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 mentions 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

A Two-step Event Clustering System

Semantic type identification

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

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

- $NE = \text{named entities in the context of } m,$
- $TE = \text{temporal expressions in the context of } m$

1. for every candidate cluster $c$ calculate similarity with $m$ and find the most similar cluster scored $\text{sim}_{m, c}$ ($\text{sim}$ is calculated by applying Jaccard index of the NE, TE, and 2 sets belonging to $m$ and $c$)
2. if $\text{sim}_{m, c} > 0$, assign $m$ to the most similar cluster
3. else if $\text{sim}_{m, c} < 0$, create a singleton cluster
4. update the type index w.r.t. $S$

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

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.”

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.