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

Jiwei Li$^1$, Alan Ritter$^2$, Claire Cardie$^3$ and Eduard Hovy$^4$

Department of Computer Science
1Stanford University 2Ohio State University
3Cornell University 4Carnegie Mellon University

June 22nd, 2014
Life Events

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

Jiwei Li¹, Alan Ritter², Claire Cardie³ and Eduard Hovy⁴
Life Events on Social Media

Jiwei Li, Alan Ritter, Claire Cardie and Eduard Hovy

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

Jessica Jones @jonesalgebra · Sep 27
We're engaged!!! I could not be more thrilled! We are getting married June 12, 2015!
Life Events on Social Media

Jessica Jones @jonesalgebra · Sep 27
We’re engaged!!!! I could not be more thrilled! We are getting married June 12, 2015!

$002$ @susiezenanrio · Dec 17
Haha love school: I just got accepted by Harvard
Life Events on Social Media

Challenges
Response based Data Harvesting
System Overview
Algorithms
Experiments
Conclusion

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

Jiwei Li, Alan Ritter, Claire Cardie and Eduard Hovy
Life Events on Social Media
Life Events on Social Media

accepted to MIT. no words can describe how happy i am. guess hard work really does pay off.

Life Event: University Admission
Event Property (University): MIT
Life Events on Social Media

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

- Life Event: University Admission
  - Event Property (University): MIT

- Life Event: Engagement
  - Event Property (Engaged to): kyloatoast
Life Events on Social Media

- **Life Event**: University Admission
  - Event Property (University): MIT

- **Life Event**: Engagement
  - Event Property (Engaged to): kyloatoast

- **Life Event**: Receiving Award
  - Event Property (From): Norway-America Association
Life Events on Social Media

Why?
Life Events on Social Media

Why?

- Better understanding of users
Life Events on Social Media

Why?

- Better understanding of users
- Friend Recommendation
Life Events on Social Media

Why?

- Better understanding of users
- Friend Recommendation
- Online advertising
Outline

- Challenges
- System Overview
- Algorithms
- Experiments
- Conclusion
Challenges
Challenge 1: Major life event is an ambiguous concept!
**Challenge 1:** Major life event is an ambiguous concept!
Challenge 1: Major life event is an ambiguous concept!
Challenges

**Challenge 1:** Major life event is an ambiguous concept!
Challenge 1: Major life event is an ambiguous concept!
Challenges

Challenge 1: What are life events?
Challenge 1: What are life events?
Challenge 1: What are life events?
Challenge 1: What are life events?
Challenge 2: Noisy Data
Challenge 2: Noisy Data
Challenge 2: Noisy Data
Challenge 2: Noisy Data
Challenges

Challenge 2: Noisy Data

Love Quotes @LoveQuotes - 21h
I want to get married once. No divorce & no cheating, just us two till the end.

Random Imagination/ Wish

Marquita Brown @mbrownNR - 25m
I'm at the #GSO register of deeds office. Two couples are here to get married.

Some other guys
Challenge 2: Noisy Data

Retweeted 618 times

Love Quotes @LoveQuotes · 21h
I want to get married once. No divorce & no cheating, just us two till the end.

Random Imagination/ Wish

Marquita Brown @mbrownNR · 25m
I'm at the #GSO register of deeds office. Two couples are here to get married.

Some other guys

Single Dad @Lonely_Dad · Oct 7
my dreams died when I got married.
past tense
Challenge 3: Lack of labeled data
Challenge 3: Lack of labeled data

- No labeling criteria
Challenges

Challenge 3: Lack of labeled data

- No labeling criteria
- Life events sparsely distributed
Challenges

**Challenge 3**: Lack of labeled data

- No labeling criteria
- Life events sparsely distributed
- Rare events
Challenges

HOW ??
Challenges

Jiwei Li\textsuperscript{1}, Alan Ritter\textsuperscript{2}, Claire Cardie\textsuperscript{3} and Eduard Hovy\textsuperscript{4}

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

I say

I got accepted by Harvard!!

What you would say?
Challenges

I say

I got accepted by Harvard!!

Congratulations!
Challenges

Congratulations!
great!
Fantastic!
Awesome

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

Jiwei Li¹, Alan Ritter², Claire Cardie³ and Eduard Hovy⁴
Challenges

Congratulations!
great!
Fantastic!
Awesome
I'm so sorry to hear that.

"THAT'S TERRIBLE"
Responses based Data Harvesting

Seeds:
congrats, fantastic, cool, ....
Responses based Data Harvesting

Seeds: 
congrats, fantastic, cool, ....

collect

Conversation Text
Responses based Data Harvesting

Seeds: congrats, fantastic, cool, ....

collect

Conversation Text

LDA

Word Clusters (topics)
Responses based Data Harvesting

Seeds: congrats, fantastic, cool, ....

collect

Conversation Text

LDA

Word Clusters (topics)

manual identification

Meaningful Word Clusters
Responses based Data Harvesting

- Seeds: congrats, fantastic, cool, ....
- Collect
- Conversation Text
- LDA
- Word Clusters (topics)
- Manual identification
- Semi-supervised Data harvesting
- Meaningful Word Clusters

Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy

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

Semi-supervised Data harvesting

(Kozareva and Hovy, 2010;
Davidov et al, 2007;
Igo and Riloff, 2009)
Responses based Data Harvesting

(Kozareva and Hovy, 2010; Davidov et al, 2007; Igo and Riloff, 2009)

Semi-supervised Data harvesting

Stream-LDA
(Yao et al, 2009)

More Texts

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

Semi-supervised Data harvesting

(Kozareva and Hovy, 2010; Davidov et al, 2007; Igo and Riloff, 2009)

Stream-LDA

(Yao et al, 2009)

More Texts

collect

More Expression Seeds
Responses based Data Harvesting

Semi-supervised Data harvesting
(Yao et al, 2009)

Stream-LDA

More Texts

collect

More Expression Seeds

More Texts

(Kozareva and Hovy, 2010; Davidov et al, 2007; Igo and Riloff, 2009)
Responses based Data Harvesting

Semi-supervised Data harvesting
(Yao et al, 2009)

Stream-LDA

More Texts

collect

More Expression Seeds

Word Clusters

LDA

Manual identification

More Texts

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

Semi-supervised Data harvesting
(Kozareva and Hovy, 2010; Davidov et al, 2007; Igo and Riloff, 2009)

Stream-LDA
(Yao et al, 2009)

More Texts

collect

More Expression Seeds

Word Clusters

LDA
Manual identification

More Texts
Responses based Data Harvesting

![Graph showing data retrieval over bootstrapping steps](image)

- **Num of Data Retrieved**
- **Num of Bootstrapping**

Legend:
- replies
- topics
- tweet *10^4

---

Jiwei Li\(^1\), Alan Ritter\(^2\), Claire Cardie\(^3\) and Eduard Hovy\(^4\)

Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts
### Challenges
- Response based Data Harvesting

### System Overview
- Algorithms
- Experiments
- Conclusion

## Responses based Data Harvesting

| Life Event | Proportion |
|------------|------------|
| Birthday   | 9.78       |
| Job        | 8.39       |
| Wedding    | 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  | 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       |

### Table 1: List of automatically discovered life event types.

Jiwei Li\(^1\), Alan Ritter\(^2\), Claire Cardie\(^3\) and Eduard Hovy\(^4\)

Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts
## Responses based Data Harvesting

| Human Label     | Top words                                                                 |
|-----------------|---------------------------------------------------------------------------|
| Wedding         | wedding, love, ring, engagement, engaged, bride, video, marrying          |
| Graduation      | graduation, school, college, graduate, graduating, year, grad             |
| Relationship    | boyfriend, girlfriend, date, check, relationship, see, look               |
| Anniversary     | anniversary, years, year, married, celebrating, wife, celebrate, love     |
| 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       | 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        | vacation, family, trip, country, go, flying, visited, holiday, Hawaii    |
| Winning Award   | won, award, support, awards, winning, honor, scholarship, prize           |
| Election        | 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, 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           |

**Table 2**: Example event types with top words discovered by our model.
System Overview

Challenges Response based Data Harvesting System Overview Algorithms Experiments Conclusion

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

System Overview

- **Input:** 'I had beef jerky for lunch'
  - **Output:** 'I got married to Tom', 'My friend Chris got married'

- **Pipeline 1:** Personal Life Event Identification
  - **Input:** Tweets: 'I had beef jerky for lunch', 'I got married to Tom', 'My friend Chris got married'
  - **Output:** 'I got married to Tom', 'Event Category: marriage', 'My friend Chris got married', 'Event Category: marriage'

- **Pipeline 2:** Self-reported Information Identification
  - **Input:** 'My friend Chris got married'
  - **Output:** 'My friend Chris got married'

- **Pipeline 3:** Event Property Extraction
  - **Input:** Tweets: 'I got married to Tom', 'Event Category: marriage'
  - **Output:** 'I got married to Tom', 'Event Category: marriage'

- **Output:** 'I had beef jerky for lunch', 'throw away'
System Overview

Challenges
Response based Data Harvesting
System Overview
Algorithms
Experiments
Conclusion

System Overview

Jiwei Li¹, Alan Ritter², Claire Cardie³ and Eduard Hovy⁴

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

 Pipeline 1: Personal Life Event Identification

 tweets:
I had beef jerky for lunch.
I got married to Tom
My friend Chris got married.

I got married to Tom
Event Category: marriage
My friend Chris got married.
Event Category: marriage

 Pipeline 2: Self-reported Information Identification

 I got married to Tom
Event Category: marriage

 My friend Chris got married
 throw away

 Pipeline 3: Event Property Extraction

 I got married to Tom
Event Category: marriage
Married to (property): Tom

 My friend Chris got married
 throw away

Jiwei Li¹, Alan Ritter², Claire Cardie³ and Eduard Hovy⁴

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

**Challenges**

**Response based Data Harvesting**

**System Overview**

**Algorithms**

**Experiments**

**Conclusion**

**System Overview**

Jiwei Li\(^1\), Alan Ritter\(^2\), Claire Cardie\(^3\) and Eduard Hovy\(^4\)

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

Challenges Response based Data Harvesting System Overview Algorithms Experiments Conclusion

Jiwei Li¹, Alan Ritter², Claire Cardie³ and Eduard Hovy⁴

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

Pipeline 1: Personal Life Event Identification

I had beef jerky for lunch

I got married to Tom
Event Category: marriage
My friend Chris got married.
Event Category: marriage

Pipeline 2: Self-reported Information Identification

I got married to Tom
Event Category: marriage

Pipeline 3: Event Property Extraction

tweets:
I had beef jerky for lunch.
I got married to Tom
My friend Chris got married.

I got married to Tom
Event Category: marriage
Married to (property): Tom

My friend Chris got married

throw away

input

output
Personal Event Identification

Multi-Class Classifier based on SVM
Personal Event Identification

Multi-Class Classifier based on SVM
Positive Examples for each category: Pre-identified data
Personal Event Identification

Multi-Class Classifier based on SVM
Positive Examples for each category: Pre-identified data
Negative Examples: Random Tweets
Personal Event Identification

Multi-Class Classifier based on SVM
Positive Examples for each category: Pre-identified data
Negative Examples: Random Tweets
Personal Event Identification

Multi-Class Classifier based on SVM
Positive Examples for each category: Pre-identified data
Negative Examples: Random Tweets
  - Topic-Tweet probability
Personal Event Identification

Multi-Class Classifier based on SVM
Positive Examples for each category: Pre-identified data
Negative Examples: Random Tweets
- Topic-Tweet probability
- Dictionary
Personal Event Identification

Multi-Class Classifier based on SVM
Positive Examples for each category: Pre-identified data
Negative Examples: Random Tweets
- Topic-Tweet probability
- Dictionary
- Word, NER, POS
- Window Context
Personal Event Identification

Multi-Class Classifier based on SVM:

Split harvested data, training and testing

| Feature Setting               | Precision | Recall |
|-------------------------------|-----------|--------|
| Word+NER                      | 0.204     | 0.326  |
| Word+NER+Dictionary           | 0.362     | 0.433  |
| All                           | 0.382     | 0.487  |
Self Information Identification

- **Input**: I had beef jerky for lunch
- **Output**: I got married to Tom
  - Event Category: marriage
  - Event Property: Tom

**Pipelines**

1. **Pipeline 1**: Personal Life Event Identification
   - Input: I got married to Tom
   - Output: My friend Chris got married
     - Event Category: marriage

2. **Pipeline 2**: Self-reported Information Identification
   - Input: I got married to Tom
   - Output: My friend Chris got married
     - Event Category: marriage

3. **Pipeline 3**: Event Property Extraction
   - Input: I got married to Tom
   - Output: Married to (property): Tom
Self Information Identification

Negative Examples
Self Information Identification

Negative Examples

- Not self
Self Information Identification

Negative Examples

- Not self
- Random Thought
Self Information Identification

Negative Examples

- Not self
- Random Thought
- Past Tense
Self Information Identification

Dataset:
Positive: selected from harvested data
Negative: selected from harvested data
Self Information Identification

**Dataset:**
Positive: selected from harvested data
Negative: selected from harvested data

**Binary SVM Classifier**
Self Information Identification

**Dataset:**
Positive: selected from harvested data
Negative: selected from harvested data

**Binary SVM Classifier**
- Tense
Self Information Identification

**Dataset:**
Positive: selected from harvested data
Negative: selected from harvested data

**Binary SVM Classifier**
- Tense
- Factuality (could, would, can ... ) (Saurf and Pustejovsky, 2007)
Self Information Identification

**Dataset:**
Positive: selected from harvested data  
Negative: selected from harvested data

**Binary SVM Classifier**
- Tense
- Factuality (could, would, can ... ) (Saurf and Pustejovsky, 2007)
- I
Self Information Identification

**Dataset:**
Positive: selected from harvested data  
Negative: selected from harvested data

**Binary SVM Classifier**
- Tense
- Factuality (could, would, can ... ) (Saurf and Pustejovsky, 2007)
- I
- Dependency (Kong et al., 2014)
Self Information Identification

**Dataset:**
Positive: selected from harvested data
Negative: selected from harvested data

**Binary SVM Classifier**
- Tense
- Factuality (could, would, can ... ) (Saurf and Pustejovský, 2007)
- I
- Dependency (Kong et al., 2014)
- Token, NER, POS, window context
Self Information Identification

| 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|
Event Property Identification

- **Pipeline 1:** Personal Life Event Identification
  - Input: Tweets
  - Output: Event: I got married to Tom
  - Event Category: marriage
  - Married to (property): Tom

- **Pipeline 2:** Self-reported Information Identification
  - Input: Tweets
  - Output: Event: My friend Chris got married
  - Event Category: marriage

- **Pipeline 3:** Event Property Extraction
  - Input: Tweets
  - Output: Event: I had beef jerky for lunch
  - Event Category: not identified

The process involves identifying events and properties from tweets, categorizing events, and extracting properties related to the events.
### Human Labeling

| 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                          |
Event Property Identification

**Sequence Labeling Task, CRF** (Lafferty, et al., 2001)
- Word token, Capitalization, POS, word shape, NER
- A gazetteer of universities and companies
- Context
What benefits brought from Congratulations/Condolences Speech Acts?

- Clean Data
What benefits brought from Congratulations/Condolences Speech Acts?

- Clean Data
- Personal Topic Identification
System

What benefits brought from Congratulations/Condolences Speech Acts?

- Clean Data
  - Personal Topic Identification
  - Self Report Information
What benefits brought from Congratulations/Condolences Speech Acts?

- Clean Data
  - Personal Topic Identification
  - Self Report Information
    User 1: I wish to get married
    User 2: Congratulations!!
Experiments

- End-to-End Experiments
Experiments

Gold-standard life event dataset
Gold-standard life event dataset

- Ask Twitter users to label their own tweets
Experiments

Gold-standard life event dataset

- Ask Twitter users to label their own tweets
- Ask Turkers to label other people’s tweets.
Experiments

Gold-standard life event dataset

- Ask Twitter users to label their own tweets
- Ask Turkers to label other people’s tweets.
  - 2 Turkers 1 tweet

Inter-rater agreement is 0.54 (Cohen's kappa)
Authors make final decision

900 positive tweets
60,000 negative tweets
Experiments

Gold-standard life event dataset

- Ask Twitter users to label their own tweets
- Ask Turkers to label other people’s tweets.
  - 2 Turkers 1 tweet
  - Inter-rater agreement is 0.54 (cohen’s kappa)
Gold-standard life event dataset

- Ask Twitter users to label their own tweets
- Ask Turkers to label other people’s tweets.
  - 2 Turkers 1 tweet
  - Inter-rater agreement is 0.54 (cohen’s kappa)
  - Authors make final decision
Experiments

Gold-standard life event dataset

- Ask Twitter users to label their own tweets
- Ask Turkers to label other people’s tweets.
  - 2 Turkers 1 tweet
  - Inter-rater agreement is 0.54 (cohen’s kappa)
  - Authors make final decision
- 900 positive tweets
Experiments

Gold-standard life event dataset

- Ask Twitter users to label their own tweets
- Ask Turkers to label other people’s tweets.
  - 2 Turkers 1 tweet
  - Inter-rater agreement is 0.54 (cohen’s kappa)
  - Authors make final decision
- 900 positive tweets
- 60,000 negative tweets
Experiments

Baselines
Experiments

Baselines

- Supervised

Table 3: Performance for different approaches for identifying life events.
Experiments

Baselines

- Supervised
- Supervised + Self

Table 3: Performance for different approaches for identifying life events.
Experiments

Baselines

- Supervised
- Supervised + Self

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

Table 3: Performance for different approaches for identifying life events.
Experiments

Does bootstrapping help?
Does bootstrapping help?

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

Table 4: Performance for different steps of bootstrapping for identifying.
Conclusion
We study the life event extraction problem on Twitter. We propose a framework based on Congratulations/Condolences Speech Acts for data harvesting. We explore different types of features and algorithms for this task.
We study the life event extraction problem on Twitter.
Conclusion

- We study the life event extraction problem on Twitter
- We propose a framework based on Congratulations/Condolences Speech Acts for data harvesting
We study the life event extraction problem on Twitter
We propose a framework based on Congratulations/Condolences Speech Acts for data harvesting
We explore different types features and algorithms for this task
Key idea: solve this problem based on minimum human efforts.
Conclusion

Key idea: solve this problem based on minimum human efforts.

Problems
Conclusion

Key idea: solve this problem based on minimum human efforts.

Problems

- Restricted to event types identified by Congratulations/Condolences Speech Acts.
Conclusion

Key idea: solve this problem based on minimum human efforts.

Problems

- Restricted to event types identified by Congratulations/Condolences Speech Acts.
- No all responses correspond to life events
Conclusion

Key idea: solve this problem based on minimum human efforts.

Problems

- Restricted to event types identified by Congratulations/Condolences Speech Acts.
- No all responses correspond to life events
- Error accumulations.
Thank you!
Thank you!

Questions, Suggestions
Thank you!

Questions, Suggestions

Joint work with

Alan Ritter  Claire Cardie  Eduard Hovy