Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts

Jiwei Li[1], **Alan Ritter**[2], **Claire Cardie**[3] **and Eduard Hovy**[4]
[1]Computer Science Department, Stanford University, Stanford, CA 94305, USA
[2]Department of Computer Science and Engineering, the Ohio State University, OH 43210, USA
[3]Computer Science Department, Cornell University, Ithaca, NY 14853, USA
[4]Language Technology Institute, Carnegie Mellon University, PA 15213, USA
jiweil@stanford.edu ritter.1492@osu.edu
cardie@cs.cornell.edu ehovy@andrew.cmu.edu

Abstract

Social media websites provide a platform for anyone to describe significant events taking place in their lives in realtime. Currently, the majority of personal news and life events are published in a textual format, motivating information extraction systems that can provide a structured representations of major life events (weddings, graduation, etc...). This paper demonstrates the feasibility of accurately extracting major life events. Our system extracts a fine-grained description of users' life events based on their published tweets. We are optimistic that our system can help Twitter users more easily grasp information from users they take interest in following and also facilitate many downstream applications, for example realtime friend recommendation.

1 Introduction

Social networking websites such as Facebook and Twitter have recently challenged mainstream media as the freshest source of information on important news events. In addition to an important source for breaking news, social media presents a unique source of information on private events, for example a friend's engagement or college graduation (examples are presented in Figure 1). While a significant amount of previous work has investigated event extraction from Twitter (e.g., (Ritter et al., 2012; Diao et al., 2012)), existing approaches mostly focus on public bursty event extraction, and little progress has been made towards the problem of automatically extracting the major life events of ordinary users.

A system which can automatically extract major life events and generate fine-grained descriptions as in Figure 1 will not only help Twitter users with the problem of information overload by summarizing important events taking place in their friends lives, but could also facilitate downstream applications such as friend recommendation (e.g., friend recommendation in realtime to people who were just admitted into the same university, get the same jobs or internships), targeted online advertising (e.g., recommend baby care products to newly expecting mothers, or wedding services to new couples), information extraction, etc.

Before getting started, we first identify a number of key challenges in extracting significant life events from user-generated text, which account the reason for the lack of previous work in this area:

Challenge 1: Ambiguous Definition for Major Life Events Major life event identification is an open-domain problem. While many types of events (e.g., marriage, engagement, finding a new job, giving birth) are universally agreed to be important, it is difficult to robustly predefine a list of characteristics for important life events on which algorithms can rely for extraction or classification.

Challenge 2: Noisiness of Twitter Data: The user-generated text found in social media websites such as Twitter is extremely noisy. The language used to describe life events is highly varied and ambiguous and social media users frequently discuss public news and mundane events from their daily lives, for instance what they ate for lunch.

Even for a predefined life event category, such as marriage, it is still difficult to accurately identify mentions. For instance, a search for the keyphrase "get married" using Twitter Search[1] results in a large number of returned results that do not correspond to a personal event:

- I want to **get married** once. No divorce & no cheating, just us two till the end.
  (error: wishes)

[1]https://twitter.com/search?q=get~married

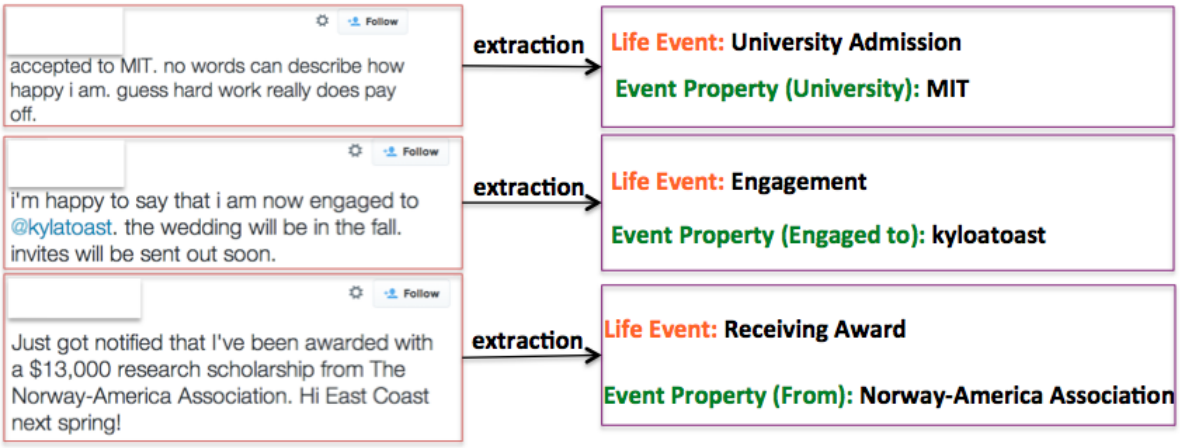

Figure 1: Examples of users mentioning personal life events on Twitter.

- Can Adam Sandler and Drew Barrymore just drop the pretense and **get married** already? (error: somebody else)
- I **got married** and had kids on purpose (error: past)

Challenge 3: the Lack of Training Data Collecting sufficient training data in this task for machine learning models is difficult for a number of reasons: (1) A traditional, supervised learning approach, requires explicit annotation guidelines for labeling, though it is difficult to know which categories are most representative in the data apriori. (2) Unlike public events which are easily identified based on message volume, significant private events are only mentioned by one or several users directly involved in the event. Many important categories are relatively infrequent, so even a large annotated dataset may contain just a few or no examples of these categories, making classification difficult.

In this paper, we present a pipelined system that addresses these challenges and extracts a structured representation of individual life events based on users' Twitter feeds. We exploit the insight to automatically gather large volumes of major life events which can be used as training examples for machine learning models. Although personal life events are difficult to identify using traditional approaches due to their highly diverse nature, we noticed that users' followers often directly reply to such messages with CONGRATULATIONS or CONDOLENCES speech acts, for example:

User1: *I got accepted into Harvard !*
User2: *Congratulations !*

These speech acts are easy to identify with high precision because the possible ways to express them are relatively constrained. Instead of directly inspecting tweets to determine whether they correspond to major life events, we start by identifying replies corresponding to CONGRATULATIONS or CONDOLENCES, and then retrieve the message they are in response to, which we assume refer to important life events.

The proposed system automatically identifies major life events and then extracts correspondent event properties. Through the proposed system, we demonstrate that it is feasible to automatically reconstruct a detailed list of individual life events based on users' Twitter streams. We hope that work presented in this paper will facilitate downstream applications and encourage follow-up work on this task.

2 System Overview

An overview of the components of the system is presented in Figure 2. **Pipeline1** first identifies the major life event category the input tweet talks about and filters out the irrelevant tweets and will be described in Section 4. Next, **Pipeline2**, as, demonstrated in Section 5, identifies whether the speaker is directly involved in the life event. Finally, **Pipeline3** extracts the property of event and will be illustrated in Section 6.

Section 3 serves as the preparing step for the pipelined system, describing how we collect training data in large-scale. The experimental evaluation regarding each pipeline of the system is presented in the corresponding section (i.e., Section 4,5,6) and the end-to-end evaluation will be pre-

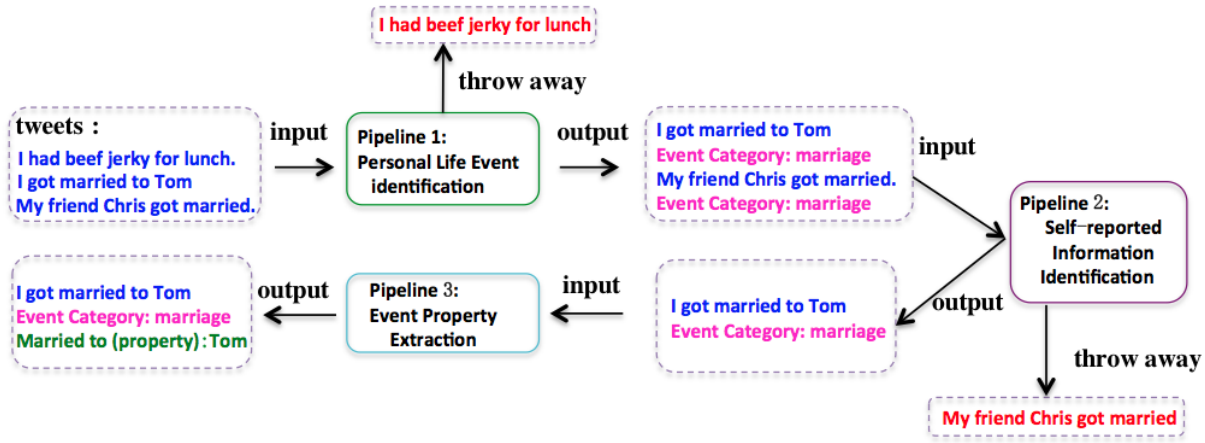

Figure 2: System Overview. Blue: original input tweets. Red: filtered out tweets. Magenta: life event category. Green: life event property. **Pipeline 1** identifies the life category the input tweet talks about (e.g., marriage, graduation) and filter out irrelevant tweets (e.g., I had beef stick for lunch). **Pipeline 2** identifies whether the speaker is directly involved in the event. It will preserve self-reported information (i.e. "I got married") and filtered out unrelated tweets (e.g., "my friend Chris got married"). **Pipeline 3** extracts the property of event (e.g. to whom the speaker married or the speaker admitted by which university).

sented in Section 7.

3 Personal Life Event Clustering

In this section, we describe how we identify common categories of major life events by leveraging large quantities of unlabeled data and obtain a collection of tweets corresponding to each type of identified event.

3.1 Response based Life Event Detection

While not all major life events will elicit CONGRATULATIONS or CONDOLENCES from a user's followers, this technique allows us to collect large volumes of high-precision personal life events which can be used to train models to recognize the diverse categories of major life events discussed by social media users.

3.2 Life Event Clustering

Based on the above intuition, we develop an approach to obtain a list of individual life event clusters. We first define a small set of seed responses which capture common CONGRATULATIONS and CONDOLENCES, including the phrases: "Congratulations", "Congrats", "Sorry to hear that", "Awesome", and gather tweets that were observed with seed responses. Next, an LDA (Blei et al., 2003)[2] based topic model is used to cluster the gathered tweets to automatically identify important categories of major life events in an unsupervised way. In our approach, we model the whole conversation dialogue as a document[3] with the response seeds (e.g., congratulation) masked out. We furthermore associate each sentence with a single topic, following strategies adopted by (Ritter et al., 2010; Gruber et al., 2007). We limit the words in our document collection to verbs and nouns which we found to lead to clearer topic representations, and used collapsed Gibbs Sampling for inference (Griffiths and Steyvers, 2004).

Next one of the authors manually inspected the resulting major life event types inferred by the model, and manually assigned them labels such as "getting a job", "graduation" or "marriage" and discarded incoherent topics[4]. Our methodology is inspired by (Ritter et al., 2012) that uses a LDA-CLUSTERING+HUMAN-IDENTIFICATION strategy to identify public events from Twitter. Similar strategies have been widely used in unsupervised information extraction (Bejan et al., 2009; Yao et al., 2011) and selectional preference

[2]Topic Number is set to 120.

[3]Each whole conversation usually contains multiple tweets and users.

[4]While we applied manual labeling and coherence evaluation in this work, an interesting direction for future work is automatically labeling major life event categories following previous work on labeling topics in traditional document-based topic models (Mimno et al., 2011; Newman et al., 2010).

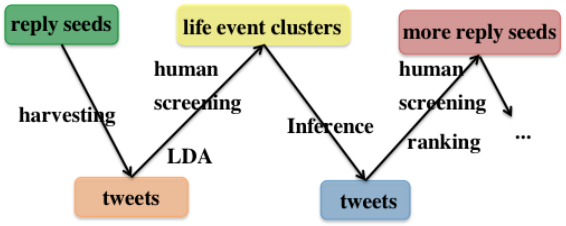

Figure 3: Illustration of bootstrapping process.

Input: Reply seed list $E = \{e\}$, Tweet conversation collection $T = \{t\}$, Retrieved Tweets Collection $D = \phi$. Identified topic list L=ϕ
Begin
While not stopping:

1. For unprocessed conversation $t \in T$ if t contains reply $e \in E$,
   - add t to D: $D = D + t$.
   - remove t from T: $T = T - t$
2. Run streaming LDA (Yao et al., 2009) on newly added tweets in D.
3. Manually Identify meaningful/trash topics, giving label to meaningful topics.
4. Add newly detected meaningful topic l to L.
5. For conversation t belonging to trash topics
   - remove t from D: $D = D - t$
6. Harvest more tweets based on topic distribution.
7. Manually identify top 20 responses to tweets harvested from Step 6.
8. Add meaningful responses to E.

End
Output: Identified topic list L. Tweet collection D.

Figure 4: Bootstrapping Algorithm for Response-based Life event identification.

modeling (Kozareva and Hovy, 2010a; Roberts and Harabagiu, 2011).

Conversation data was extracted from the CMU Twitter Warehouse of 2011 which contains a total number of 10% of all published tweets in that year.

3.3 Expanding dataset using Bootstrapping

While our seed patterns for identifying messages expressing CONGRATULATIONS and CONDOLENCES are very high precision, they don't cover all the possible ways these speech acts can be expressed. We therefore adopt a semi-supervised bootstrapping approach to expand our reply seeds and event-related tweets. Our bootstrapping approach is related to previous work on semi-supervised information harvesting (e.g., (Kozareva and Hovy, 2010b; Davidov et al., 2007)). To preserve the labeled topics from the first iteration, we apply a streaming approach to inference (Yao et al., 2009) over unlabeled tweets (those which did not match one of the response

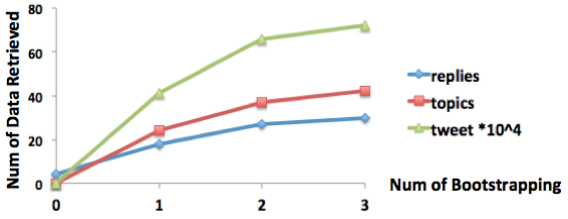

Figure 5: Illustration of data retrieved in each step of bootstrapping.

congratulations (cong, congrats); (that's) fantastic; (so) cool; (I'm) (very) sorry to hear that; (that's) great (good) new; awesome; what a pity; have fun; great; that sucks; too bad; (that's) unfortunate; how sad; fabulous; (that's) terrific; (that's) (so) wonderful; my deepest condolences;

Table 1: Responses retrieved from Bootstrapping.

seeds). We collect responses to the newly added tweets, then select the top 20 frequent replies[5]. Next we manually inspect and filter the top ranked replies, and use them to harvest more tweets. This process is then repeated with another round of inference in LDA including manual labeling of newly inferred topics, etc... An illustration of our approach is presented in Figure 3 and the details are presented in Figure 4. The algorithm outputs a collection of personal life topics L, and a collection of retrieved tweets D. Each tweet $d \in D$ is associated with a life event topic l, $l \in L$.

We repeat the bootstrapping process for 4 iterations and end up with 30 different CONGRATULATIONS and CONDOLENCES patterns (shown in Table 1) and 42 coherent event types which refer to significant life events (statistics for harvested data from each step is shown in Figure 5). We show examples of the mined topics with correspondent human labels in Table 3, grouped according to a specific kind of resemblance.

3.4 Summary and Discussion

The objective of this section is (1) identifying a category of life events (2) identifying tweets associated with each event type which can be used as candidates for latter self reported personal information and life event category identification.

We understand that the event list retrieved from our approach based on replies in the conversation is far from covering all types of personal events (especially the less frequent life events). But our

[5]We only treat the first sentence that responds to the beginning of the conversation as replies.

Life Event	Proportion
Birthday	9.78
Job	8.39
Wedding Engagement	7.24
Award	6.20
Sports	6.08
Anniversary	5.44
Give Birth	4.28
Graduate	3.86
Death	3.80
Admission	3.54
Interview Internship	3.44
Moving	3.26
Travel	3.24
Illness	2.45

Life Event	Proportion
Vacation	2.24
Relationship	2.16
Exams	2.02
Election	1.85
New Car	1.65
Running	1.42
Surgery	1.20
Lawsuit	0.64
Acting	0.50
Research	0.48
Essay	0.35
Lost Weight	0.35
Publishing	0.28
Song	0.22
OTHER	15.31

Table 2: List of automatically discovered life event types with percentage (%) of data covered.

list is still able to cover a large proportion of IMPORTANT and COMMON life events. Our latter work is focused on given a random tweet, identifying whether it corresponds to one of the 42 types of life events in our list.

Another thing worth noting here is that, while current section is not focused on self-reported information identification, we have already obtained a relatively clean set of data with a large proportion of non self-reported information related tweets being screened: people do not usually respond to non self-reported information with commonly used replies, or in other words, with replies that will pass our next step human test[6]. These non self-reported tweets would therefore be excluded from training data.

4 Life Event Identification

In this section, we focused on deciding whether a given tweet corresponds to one of the 42 predefined life events.

Our training dataset consists of approximately 72,000 tweets from 42 different categories of life events inferred by our topic model as described in Section 3. We used the top 25% of tweets for which our model assigned highest probability to each topic. For sparsely populated topics we used the top 50% of tweets to ensure sufficient coverage.

We further collected a random sample of about 10 million tweets from Twitter API[7] as non-life

[6]For example, people don't normally respond to "I want to **get married** once" (example in Challenge 2, Section 1) with "Congratulations".

[7]https://dev.twitter.com/

Human Label	Top words
Wedding &engagement	wedding, love, ring, engagement, engaged, bride, video, marrying
Relationship Begin	boyfriend, girlfriend, date, check, relationship, see, look
Anniversary	anniversary, years, year, married, celebrating, wife, celebrate, love
Relation End/ Devoice	relationship, ended, hurt, hate, devoice, blessings, single
Graduation	graduation, school, college, graduate, graduating, year, grad
Admission	admitted, university, admission, accepted, college, offer, school
Exam	passed, exam, test, school, semester, finished, exams, midterms
Research	research, presentation, journalism, paper, conference, go, writing
Essay & Thesis	essay, thesis, reading, statement, dissertation, complete, project
Job	job, accepted, announce, join, joining, offer, starting, announced, work
Interview& Internship	interview, position, accepted, internship, offered, start, work
Moving	house, moving, move, city, home, car, place, apartment, town, leaving
Travel	leave, leaving, flight, home, miss, house, airport, packing, morning
Vacation	vocation, family, trip, country, go, flying, visited, holiday, Hawaii
Winning Award	won, award, support, awards, winning, honor, scholarship, prize
Election/ Promotion/ Nomination	president, elected, run, nominated, named, promotion, cel, selected, business, vote
Publishing	book, sold, writing, finished, read, copy, review, release, books, cover
Contract	signed, contract, deal, agreements, agreed, produce, dollar, meeting
song/ video/ album release	video, song, album, check, show, see, making, radio, love
Acting	play, role, acting, drama, played, series, movie, actor, theater
Death	dies, passed, cancer, family, hospital, dad, grandma, mom, grandpa
Give Birth	baby, born, boy, pregnant, girl, lbs, name, son, world, daughter, birth
Illness	ill, hospital, feeling, sick, cold, flu, getting, fever, doctors, cough
Surgery	surgery, got, test, emergency, blood, tumor, stomachs, hospital, pain, brain
Sports	win, game, team, season, fans, played, winning, football, luck
Running	run, race, finished, race, marathon, ran, miles, running, finish, goal
New Car	car, buy, bought, cars, get, drive, pick, seat, color, dollar, meet
Lost Weight	weight, lost, week, pounds, loss, weeks, gym, exercise, running
Birthday	birthday, come, celebrate, party, friends, dinner, tonight, friend
Lawsuit	sue, sued, file, lawsuit, lawyer, dollars, illegal, court, jury.

Table 3: Example event types with top words discovered by our model.

event examples and trained a 43-class maximum entropy classifier based on the following features:

- **Word**: The sequence of words in the tweet.
- **NER**: Named entity Tag.
- **Dictionary**: Word matching a dictionaries of the top 40 words for each life event category (automatically inferred by the topic model). The feature value is the term's probability generated by correspondent event.
- **Window**: If a dictionary term exists, left and right context words within a window of 3 words and their part-of-speech tags.

Name entity tag is assigned from Ritter et al's Twitter NER system (Ritter et al., 2011). Part-of-Speech tags are assigned based on Twitter POS package (Owoputi et al., 2013) developed by CMU ARK Lab. **Dictionary** and **Window** are constructed based on the topic-term distribution obtained from the previous section.

The average precision and recall are shown in Table 4. And as we can observe, the dictionary (with probability) contributes a lot to the performance and by taking into account a more comprehensive set of information around the key word, classifier on **All** feature setting generate significantly better performance, with 0.382 prevision and 0.48 recall, which is acceptable considering (1) This is is a 43-way classification with much more negative data than positive (2) Some types of events are very close to each other (e.g., Leaving and Vocation). Note that recall is valued more than precision here as false-positive examples will be further screened in self-reported information identification process in the following section.

Feature Setting	Precision	Recall
Word+NER	0.204	0.326
Word+NER+Dictionary	0.362	0.433
All	0.382	0.487

Table 4: Average Performance of Multi-Class Classifier on Different Feature Settings. Negative examples (non important event type) are not considered.

5 Self-Reported Information Identification

Although a message might refer to a topic corresponding to a life event such as marriage, the event still might be one in which the speaker is not directly involved. In this section we describe the self reported event identification portion of our pipeline, which takes output from Section 4 and further identifies whether each tweet refers to an event directly involving the user who publishes it.

Direct labeling of randomly sampled Twitter messages is infeasible for the following reasons: (1) Class imbalance: self-reported events are relatively rare in randomly sampled Twitter messages. (2) A large proportion of self-reported information refers to mundane, everyday topics (e.g., "I just finished dinner!"). Fortunately, many of the tweets retrieved from Section 3 consist of self-reported information and describe major life events. The candidates for annotation are therefore largely narrowed down.

We manually annotated 800 positive examples of self-reported events distributed across the event categories identified in Section 3. We ensured good coverage by first randomly sampling 10 examples from each category, the remainder were sampled from the class distribution in the data. Negative examples of self-reported information consisted of a combination of examples from the original dataset[8] and randomly sampled messages gathered by searching for the top terms in each of the pre-identified topics using the Twitter Search interface [9]. Due to great varieties of negative scenarios, the negative dataset constitutes about 2500 tweets.

5.1 Features

Identifying self-reported tweet requires sophisticated feature engineering. Let u denote the term within the tweet that gets the highest possibility generated by the correspondent topic. We experimented with combinations of the following types of features (results are presented in Table **??**):

- **Bigram**: Bigrams within each tweet (punctuation included).
- **Window**: A window of $k \in \{0, 1, 2\}$ words adjacent to u and their part-of-speech tags.
- **Tense**: A binary feature indicating past tense identified in by the presence of past tense verb (VBD).
- **Factuality**: Factuality denotes whether one expression is presented as corresponding to real situations in the world (Saurí and Pustejovsky, 2007). We use Stanford PragBank[10],

[8] Most tweets in the bootstrapping output are positive.

[9] The majority of results returned by Twitter Search are negative examples.

[10] http://compprag.christopherpotts.net/factbank.html

an extension of FactBank (Saurí and Pustejovsky, 2009) which contains a list of modal words such as "might", "will", "want to" etc[11].

- **I**: Whether the subject of the tweet is first person singular.
- **Dependency**: If the subject is first person singular and the u is a verb, the dependency path between the subject and u (or non-dependency).

Tweet dependency paths were obtained from (Lingpeng Kong and Smith, 2014). As the tweet parser we use only supports one-to-one dependency path identification but no dependency properties, **Dependency** is a binary feature. The subject of each tweet is determined by the dependency link to the root of the tweet from the parser.

Among the features we explore, **Word** encodes the general information within the tweet. **Window** addresses the information around topic key word. The rest of the features specifically address each of the negative situations described in Challenge 2, Section 1: **Tense** captures past event description, **Factuality** filters out wishes or imagination, **I** and **Dependency** correspond to whether the described event involves the speaker. We built a linear SVM classifier using SVM$_{light}$ package (Joachims, 1999).

5.2 Evaluation

Feature Setting	Acc	Pre	Rec
Bigram+Window	0.76	0.47	0.44
Bigram+Window +Tense+Factuality	0.77	0.47	0.46
all	0.82	0.51	0.48

Table 5: Performance for self-report information identification regarding different feature settings.

We report performance on the task of identifying self-reported information in this subsection. We employ 5-fold cross validation and report Accuracy (Accu), Prevision (Prec) and Recall (Rec) regarding different feature settings. The **Tense**, **Factuality**, **I** and **Dependency** features positively contribute to performance respectively and the best performance is obtained when all types of features are included.

[11]Due to the colloquial property of tweets, we also introduced terms such as "gonna", "wanna", "bona".

precision	recall	F1
0.82	0.86	0.84

Table 7: Performance for identifying properties.

6 Event Property Extraction

Thus far we have described how to automatically identify tweets referring to major life events. In addition, it is desirable to extract important properties of the event, for example the name of the university the speaker was admitted to (See Figure 1). In this section we take a supervised approach to event property extraction, based on manually annotated data for a handfull of the major life event categories automatically identified by our system. While this approach is unlikely to scale to the diversity of important personal events Twitter users are discussing, our experiments demonstrate that event property extraction is indeed feasible.

We cast the problem of event property extraction as a sequence labeling task, using Conditional Random Fields (Lafferty et al., 2001) for learning and inference. To make best use of the labeled data, we trained a unified CRF model for closely related event categories which often share properties; the full list is presented in Table 6 and we labeled 300 tweets in total. Features we used include:

- word token, capitalization, POS
- left and right context words within a window of 3 and the correspondent part-of-speech tags
- word shape, NER
- a gazetteer of universities and employers borrowed from NELL[12].

We use 5-fold cross-validation and report results in Table 7.

7 End-to-End Experiment

The evaluation for each part of our system has been demonstrated in the corresponding section. We now present a real-world evaluation: to what degree can our trained system automatically identify life events in real world.

7.1 Dataset

We constructed a gold-standard life event dataset using annotators from Amazon's Mechanical Turk (Snow et al., 2008) using 2 approaches:

[12]http://rtw.ml.cmu.edu/rtw/kbbrowser/

Life Event	Property
(a) Acceptance, Graduation	Name of University/College
(b) Wedding, Engagement, Falling love	Name of Spouse/ partner/ bf/ gf
(c) Getting a job, interview, internship	Name of Enterprise
(d) Moving to New Places, Trip, Vocation, Leaving	Place, Origin, Destination
(e) Winning Award	Name of Award, Prize

Table 6: Labeling Event Property.

- Ask Twitter users to label their own tweets (Participants include friends, colleagues of the authors and Turkers from Amazon Mechanical Turk[13]).
- Ask Turkers to label other people's tweets.

For option 1, we asked participants to directly label their own published tweets. For option 2, for each tweet, we employed 2 Turkers. Due to the ambiguity in defining life events, the value cohen's kappa[14] as a measure of inter-rater agreement is 0.54; this does not show significant inter-annotator agreement. The authors examined disagreements and also verified all positively labeled tweets. The resulting dataset contains around 900 positive tweets and about 60,000 negative tweets.

To demonstrate the advantage of leveraging large quantities of unlabeled data, the first baseline we investigate is a **Supervised** model which is trained on the manually annotated labeled dataset, and evaluated using 5 fold cross validation. Our **Supervised** baseline consists of a linear SVM classifier using bag of words, NER and POS features. We also tested a second baseline that combines **Supervised** algorithm with an our self-reported information classifier, denoted as **Supervised+Self**.

Results are reported in Table 8; as we can observe, the fully supervised approach is not suitable for this task with only one digit F1 score. The explanations are as follows: (1) the labeled data can only cover a small proportion of life events (2) supervised learning does not separate important event categories and will therefore classify any tweet with highly weighted features (e.g., the mention of "I" or "marriage") as positive. By using an additional self-reported information classifier in **Supervised+Self**, we get a significant boost in precision with a minor recall loss.

[13] https://www.mturk.com/mturk/welcome

[14] http://en.wikipedia.org/wiki/Cohen's_kappa

Approach	Precision	Recall
Our approach	0.62	0.48
Supervised	0.13	0.20
Supervised+Self	0.25	0.18

Table 8: Performance for different approaches for identifying life events in real world.

Approach	Precision	Recall
Step 1	0.65	0.36
Step 2	0.64	0.43
Step 3	0.62	0.48

Table 9: Performance for different steps of bootstrapping for identifying life events in real world.

Another interesting question is to what degree the bootstrapping contributes to the final results. We keep the self-reported information classifier fixed (though it's based the ultimate identified data source), and train the personal event classifier based on topic distributions identified from each of the three steps of bootstrapping[15]. Precision and recall at various stages of bootstrapping are presented in Table 9. As bootstrapping continues, the precision remains roughly constant, but recall increases as more life events and CONGRATULATIONS and CONDOLENCES are discovered.

8 Related Work

Our work is related to three lines of NLP researches. (1) user-level information extraction on social media (2) public event extraction on social media. (3) Data harvesting in Information Extraction, each of which contains large amount of related work, to which we can not do fully justice.

User Information Extraction from Twitter Some early approaches towards understanding user level information on social media is focused on user profile/attribute prediction (e.g.,(Ciot et al., 2013)) user-specific content extraction (Diao

[15] which are 24, 38, 42-class classifiers, where 24, 38, 42 denoted the number of topics discovered in each step of bootstrapping (see Figure 5).

et al., 2012; Diao and Jiang, 2013; Li et al., 2014) or user personalization (Low et al., 2011) identification.

The problem of user life event extraction was first studied by Li and Cardie's (2014). They attempted to construct a chronological timeline for Twitter users from their published tweets based on two criterion: a personal event should be personal and time-specific. Their system does not explicitly identify a global category of life events (and tweets discussing correspondent event) but identifies the topics/events that are personal and time-specific to a given user using an unsupervised approach, which helps them avoids the nuisance of explicit definition for life event characteristics and acquisition of labeled data. However, their system has the short-coming that each personal topic needs to be adequately discussed by the user and their followers in order to be detected[16].

Public Event Extraction from Twitter Twitter serves as a good source for event detection owing to its real time nature and large number of users. These approaches include identifying bursty public topics (e.g.,(Diao et al., 2012)), topic evolution (Becker et al., 2011) or disaster outbreak (Sakaki et al., 2010; Li and Cardie, 2013) by spotting the increase/decrease of word frequency. Some other approaches are focused on generating a structured representation of events (Ritter et al., 2012; Benson et al., 2011).

Data Acquisition in Information Extraction Our work is also related with semi-supervised data harvesting approaches, the key idea of which is that some patterns are learned based on seeds. They are then used to find additional terms, which are subsequently used as new seeds in the patterns to search for additional new patterns (Kozareva and Hovy, 2010b; Davidov et al., 2007; Riloff et al., 1999; Igo and Riloff, 2009; Kozareva et al., 2008). Also related approaches are distant or weakly supervision (Mintz et al., 2009; Craven et al., 1999; Hoffmann et al., 2011) that rely on available structured data sources as a weak source of supervision for pattern extraction from related text corpora.

[16]The reason is that topic models use word frequency for topic modeling.

9 Conclusion and Discussion

In this paper, we propose a pipelined system for major life event extraction from Twitter. Experimental results show that our model is able to extract a wide variety of major life events.

The key strategy adopted in this work is to obtain a relatively clean training dataset from large quantity of Twitter data by relying on minimum efforts of human supervision, and sometimes is at the sacrifice of recall. To achieve this goal, we rely on a couple of restrictions and manual screenings, such as relying on replies, LDA topic identification and seed screening. Each part of system depends on the early steps. For example, topic clustering in Section 3 not only offers training data for event identification in Section 4, but prepares the training data for self-information identification in Section 5. .

We acknowledge that our approach is not perfect due to the following ways: (1) The system is only capable of discovering a few categories of life events with many others left unidentified. (2) Each step of the system will induce errors and negatively affected the following parts. (3) Some parts of evaluations are not comprehensive due to the lack of gold-standard data. (4) Among all pipelines, event property identification in Section 6 still requires full supervision in CRF model, making it hard to scale to every event type[17]. How to address these aspects and generate a more accurate, comprehensive and fine-grained life event list for Twitter users constitute our further work.

Acknowledgement A special thanks is owned to Myle Ott for suggestions on bootstrapping procedure in data harvesting. The authors want to thank Noah Smith, Chris Dyer and Alok Kothari for useful comments, discussions and suggestions regarding different steps of the system and evaluations. We thank Lingpeng Kong and members of Noah's ARK group at CMU for providing the tweet dependency parser. All data used in this work is extracted from CMU Twitter Warehouse maintained by Brendan O'Connor, to whom we want to express our gratitude.

References

Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond trending topics: Real-world event identifi-

[17]We view weakly supervised life event property extraction as an interesting direction for future work.

cation on twitter. *ICWSM*, 11:438–441.

Cosmin Adrian Bejan, Matthew Titsworth, Andrew Hickl, and Sanda M Harabagiu. 2009. Nonparametric bayesian models for unsupervised event coreference resolution. In *NIPS*, pages 73–81.

Edward Benson, Aria Haghighi, and Regina Barzilay. 2011. Event discovery in social media feeds. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 389–398. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-english contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash*, pages 18–21.

Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.

Dmitry Davidov, Ari Rappoport, and Moshe Koppel. 2007. Fully unsupervised discovery of concept-specific relationships by web mining. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 232.

Qiming Diao and Jing Jiang. 2013. A unified model for topics, events and users on twitter. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1869–1879.

Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. 2012. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 536–544. Association for Computational Linguistics.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In *International Conference on Artificial Intelligence and Statistics*, pages 163–170.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.

Sean P Igo and Ellen Riloff. 2009. Corpus-based semantic lexicon induction with web-based corroboration. In *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, pages 18–26. Association for Computational Linguistics.

Thorsten Joachims. 1999. Making large scale svm learning practical.

Zornitsa Kozareva and Eduard Hovy. 2010a. Learning arguments and supertypes of semantic relations using recursive patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1482–1491. Association for Computational Linguistics.

Zornitsa Kozareva and Eduard Hovy. 2010b. Not all seeds are equal: Measuring the quality of text mining seeds. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 618–626. Association for Computational Linguistics.

Zornitsa Kozareva, Ellen Riloff, and Eduard H Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In *ACL*, volume 8, pages 1048–1056.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Jiwei Li and Claire Cardie. 2013. Early stage influenza detection from twitter. *arXiv preprint arXiv:1309.7340*.

Jiwei Li and Claire Cardie. 2014. Timeline generation: Tracking individuals on twitter. *WWW, 2014*.

Jiwei Li, Alan Ritter, and Eduard Hovy. 2014. Weakly supervised user profile extraction from twitter. ACL.

Swabha Swayamdipta Archna Bhatia Chris Dyer Lingpeng Kong, Nathan Schneider and Noah Smith. 2014. A dependency parser for tweets. In *EMNLP*.

Yucheng Low, Deepak Agarwal, and Alexander J Smola. 2011. Multiple domain user personalization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 123–131. ACM.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the*

Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, pages 1003–1011. Association for Computational Linguistics.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390.

Ellen Riloff, Rosie Jones, et al. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, pages 474–479.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM.

Kirk Roberts and Sanda M Harabagiu. 2011. Unsupervised learning of selectional restrictions and detection of argument coercions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 980–990. Association for Computational Linguistics.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Roser Saurí and James Pustejovsky. 2007. Determining modality and factuality for text entailment. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 509–516. IEEE.

Roser Saurí and James Pustejovsky. 2009. Factbank: A corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227–268.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1456–1466. Association for Computational Linguistics.