Background

The study of political behavior using event data automatically extracted from news text goes back to early 1990s [9]. Event data have been used in data journalism and near real-time event monitoring [1, 7]. Our interest lies in the public protest domain. This includes events like strikes, demonstrations, riots, terrorist attacks, on-line campaigns, symbolic protest actions (e.g. shoe throwing). We want to learn about protest forms, actors, locations and times, and intensity, the evolution of protest stories in time.

Figure 1: From a news report to structured event data representation

Related work: Systems

All widely used systems [8, 5] for political event data extraction
• are primarily oriented toward international relations and conflicts and much less so public protest,
• extract who did what to whom events; we need claims and grievances of protesters, numbers of participants – the whys and how manys,
• use pattern matching with large dictionaries of hand-crafted patterns, infamous for brittleness and low portability, e.g. from [8, 2]
  - ⟨content⟩[compete][oppose][protest][contest]⟨/⟩ against discrimination ⇒ Engage in diplomatic cooperation
  - ⟨content⟩[compete][oppose][protest][contest]⟨/⟩ nomination of candidate ⇒ Engage in political dissent (= public protest)
• use complex event/actor ontologies [2] with dozens of event and hundreds of actor types. But data reliability goes down with the complexity of the ontology [6, 5].

Related work: Corpora

There are no corpora of political event data for training statistical models and system evaluation. Also traditionally, the manual extraction of political event data has been performed at the level of document, not tokens. The Automated Content Extraction (ACE) ’05 corpus [11] covers some of this ground. The ACE ’05 corpus
• is a standard benchmark for event extraction,
• comes with rich token-level annotations, however
• does not include much protest (Protest events are primarily demonstrations, Attack events overlap with political violence),
• does not annotate the whys and how manys.

Our corpus

• We construct an English-language corpus of protest events with a budget to annotate about 300 documents (half the en ACE’05 corpus),
• annotate at the level of tokens, including event co-reference,
• work on a portion of the LDC English Gigaword corpus [3] and will subsequently release all annotations.

Corpus features

• Our annotators are political scientists familiar with manual event extraction (=coding in the social sciences), which is document-level annotation.
• The annotation process involves traditional coding and token-level annotation as is practised in NLP.
• We ask the annotators to think of token-level annotation as a means of explaining their coding decisions with the help of annotation rules.
• Our annotation guidelines borrow from the ACE guidelines. We explain the same concepts in less technical language, e.g. using particle modify to a noun instead of present-participle in the nominal pre-modifier position. We have introduced many simplifications, e.g. the avoidance of syntactic-phrase annotations or annotations embedded within other annotations.
• https://pub.c1.uzh.ch/projects/nccr/polcon/guidelines
• We experiment with the linking approach to event co-reference as opposed to defining it explicitly, which is notoriously hard [4].

Figure 2: Annotation example. We use browser-based annotation interface </br>

Intermediate results

Inter-annotator agreement results for unmasked tests on single sentences (overly optimistic)

<table>
<thead>
<tr>
<th>Component</th>
<th>F1-score μ</th>
<th>F1-score σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>event anchor (docs)</td>
<td>0.827</td>
<td>0.028</td>
</tr>
<tr>
<td>actors size</td>
<td>0.864</td>
<td>0.145</td>
</tr>
<tr>
<td>location date</td>
<td>0.897</td>
<td>0.028</td>
</tr>
<tr>
<td>size</td>
<td>0.724*</td>
<td>0.046</td>
</tr>
<tr>
<td>location</td>
<td>0.872</td>
<td>0.026</td>
</tr>
<tr>
<td>date</td>
<td>0.758*</td>
<td>0.056</td>
</tr>
</tbody>
</table>

Future work

• Some documents are on related topics and different dates. We shall post-hoc add some annotation of cross-document co-reference.

• A beautiful structured prediction problem awaiting neat handling.

References