# 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/

# What is an Event?

"An event instance is something that happens at a specific time and place“ (Topic Detection and Tracking, 2002)

**Event**


USAir 427 crash
on 08 September 1994
in Pittsburg

**Topic**


2

# Information Extraction (IE)

- 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.q., 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)

REES: 61 types of events

| 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.q., 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:

| 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 |

(Ritter et al, 2012)

- Huge amount of labeled data requires human effort and labor.

# Examples of events in our repository

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,

Social media giant Facebook has just monopolised the messaging market with its whopping $19 billion cash and stock

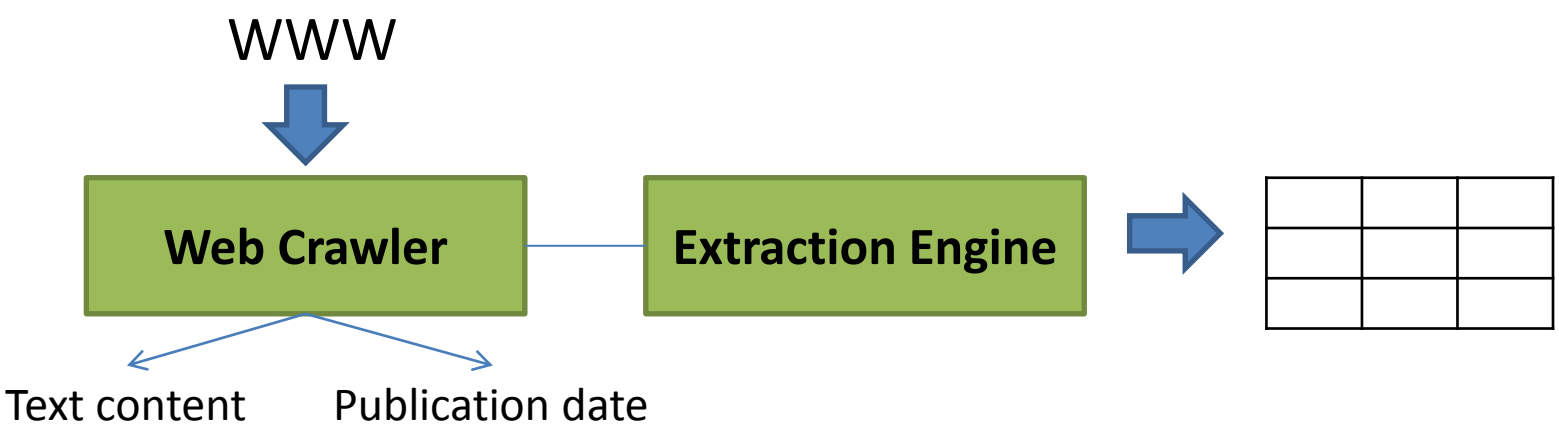
MENLO PARK, CALIF. – February 19, 2014 – Facebook today announced that it has reached a definitive agreement to acquire WhatsApp, a rapidly growing cross-platform mobile messaging

| Id | Timestamp | Place | Representative | Textual References |
|---|---|---|---|---|
| 1 | 2014-02-19 | Menlo Park | Facebook buy Whatsapp | social media giant Facebook monopolise messaging market |
| | | | | Facebook acquire WhatsApp |
| | | | | social media company purchase WhatsApp |

6

# News Cluster (Google News)

- Each cluster typically represent *one news item*

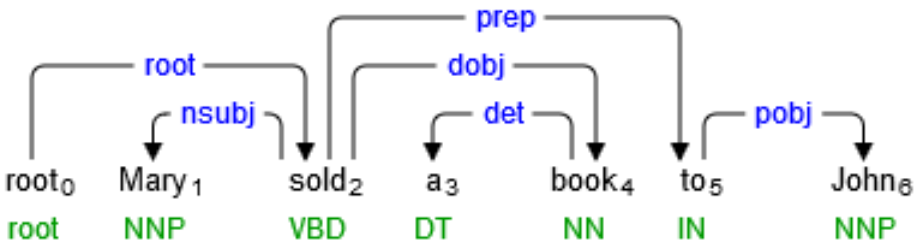



# Temporal Tagger

- 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.

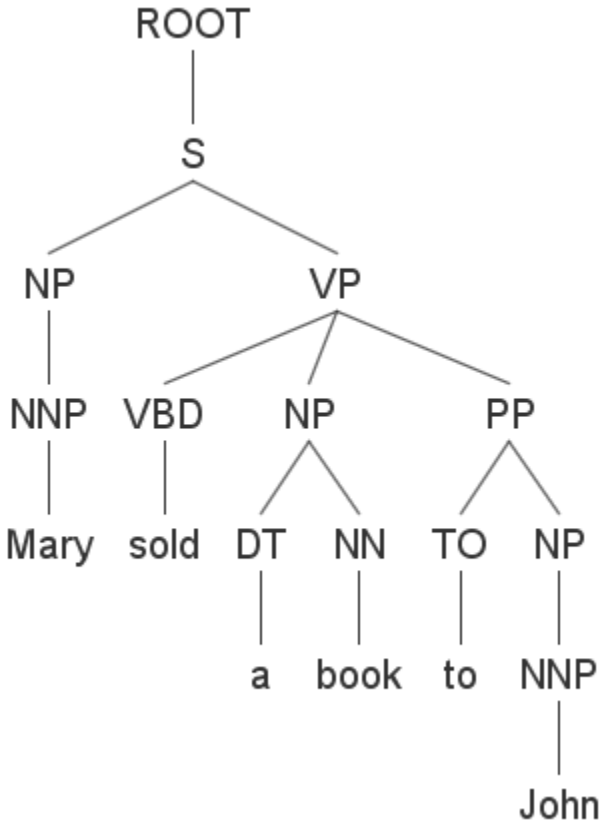
| Expression | Type | Value |
|---|---|---|
| 12 December 1974 | DATE | 1974-12-12 |
| Last Monday | DATE | 2014-02-17 |
| Christmast Eve | DATE | 2014-12-24 |
| Winter of 2014 | DATE | 2014-WI |
| Saturday Morning | TIME | 2014-02-22TMO |
| 7:18 p.m. | TIME | 2014-08-21T19:18 |
| 3 days | DURATION | P3D |
| from 2010 to 2014 | RANGE | 2010/2014 |
| every third Saturday | SET | XXXX-WXX-6 |

## Dependency Parsing




- Easier to understand
- faster
- binds grammaticality to words

## Constituent Parsing



# Semantic Role Labeler (SRL)

- "Who did What to Whom, When and Where?"

Mary sold a book to John in the library yesterday

↓

Mary sold a book to John in the library yesterday

**ARG0** **ARG1** **ARG2** **ARGM-LOC** **ARGM-TMP**

↓

**Predicate**: sell

**Roles**:

ARG0: Seller

ARG1: thing sold

ARG2: instrument/buyer

ARGM-LOC: location modifier

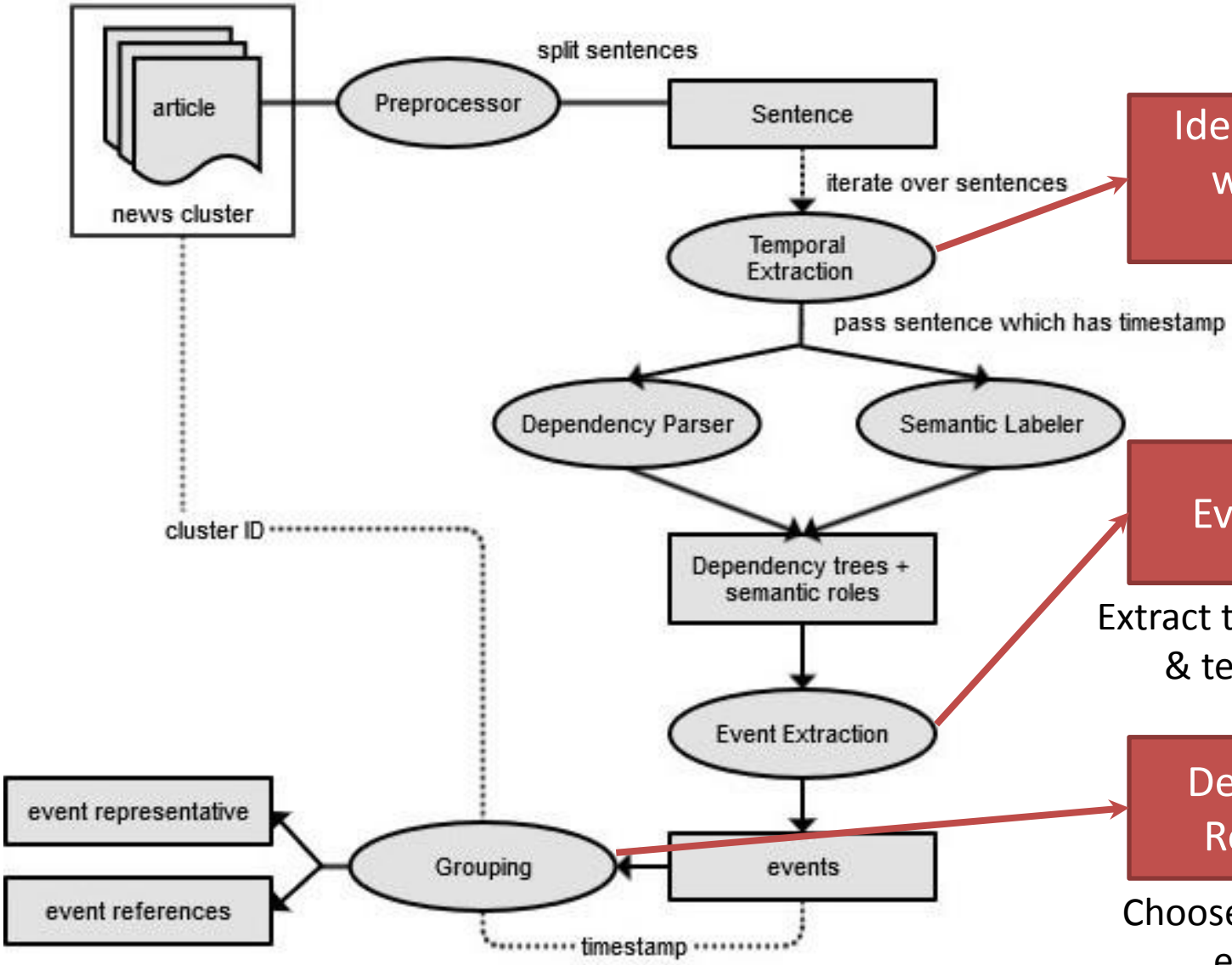
ARGM-TMP: temporal modifier

# Main Hypothesis

**Hypothesis: "same cluster + same timestamp = same event"**

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

# Overall Framework




Identify sentences with temporal expressions

Choose unit "day" as the granularity

Event Extraction

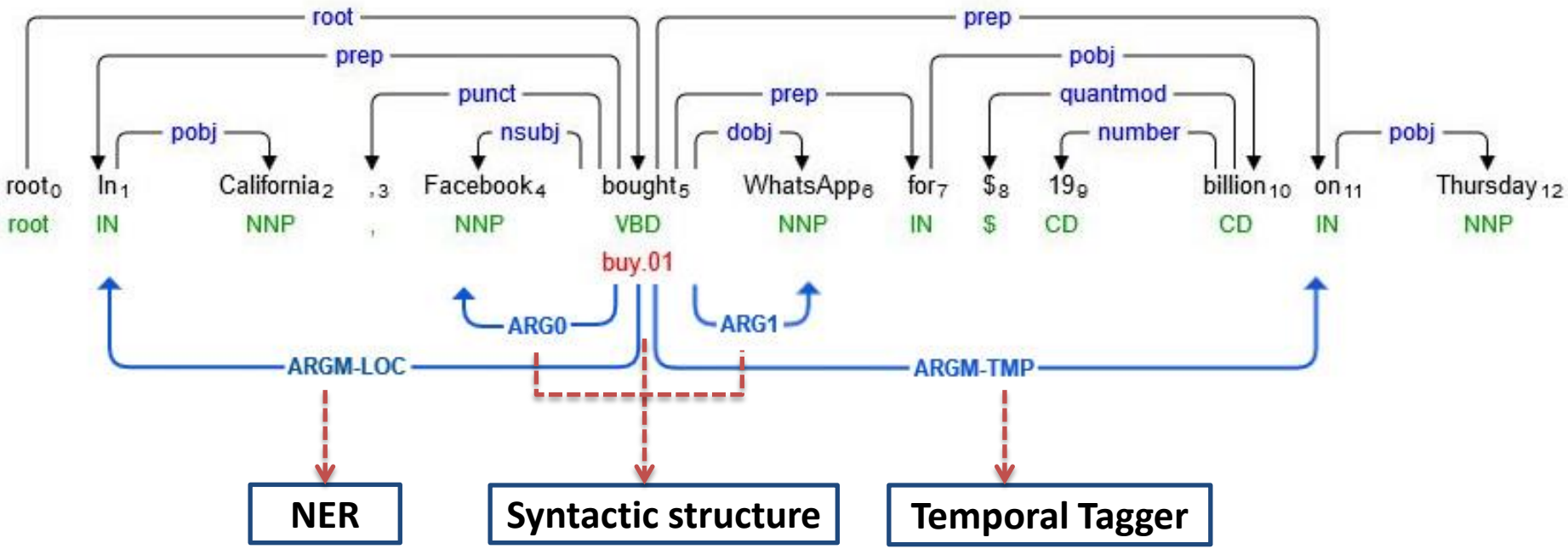
Extract timestamp, location & textual reference

Determine Event Representative

Choose the most frequent event 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

| Expression | Value |
|---|---|
| **A few months ago** | 2014-05-20 |
| **A few years ago** | 2013-06-20 |
| day and **night** | 2014-06-20TNI |
| man's **last minute** will | 2014-06-19TXX |
| **the same day** that Facebook celebrated | 2014-06-20 |

- 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"

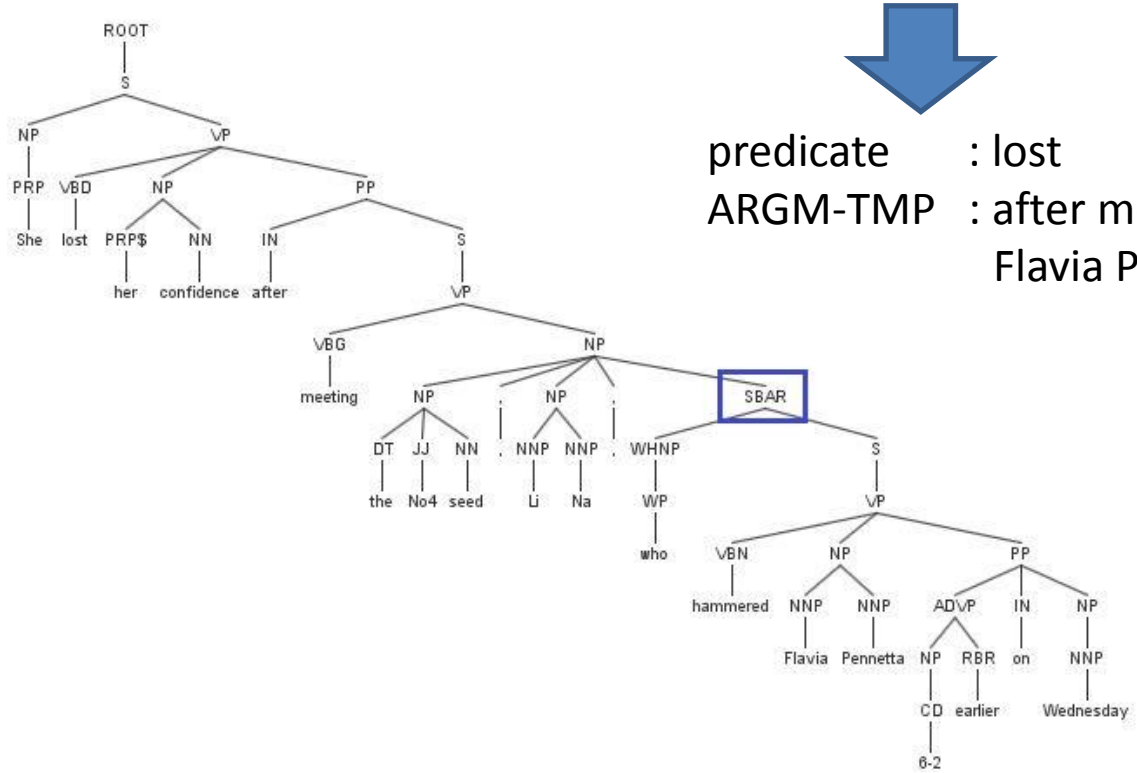
VBD
(verb, past tense)

→ last Wednesday → **2014-02-19**

# Temporal Extraction (Cont)

- 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

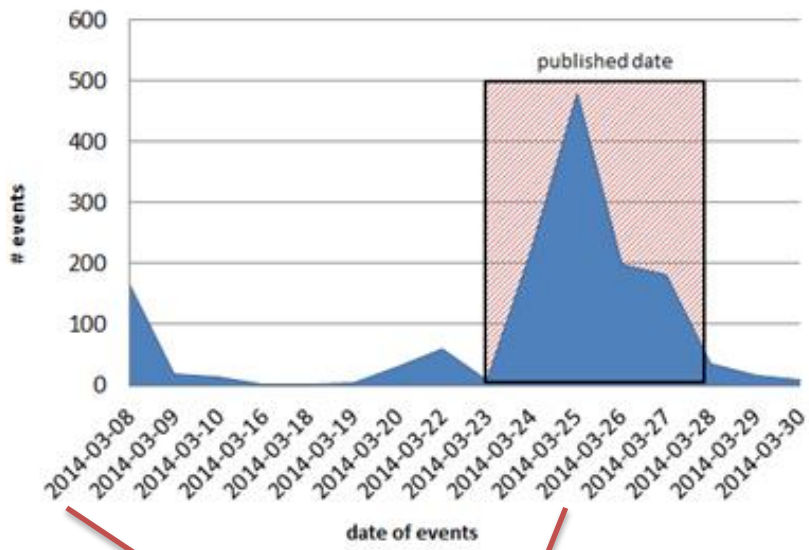



predicate : lost
ARGM-TMP : after meeting Li Na, who **hammered** Flavia Pennetta 6-2 on **Wednesday**

"lost" : Main clause
"Wednesday" : SBAR

**removed!**

# Preliminary Inspection of Result




**Hypothesis**
**"same cluster + same timestamp = same event"**
**fails in some cases!**

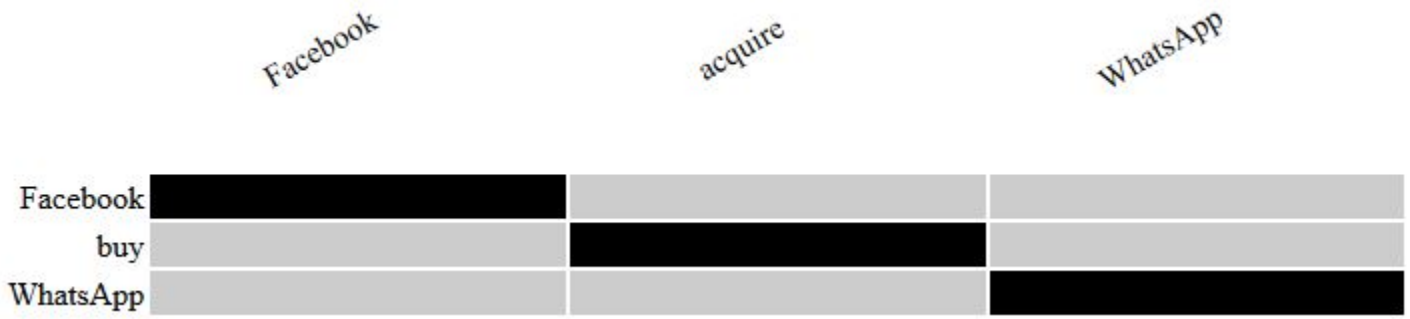
| Cluster ID | Timestamp | Textual Reference |
|---|---|---|
| 43988874 | 2014-03-08 | Malaysian Flight MH370 lose contact |
| | : | |
| | 2014-03-27 | Thailand spot 300 floating objects |
| | | Thaichote satellite acquire imagery of flotsam |
| | | Thai satellite detect 300 objects floating |
| | | Bangladesh join international search operations |

16

# 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

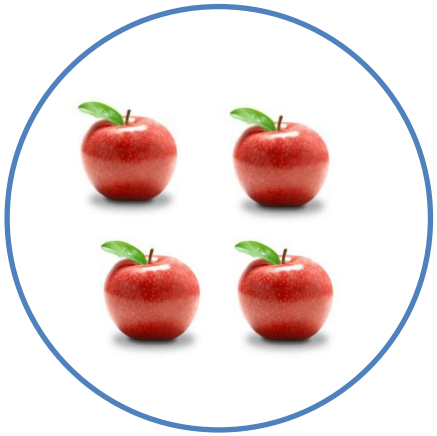
Facebook buy mobile messaging service
Facebook acquire Whatsapp
Facebook ink deal

venture capital firm hit 3 billion
jackpot company IPO hit jackpot

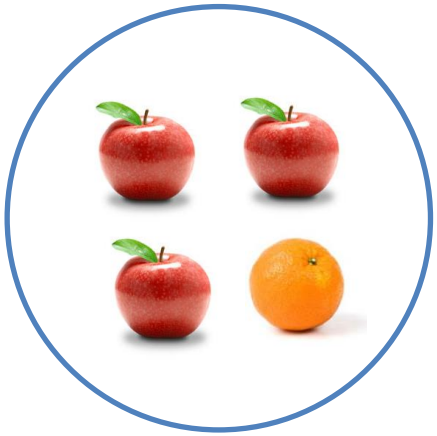
17

# 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

# Evaluation Method

- 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.

# Precision and Purity

| Method | Total Events References | Correct Events References | Precision of Event References | Event Purity |
|---|---|---|---|---|
| BASE | **2052** | 1711 | 0.834 | **0.798** |
| TIME | **339** | 274 | 0.808 | 0.788 |
| ALIGN | 2052 | 1793 | 0.874 | **0.917** |
| ALL | 339 | 277 | 0.817 | 0.893 |

- 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.

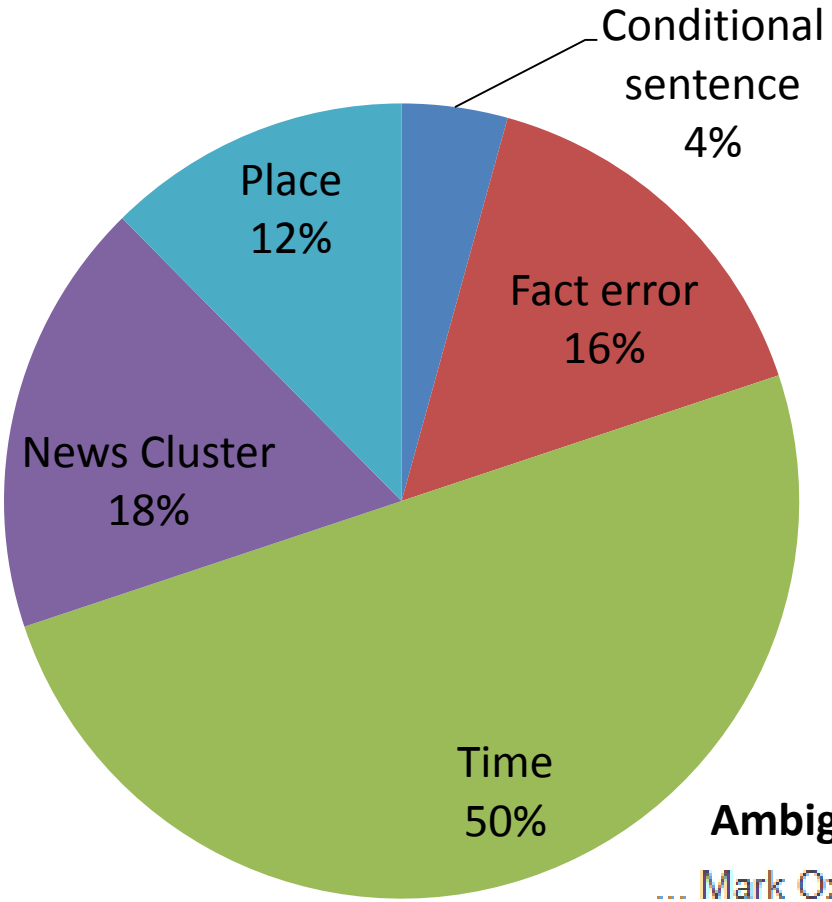
# Recall

- 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.

| | # Events | Percent |
|---|---|---|
| Noun phrase | 21 | 4.72 % |
| Coreference | 37 | 8.29 % |
| Common predicate | 77 | 17.27 % |
| Insufficient semantic roles | 82 | 18.38 % |
| **True positives** | **229** | **51.34 %** |

# Error Analysis

## Source of Errors




**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.

22

# Result Presentation

- Top 3 events for each day based on their frequency

« January 2014 »

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

# Summary

- 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.

# Publication

- 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


- Oral and poster presentation at Coling Conference 2014 in Dublin, 23 August 2014.

aha! of course is A (a) workshop.

Coling
Dublin | 2014
23-29 August

# 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


- Chinatsu Aone and Mila Ramos-Santacruz. 2000. Rees: a large-scale relation and event extraction system. In Proceedings of the sixth conference on Applied natural language processing, pages 76–83. Association for Computational Linguistics.

- Alan Ritter, Oren Etzioni, Sam Clark, et al. 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, 2012.

- James Allan. Topic detection and tracking: event-based information organization, volume 12. Springer, 2002.

- Angel X Chang and Christopher Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In LREC, pages 3735–3740.

- Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In ACL, pages 254–262

- Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. A lightweight and high performance monolingual word aligner. In Proceedings of ACL.