Extracting a Repository of Events and Event References from News Clusters

Presented by: Silvia Julinda
Thesis Advisor: Alan Akbik
Thesis Supervisor: Prof. Dr. Volker Markl

Barcelona, 04 September 2014

Fachgebiet Datenbanksysteme und Informationsmanagement
Technische Universität Berlin

http://www.dima.tu-berlin.de/
"An event instance is something that happens at a specific time and place“ (Topic Detection and Tracking, 2002)

**Event**

**Topic**

**Airplane crash**

USAir 427 crash  
on 08 September 1994  
in Pittsburg
Identification in text, of specified classes of:
- Names / Entities
- Relations
- Events

Semi-structured text: typical of web pages, includes markup and limited free text (e.g., XML)

Unstructured text: extended text with little markup, such as news stories.
### How is event extraction done?

- **Define Event Templates**
  
  **ACE: 33 types of events**
  
  **Personnel_End-Position**
  
  - Event type
  - Subtype
  - Person
  - Organization
  - Position
  - Time-within

  *(Jiand Grishman, 2008)*

  **Attack_Target**
  
  - Event type
  - Subtype
  - Attacker
  - Target
  - Weapon
  - Time
  - Place
  - Comment

  *(Aone and Ramos-Santacruz, 2000)*

- **Hard to scale-up** because it is hand-crafted.

- **Mostly domain dependent** (e.g., biomedical events, infectious disease outbreaks, and violent and natural disaster events).
Open Domain Event Extraction from Twitter

- **TWICAL**
  - Apple to *Announce* iPhone 5 on October 4th?! YES!
  - iPhone 5 *announcement* coming Oct 4th
  - WOOOHOO *NEW* IPHONE TODAY! CAN’T WAIT!

4-tuple representation:

<table>
<thead>
<tr>
<th>Entity</th>
<th>Event Phrase</th>
<th>Date</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steve Jobs</td>
<td>died</td>
<td>10/6/11</td>
<td>DEATH</td>
</tr>
<tr>
<td>iPhone</td>
<td>announcement</td>
<td>10/4/11</td>
<td>PRODUCTLAUNCH</td>
</tr>
<tr>
<td>GOP</td>
<td>debate</td>
<td>9/7/11</td>
<td>POLITICALEVENT</td>
</tr>
<tr>
<td>Amanda Knox</td>
<td>verdict</td>
<td>10/3/11</td>
<td>TRIAL</td>
</tr>
</tbody>
</table>

(Ritter et al, 2012)

- Huge amount of labeled data requires human effort and labor.
Aim: extract unrestricted set of events and all their possible event mentions

Facebook, on Thursday, agreed to buy mobile messaging service WhatsApp for $19 billion,

MENLO PARK, CALIF. – February 19, 2014 – Facebook today announced that it has reached a definitive agreement to acquire WhatsApp, a rapidly growing cross-platform mobile messaging

<table>
<thead>
<tr>
<th>Id</th>
<th>Timestamp</th>
<th>Place</th>
<th>Representative</th>
<th>Textual References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2014-02-19</td>
<td>Menlo Park</td>
<td>Facebook buy Whatsapp</td>
<td>social media giant Facebook monopolise messaging market</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Facebook acquire WhatsApp</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>social media company purchase WhatsApp</td>
</tr>
</tbody>
</table>
Each cluster typically represent one news item
Temporal Tagger

- Recognize temporal expressions and normalized to a specific points in time.
- Use article’s publication date as reference point in which the statement was made.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 December 1974</td>
<td>DATE</td>
<td>1974-12-12</td>
</tr>
<tr>
<td>Last Monday</td>
<td>DATE</td>
<td>2014-02-17</td>
</tr>
<tr>
<td>Christmast Eve</td>
<td>DATE</td>
<td>2014-12-24</td>
</tr>
<tr>
<td>Winter of 2014</td>
<td>DATE</td>
<td>2014-WI</td>
</tr>
<tr>
<td>Saturday Morning</td>
<td>TIME</td>
<td>2014-02-22TMO</td>
</tr>
<tr>
<td>7:18 p.m.</td>
<td>TIME</td>
<td>2014-08-21T19:18</td>
</tr>
<tr>
<td>3 days</td>
<td>DURATION</td>
<td>P3D</td>
</tr>
<tr>
<td>from 2010 to 2014</td>
<td>RANGE</td>
<td>2010/2014</td>
</tr>
<tr>
<td>every third Saturday</td>
<td>SET</td>
<td>XXXX-WXX-6</td>
</tr>
</tbody>
</table>
Analyzing Syntactic Structure of Sentence

Dependency Parsing

- Easier to understand
- Faster
- Binds grammaticality to words

Constituent Parsing
"Who did What to Whom, When and Where?"

Mary sold a book to John in the library yesterday

**Predicate**: sell

**Roles**:
- **ARG0**: Seller
- **ARG1**: thing sold
- **ARG2**: instrument/buyer
- **ARGM-LOC**: location modifier
- **ARGM-TMP**: temporal modifier
### Hypothesis: “same cluster + same timestamp = same event”

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-02-19</td>
<td>Facebook inked a deal <strong>late Wednesday</strong> to buy popular texting service WhatsApp.</td>
</tr>
<tr>
<td></td>
<td><strong>Yesterday</strong>, Facebook Chief Executive Officer Mark Zuckerberg bought their five-year-old company.</td>
</tr>
<tr>
<td>2014-07-17</td>
<td><strong>19 February 2014</strong> Facebook Inc will buy fast-growing mobile-messaging startup WhatsApp.</td>
</tr>
<tr>
<td></td>
<td>Malaysia Airlines flight MH17 went down over Ukraine shortly <strong>after 2 PM.</strong></td>
</tr>
<tr>
<td></td>
<td>MH17, flying from Amsterdam to Kuala Lumpur <strong>today</strong>, was shot down over Eastern Ukraine.</td>
</tr>
<tr>
<td></td>
<td>Malaysia Airlines MH17 carrying 298 on board crashed in eastern Ukraine <strong>July 17, 2014.</strong></td>
</tr>
</tbody>
</table>
Overall Framework

1. Identify sentences with temporal expressions
2. Choose unit “day” as the granularity
3. Event Extraction
   - Extract timestamp, location & textual reference
4. Determine Event Representative
   - Choose the most frequent event reference

Diagram:
- Preprocessor
- Split sentences
- Temporal Extraction
- Iterate over sentences
- Dependency Parser
- Semantic Labeler
- Dependency trees + semantic roles
- Event Extraction
- Grouping
- Event references
- Event representative
- News cluster
- Article
- Cluster ID
- Timestamp
- Location
- Textual reference
Event Extraction

- Top arcs: typed dependencies (nsubj, dobj, prep,..)
- Bottom arcs: semantic roles (ARG0, ARG1, ARGM-TMP, ARGM-LOC)

(California) (Facebook buy WhatsApp) (2014-02-19)
Temporal Extraction

- Eliminate Ambiguous Temporal Expression

<table>
<thead>
<tr>
<th>Expression</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A few months ago</td>
<td>2014-05-20</td>
</tr>
<tr>
<td>A few years ago</td>
<td>2013-06-20</td>
</tr>
<tr>
<td>day and night</td>
<td>2014-06-20TNI</td>
</tr>
<tr>
<td>man’s last minute will</td>
<td>2014-06-19TXX</td>
</tr>
<tr>
<td>the same day that Facebook celebrated</td>
<td>2014-06-20</td>
</tr>
</tbody>
</table>

- Using tenses of the verb

Given a reference date of "2014-02-23", a Monday, "Wednesday" → last Wednesday or next Wednesday?

"Facebook bought ... on Wednesday"  
VBD  
(verb, past tense)  

2014-02-19  
last Wednesday
Using dependent clause

- ARGM-TMP argument is a dependent clause (contain subject and predicate)

"She **lost** her confidence after meeting Li Na, who hammered Flavia Pennetta 6-2 on Wednesday"

**Predicate**: lost
**ARGM-TMP**: after meeting Li Na, who **hammered** Flavia Pennetta 6-2 on **Wednesday**

"**lost**" : Main clause
"**Wednesday**" : SBAR

removed!
Preliminary Inspection of Result

Hypothesis
“same cluster + same timestamp = same event” fails in some cases!

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Timestamp</th>
<th>Textual Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>43988874</td>
<td>2014-03-08</td>
<td>Malaysian Flight MH370 lose contact</td>
</tr>
<tr>
<td></td>
<td>2014-03-27</td>
<td>Thailand spot 300 floating objects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thaichote satellite acquire imagery of flotsam</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thai satellite detect 300 objects floating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bangladesh join international search operations</td>
</tr>
</tbody>
</table>
Improving Event Quality

- **Time Window Filter**
  - Filter out events occur within 2 days of the article’s publishing date.

- **Word Alignment Condition**
  - Determine the “similarity” of two event references.

---

(Yao et al., 2013)

2014-02-19

Facebook buy mobile messaging service
Facebook acquire Whatsapp
Facebook ink deal

venture capital firm hit 3 billion jackpot company IPO hit jackpot
Evaluation Metrics

- Precision: fraction of extracted event references that are correct.

- Recall: fraction of correct event references that are extracted.

- Purity: fraction of the biggest group of event references per detected event that refer to the same actual event.

Purity = 1.0

Purity = 3/4 = 0.75
Evaluated in 4 setups on a sample of 40 news clusters with an average size of 300 news articles

- BASE: Only the “same cluster + same timestamp = same event” hypothesis.
- TIME: BASE + the time window filter.
- ALIGN: BASE + the word alignment condition.
- ALL: BASE + the time window filter and the word alignment condition.
Hypothesis produces promising results with a precision of 0.84.

Time Window Filter provides no significant contribution but greatly reduces number of extracted events references.

Word Alignment Condition increase both precision and purity significantly.
■ Evaluated on 202 news articles by obtaining all sentences that contain temporal expressions.
■ 446 manually extracted events, a total of 229 events were found by our system.

<table>
<thead>
<tr>
<th>Category</th>
<th># Events</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun phrase</td>
<td>21</td>
<td>4.72 %</td>
</tr>
<tr>
<td>Coreference</td>
<td>37</td>
<td>8.29 %</td>
</tr>
<tr>
<td>Common predicate</td>
<td>77</td>
<td>17.27 %</td>
</tr>
<tr>
<td>Insufficient semantic roles</td>
<td>82</td>
<td>18.38 %</td>
</tr>
<tr>
<td><strong>True positives</strong></td>
<td><strong>229</strong></td>
<td><strong>51.34 %</strong></td>
</tr>
</tbody>
</table>
Error Analysis

Source of Errors

- **Conditional sentence**: 4%
- **Fact error**: 16%
- **Place**: 12%
- **News Cluster**: 18%
- **Time**: 50%

**If construction:**

win against the Dutch that there “will be a war” if Argentina beats Germany on July 13.

**Incorrect time mention:**

NEW DELHI: Facebook, on Thursday, agreed to buy mobile messaging service WhatsApp for $19.

Facebook announced the purchase of the mobile messaging WhatsApp on Wednesday, in a $19bn deal that represents...

**Ambiguous temporal expression:**

... Mark Oxley vowed to beat Andy Goram’s goal-scoring record this season after his comical strike scuppered Livingston 2-1 at Easter Road.
Top 3 events for each day based on their frequency

<table>
<thead>
<tr>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
</tbody>
</table>
| Tehran shut its most sensitive nuclear work bombing kill 28 people
Japanese government spokesman defend annual dolphin hunt | Japanese fishermen kill 30 dolphins
Oklahoma City Thunder beat Trail Blazers 105-97
Royal Caribbean ship depart Cape Liberty | Shiite Muslims protest action to stop violence
Chief Minister Balochistan visit Alamdar Road
Boston beat Washington 113-111 | Woodbury unveil new details of Diagon Alley
Vincent Asaro plead not guilty
Anaheim Duck beat Los Angeles Kings 2-1 | Google Inc suffer service outage
Pope Francis receive Hollande
London Stock Exchange list Cairn Energy PLC |
Generate 24.808 events from 205 news cluster with 71.493 articles.

Contributions:
- A novel approach for automatically mining events and event references from news clusters.
- Develop algorithm for extracting open-domain events using Open IE techniques.
- Present the result in a repository as resource to gain a comprehensive overview of world events and also serve as a resource for event-linking efforts in future IE research.
Paper accepted at AHA! Workshop on Information Discovery in Text.

Extracting a Repository of Events and Event References from News Clusters

Silvia Julinda
TU Berlin
Einsteinufer 17
Berlin, Germany

Christoph Boden
TU Berlin
Einsteinufer 17
Berlin, Germany

Alan Akbik
TU Berlin
Einsteinufer 17
Berlin, Germany

Future Work

- Adding Coreferences Resolution.
- Improving the event representation (e.q., n-ary relations, noun-phrase events).
- Merge events across news cluster (e.q., election).
- Exploring other tools or resources for the underlying components.
References


