Conceptual and Practical Steps in
Event Coreference Analysis of Large-scale Data

Fatemeh Torabi Asr[1], Jonathan Sonntag[2], Yulia Grishina[2], Manfred Stede[2]

[1] MMCI Cluster of Excellence, Saarland University, Germany

[2] Applied Computational Linguistics, University of Potsdam, Germany

The formalism should allow for coreference between KILL and ASSASSINATE types:
Identical?
Sub-event?
Causally related event?
...

Mention 1

" President Kennedy was **killed** three days before he was to make these amendments public."

Mention 2

" Shortly after noon on November 22, 1963, President John F. Kennedy was **assassinated** as he rode in a motorcade through Dealey Plaza. "

Mentions 3

" Lushan, China (CNN) -- A strong earthquake that struck the southwestern Chinese province of Sichuan this weekend has **killed** 186 people, sent nearly 8,200 to hospitals and created a dire dearth of drinking water, Chinese state-run Xinhua reported Sunday. "

Unrelated mentions should be filtered out according to the attribute values:
People,
Time,
Location,
...

Identification of Events in the e-Identity Project

The e-Identity project: political scientists want to track interesting events in news and obtain collective information about them, i.e., the identity of every real-world event.

- **Event instances or event identities are:**
  - Finer-grained than topics
  - Coarser-grained than individual mentions in text
  - Units representative of human understanding of real-world eventualities (particular time, place and participants)
- **Data:**
  - English news 1990 – 2012
  - Avg. 1200 articles per month
  - Avg. 100 event mentions per article
- **Performance requirement:** recall is as important as precision
- **Efficiency requirement:** single-visit of mentions in the text is desired

Event **Mentions**
<what we find in text>

Real Event **Instances**
what we are interested in

Event **Types**
/the formalism/

A Two-step Event Clustering System

Semantic type identification

m = input event mention identified by ClearTK
$S = \{\}$, $R = \{\}$ //sets of synsets,
$mlem$ = head lemma of m,
$mpos$ = PoS of m

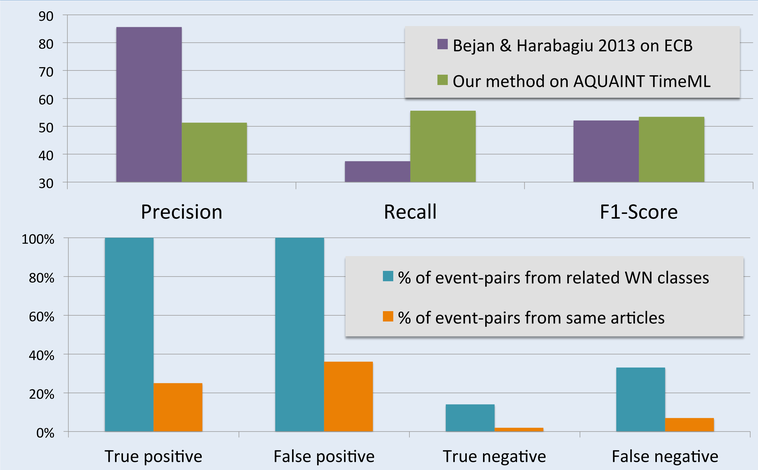
1. add all ***WordNet* synsets** for ($mlem$,$mpos$) to S
2. for every synset in S add all **hypernym** and **lexico-semantically** related synsets to R
3. nominate clusters that include some instances of the types in S *or* R

Similarity-based clustering

NE = **named entities** in the context of m,
TE = **temporal expressions** in the context of m

1. for every candidate cluster c calculate similarity with m and find the most similar cluster scored sim_{max} (sim is calculated by applying Jaccard index of the NE, TE, and S sets belonging to m and c)
2. if $sim_{max} > \theta$, assign m to the most similar cluster
3. else if $sim_{max} < \theta$, create a singleton cluster
4. update the type index w.r.t. S

Error Analysis of the Annotated Corpus



False positives:

- System considers a more flexible type identity compared with human
  - *"the immigration service **decided** the boy should go home."*
  - *"they made a reasonable decision Wednesday in **ruling** that"*
- Similar context (same article, paragraph, or sentence)
  - *"some people are **born** rich, some are **born** poor."*

False negatives:

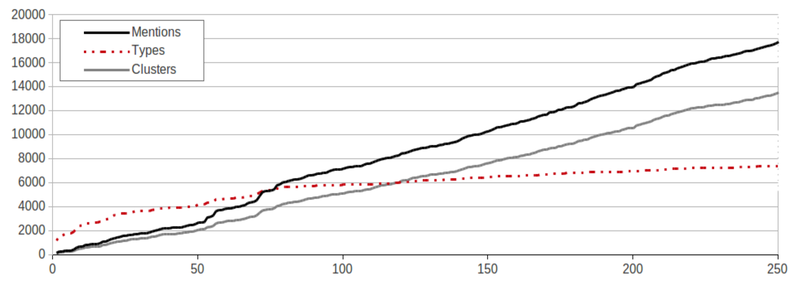
- Event types not connected in WordNet
- Sparse context (NEs and/or Timex not resolved)
  - *"the Clinton administration has pushed for the boy's **return**."*
  - *"his son said he didn't want to **go**."*

Annotating Event-pairs from TimeML AQUAINT

- Source corpus: AQUAINT TimeML
- Semi-random selection of 100 mentions, i.e., 4950 pairs
- Cross-document and cross-topic decisions
- Result: 36 corefrential and 4914 non-coreferential

http://www.coli.uni-saarland.de/~fatemeh/resources.htm

Growth of Types & Clusters (250 articles in a 2-week window)



Conclusion

- Named entities and timestamps are useful features (76.5% BLANC).
- Human decisions are more conservative than our system.
- False positives should be fixed by considering linguistic features.
- False negatives should be fixed by extending the semantic layer (e.g., more WordNet links or longer paths to be allowed).
- Temporal expressions are not very helpful in real data clustering.