categories are infrequent, there are often few or no training examples for a category, leading to low performance.

[4]To scale up to larger datasets, we performed inference in parallel on 40 cores using an approximation to the Gibbs Sampling procedure analogous to that presented by Newmann et. al. [37].

[5]After labeling some types were combined resulting in 37 distinct labels.

	Precision	Recall	F_1
TWICAL-CLASSIFY	0.85	0.55	0.67
Supervised Baseline	0.61	0.57	0.59

Table 4: Precision and recall of event type categorization at the point of maximum F_1 score.

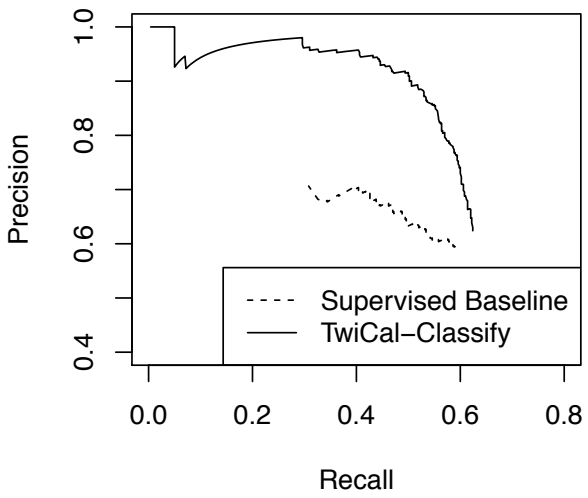

Figure 4: Precision and recall predicting event types.

Figure 4 compares the performance of our unsupervised approach to the supervised baseline, via a precision-recall curve obtained by varying the threshold on the probability of the most likely type. In addition table 4 compares precision and recall at the point of maximum F-score. Our unsupervised approach to event categorization achieves a 14% increase in maximum F_1 score over the supervised baseline.

Figure 5 plots the maximum F_1 score as the amount of training data used by the baseline is varied. It seems likely that with more data, performance will reach that of our approach which does not make use of any annotated events, however our approach both automatically discovers an appropriate set of event types and provides an initial classifier with minimal effort, making it useful as a first step in situations where annotated data is not immediately available.

7. RANKING EVENTS

Simply using frequency to determine which events are significant is insufficient, because many tweets refer to common events in user's daily lives. As an example, users often mention what they are eating for lunch, therefore entities such as *McDonalds* occur relatively frequently in association with references to most calendar days. Important events can be distinguished as those which have strong association with a unique date as opposed to being spread evenly across days on the calendar. To extract significant events of general interest from Twitter, we thus need some way to measure the strength of association between an entity and a date.

In order to measure the association strength between an

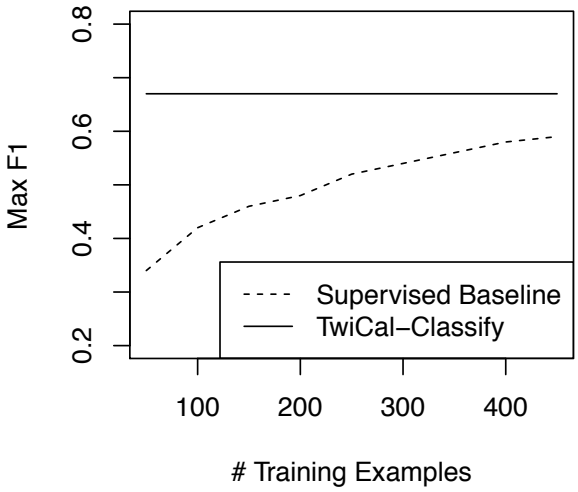

Figure 5: Maximum F_1 score of the supervised baseline as the amount of training data is varied.

entity and a specific date, we utilize the G^2 log likelihood ratio statistic. G^2 has been argued to be more appropriate for text analysis tasks than χ^2 [12]. Although Fisher's Exact test would produce more accurate p-values [34], given the amount of data with which we are working (sample size greater than 10^{11}), it proves difficult to compute Fisher's Exact Test Statistic, which results in floating point overflow even when using 64-bit operations. The G^2 test works sufficiently well in our setting, however, as computing association between entities and dates produces less sparse contingency tables than when working with pairs of entities (or words).

The G^2 test is based on the likelihood ratio between a model in which the entity is conditioned on the date, and a model of independence between entities and date references. For a given entity e and date d this statistic can be computed as follows:

$$G^2 = \sum_{x \in \{e, \neg e\}, y \in \{d, \neg d\}} O_{x,y} \times \ln\left(\frac{O_{x,y}}{E_{x,y}}\right)$$

Where $O_{e,d}$ is the observed fraction of tweets containing both e and d, $O_{e,\neg d}$ is the observed fraction of tweets containing e, but not d, and so on. Similarly $E_{e,d}$ is the expected fraction of tweets containing both e and d assuming a model of independence.

8. EXPERIMENTS

To estimate the quality of the calendar entries generated using our approach we manually evaluated a sample of the top 100, 500 and 1,000 calendar entries occurring within a 2-week future window of November 3rd.

8.1 Data

For evaluation purposes, we gathered roughly the 100 million most recent tweets on November 3rd 2011 (collected using the Twitter Streaming API[6], and tracking a broad set of temporal keywords, including "today", "tomorrow", names of weekdays, months, etc.).

We extracted named entities in addition to event phrases, and temporal expressions from the text of each of the 100M tweets. We then added the extracted triples to the dataset used for inferring event types described in §6, and performed 50 iterations of Gibbs sampling for predicting event types on the new data, holding the hidden variables in the original data constant. This *streaming* approach to inference is similar to that presented by Yao et al. [56].

We then ranked the extracted events as described in §7, and randomly sampled 50 events from the top ranked 100, 500, and 1,000. We annotated the events with 4 separate criteria:

1. Is there a significant event involving the extracted entity which will take place on the extracted date?
2. Is the most frequently extracted event phrase informative?
3. Is the event's type correctly classified?
4. Are each of (1-3) correct? That is, does the event contain a correct entity, date, event phrase, and type?

Note that if (1) is marked as incorrect for a specific event, subsequent criteria are always marked incorrect.

8.2 Baseline

To demonstrate the importance of natural language processing and information extraction techniques in extracting informative events, we compare against a simple baseline which does not make use of the Ritter et. al. named entity recognizer or our event recognizer; instead, it considers all 1-4 grams in each tweet as candidate calendar entries, relying on the G^2 test to filter out phrases which have low association with each date.

8.3 Results

The results of the evaluation are displayed in table 5. The table shows the precision of the systems at different yield levels (number of aggregate events). These are obtained by varying the thresholds in the G^2 statistic. Note that the baseline is only comparable to the third column, i.e., the precision of (entity, date) pairs, since the baseline is not performing event identification and classification. Although in some cases ngrams do correspond to informative calendar entries, the precision of the ngram baseline is extremely low compared with our system.

In many cases the ngrams don't correspond to salient entities related to events; they often consist of single words which are difficult to interpret, for example "Breaking" which is part of the movie "Twilight: Breaking Dawn" released on November 18. Although the word "Breaking" has a strong association with November 18, by itself it is not very informative to present to a user.[7]

Our high-confidence calendar entries are surprisingly high quality. If we limit the data to the 100 highest ranked calendar entries over a two-week date range in the future, the precision of extracted (entity, date) pairs is quite good (90%) - an 80% increase over the ngram baseline. As expected precision drops as more calendar entries are displayed, but

[6] `https://dev.twitter.com/docs/streaming-api`

[7] In addition, we notice that the ngram baseline tends to produce many near-duplicate calendar entries, for example: "Twilight Breaking", "Breaking Dawn", and "Twilight Breaking Dawn". While each of these entries was annotated as correct, it would be problematic to show this many entries describing the same event to a user.

November 2011						
Mon Nov 7	**Tue Nov 8**	**Wed Nov 9**	**Thu Nov 10**	**Fri Nov 11**	**Sat Nov 12**	**Sun Nov 13**
Justin *meet* *Other*	Paris *love* *Other*	EAS *test* *Other*	Robert Pattinson *show* *Performance*	iPhone *debut* *Product Release*	Sydney *perform* *Other*	Playstation *answers* *Product Release*
Motorola Pro+ *kick* *Product Release*	iPhone *holding* *Product Release*	The Feds *cut off* *Other*	James Murdoch *give evidence* *Other*	Remembrance Day *open* *Performance*	Pullman Ballroom *promoted* *Other*	Samsung Galaxy Tab *launch* *Product Release*
Nook Color 2 *launch* *Product Release*	Election Day *vote* *Political Event*	Toca Rivera *promoted* *Performance*	RTL-TVI *post* *TV Event*	France *play* *Other*	Fox *fight* *Other*	Sony *answers* *Product Release*
Eid-ul-Azha *celebrated* *Performance*	Blue Slide Park *listening* *Music Release*	Alert System *test* *Other*	Gotti Live *work* *Other*	Veterans Day *closed* *Other*	Plaza *party* *Party*	Chibi Chibi Burger *other*
MW3 *midnight release* *Other*	Hedley *album* *Music Release*	Max Day *give* *Other*	Bambi Awards *perform* *Performance*	Skyrim *arrives* *Product Release*	Red Carpet *invited* *Party*	Jiexpo Kemayoran *promoted* *TV Event*

Figure 6: Example future calendar entries extracted by our system for the week of November 7th. Data was collected up to November 5th. For each day, we list the top 5 events including the entity, *event phrase*, and *event type*. While there are several errors, the majority of calendar entries are informative, for example: the Muslim holiday eid-ul-azha, the release of several videogames: Modern Warfare 3 (MW3) and Skyrim, in addition to the release of the new playstation 3D display on Nov 13th, and the new iPhone 4S in Hong Kong on Nov 11th.

	precision				
# calendar entries	ngram baseline	entity + date	event phrase	event type	entity + date + event + type
100	0.50	0.90	0.86	0.72	0.70
500	0.46	0.66	0.56	0.54	0.42
1,000	0.44	0.52	0.42	0.40	0.32

Table 5: Evaluation of precision at different recall levels (generated by varying the threshold of the G^2 statistic). We evaluate the top 100, 500 and 1,000 (entity, date) pairs. In addition we evaluate the precision of the most frequently extracted event phrase, and the predicted event type in association with these calendar entries. Also listed is the fraction of cases where all predictions ("entity + date + event + type") are correct. We also compare against the precision of a simple ngram baseline which does not make use of our NLP tools. Note that the ngram baseline is only comparable to the entity+date precision (column 3) since it does not include event phrases or types.

remains high enough to display to users (in a ranked list). In addition to being less likely to come from extraction errors, highly ranked entity/date pairs are more likely to relate to popular or important events, and are therefore of greater interest to users.

In addition we present a sample of extracted future events on a calendar in figure 6 in order to give an example of how they might be presented to a user. We present the top 5 entities associated with each date, in addition to the most frequently extracted event phrase, and highest probability event type.

8.4 Error Analysis

We found 2 main causes for why entity/date pairs were uninformative for display on a calendar, which occur in roughly equal proportion:

Segmentation Errors Some extracted "entities" or ngrams don't correspond to named entities or are generally uninformative because they are mis-segmented. Examples include "RSVP", "Breaking" and "Yikes".

Weak Association between Entity and Date In some cases, entities are properly segmented, but are uninformative because they are not strongly associated with a specific event on the associated date, or are involved in many different events which happen to occur on that day. Examples include locations such as "New York", and frequently mentioned entities, such as "Twitter".

9. RELATED WORK

While we are the first to study open domain event extraction within Twitter, there are two key related strands of research: extracting specific types of events from Twitter, and extracting open-domain events from news [43].

Recently there has been much interest in information extraction and event identification within Twitter. Benson et al. [5] use distant supervision to train a relation extractor which identifies artists and venues mentioned within tweets of users who list their location as New York City. Sakaki et al. [49] train a classifier to recognize tweets reporting earthquakes in Japan; they demonstrate their system is capable of recognizing almost all earthquakes reported by the Japan Meteorological Agency. Additionally there is recent work on detecting events or tracking topics [29] in Twitter which does not extract structured representations, but has the advantage that it is not limited to a narrow domain. Petrović et al. investigate a streaming approach to identifying Tweets which are the first to report a breaking news story using Locally Sensitive Hash Functions [40]. Becker et al. [3], Popescu et al. [42, 41] and Lin et al. [28] investigate discovering clusters of related words or tweets which correspond to events in progress. In contrast to previous work on Twitter event identification, our approach is independent of event type or domain and is thus more widely applicable. Additionally, our work focuses on extracting a calendar of events (including those occurring in the future), extract-

ing event-referring expressions and categorizing events into types.

Also relevant is work on identifying events [23, 10, 6], and extracting timelines [30] from news articles.[8] Twitter status messages present both unique challenges and opportunities when compared with news articles. Twitter's noisy text presents serious challenges for NLP tools. On the other hand, it contains a higher proportion of references to present and future dates. Tweets do not require complex reasoning about relations between events in order to place them on a timeline as is typically necessary in long texts containing narratives [51]. Additionally, unlike News, Tweets often discus mundane events which are not of general interest, so it is crucial to exploit redundancy of information to assess whether an event is significant.

Previous work on open-domain information extraction [2, 53, 16] has mostly focused on extracting relations (as opposed to events) from web corpora and has also extracted relations based on verbs. In contrast, this work extracts events, using tools adapted to Twitter's noisy text, and extracts event phrases which are often adjectives or nouns, for example: *Super Bowl* **Party** *on Feb 5th.*

Finally we note that there has recently been increasing interest in applying NLP techniques to short informal messages such as those found on Twitter. For example, recent work has explored Part of Speech tagging [19], geographical variation in language found on Twitter [13, 14], modeling informal conversations [44, 45, 9], and also applying NLP techniques to help crisis workers with the flood of information following natural disasters [35, 27, 36].

10. CONCLUSIONS

We have presented a scalable and open-domain approach to extracting and categorizing events from status messages. We evaluated the quality of these events in a manual evaluation showing a clear improvement in performance over an ngram baseline

We proposed a novel approach to categorizing events in an open-domain text genre with unknown types. Our approach based on latent variable models first discovers event types which match the data, which are then used to classify aggregate events without any annotated examples. Because this approach is able to leverage large quantities of unlabeled data, it outperforms a supervised baseline by 14%.

A possible avenue for future work is extraction of even richer event representations, while maintaining domain independence. For example: grouping together related entities, classifying entities in relation to their roles in the event, thereby, extracting a frame-based representation of events.

A continuously updating demonstration of our system can be viewed at `http://statuscalendar.com`; Our NLP tools are available at `http://github.com/aritter/twitter_nlp`.

[8] `http://newstimeline.googlelabs.com/`

11. ACKNOWLEDGEMENTS

The authors would like to thank Luke Zettlemoyer and the anonymous reviewers for helpful feedback on a previous draft. This research was supported in part by NSF grant IIS-0803481 and ONR grant N00014-08-1-0431 and carried out at the University of Washington's Turing Center.

12. REFERENCES

[1] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *SIGIR*, 1998.

[2] M. Banko, M. J. Cafarella, S. Soderl, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *In IJCAI*, 2007.

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

[4] C. Bejan, M. Titsworth, A. Hickl, and S. Harabagiu. Nonparametric bayesian models for unsupervised event coreference resolution. In *NIPS*. 2009.

[5] E. Benson, A. Haghighi, and R. Barzilay. Event discovery in social media feeds. In *ACL*, 2011.

[6] S. Bethard and J. H. Martin. Identification of event mentions and their semantic class. In *EMNLP*, 2006.

[7] N. Chambers and D. Jurafsky. Template-based information extraction without the templates. In *Proceedings of ACL*, 2011.

[8] N. Chambers, S. Wang, and D. Jurafsky. Classifying temporal relations between events. In *ACL*, 2007.

[9] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais. Mark my words! Linguistic style accommodation in social media. In *Proceedings of WWW*, pages 745–754, 2011.

[10] A. Das Sarma, A. Jain, and C. Yu. Dynamic relationship and event discovery. In *WSDM*, 2011.

[11] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. *LREC*, 2004.

[12] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 1993.

[13] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *EMNLP*, 2010.

[14] J. Eisenstein, N. A. Smith, and E. P. Xing. Discovering sociolinguistic associations with structured sparsity. In *ACL-HLT*, 2011.

[15] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *PNAS*, 2004.

[16] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP*, 2011.

[17] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005.

[18] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *WWW*, 2004.

[19] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging

for twitter: Annotation, features, and experiments. In *ACL*, 2011.

[20] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1, 2004.

[21] R. Grishman and B. Sundheim. Message understanding conference - 6: A brief history. In *Proceedings of the International Conference on Computational Linguistics*, 1996.

[22] Z. Kozareva and E. Hovy. Learning arguments and supertypes of semantic relations using recursive patterns. In *ACL*, 2010.

[23] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *SIGIR*, 2004.

[24] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[25] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin. Automatic labelling of topic models. In *ACL*, 2011.

[26] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, 2009.

[27] W. Lewis, R. Munro, and S. Vogel. Crisis mt: Developing a cookbook for mt in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011.

[28] C. X. Lin, B. Zhao, Q. Mei, and J. Han. PET: a statistical model for popular events tracking in social communities. In *KDD*, 2010.

[29] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: Topic tracking in tweet streams. In *KDD*, 2011.

[30] X. Ling and D. S. Weld. Temporal information extraction. In *AAAI*, 2010.

[31] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *ACL*, 2011.

[32] I. Mani, M. Verhagen, B. Wellner, C. M. Lee, and J. Pustejovsky. Machine learning of temporal relations. In *ACL*, 2006.

[33] I. Mani and G. Wilson. Robust temporal processing of news. In *ACL*, 2000.

[34] R. C. Moore. On log-likelihood-ratios and the significance of rare events. In *EMNLP*, 2004.

[35] R. Munro. Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol. In *CoNLL*, 2011.

[36] G. Neubig, Y. Matsubayashi, M. Hagiwara, and K. Murakami. Safety information mining - what can NLP do in a disaster -. In *IJCNLP*, 2011.

[37] D. Newman, A. U. Asuncion, P. Smyth, and M. Welling. Distributed inference for latent dirichlet allocation. In *NIPS*, 2007.

[38] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *HLT-NAACL*, 2010.

[39] D. Ó Séaghdha. Latent variable models of selectional preference. In *ACL*, ACL '10, 2010.

[40] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *HLT-NAACL*, 2010.

[41] A.-M. Popescu and M. Pennacchiotti. Dancing with the stars, nba games, politics: An exploration of twitter users' response to events. In *ICWSM*, 2011.

[42] A.-M. Popescu, M. Pennacchiotti, and D. A. Paranjpe. Extracting events and event descriptions from twitter. In *WWW*, 2011.

[43] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003*, 2003.

[44] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *HLT-NAACL*, 2010.

[45] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *EMNLP*, 2011.

[46] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. *EMNLP*, 2011.

[47] A. Ritter, Mausam, and O. Etzioni. A latent dirichlet allocation method for selectional preferences. In *ACL*, 2010.

[48] K. Roberts and S. M. Harabagiu. Unsupervised learning of selectional restrictions and detection of argument coercions. In *EMNLP*, 2011.

[49] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, 2010.

[50] R. Saurí, R. Knippen, M. Verhagen, and J. Pustejovsky. Evita: a robust event recognizer for qa systems. In *HLT-EMNLP*, 2005.

[51] F. Song and R. Cohen. Tense interpretation in the context of narrative. In *Proceedings of the ninth National conference on Artificial intelligence - Volume 1*, AAAI'91, 1991.

[52] B. Van Durme and D. Gildea. Topic models for corpus-centric knowledge generalization. In *Technical Report TR-946, Department of Computer Science, University of Rochester, Rochester*, 2009.

[53] D. S. Weld, R. Hoffmann, and F. Wu. Using wikipedia to bootstrap open information extraction. *SIGMOD Rec.*, 2009.

[54] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, 1998.

[55] L. Yao, A. Haghighi, S. Riedel, and A. McCallum. Structured relation discovery using generative models. In *EMNLP*, 2011.

[56] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *KDD*, 2009.

[57] F. M. Zanzotto, M. Pennaccchiotti, and K. Tsioutsiouliklis. Linguistic redundancy in twitter. In *EMNLP*, 2011.