Political Conflict in Europe
in the Shadow of the Great Recession

An Annotated Corpus for Protest Event Mining

Peter Makarov

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, topics and intensity, the evolution of protest stories in time.

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

Trade unions *are satisfied with the course of* today's blockade *of the Czech-Slovak Drietoma-Stary Hrozenkov border crossing aimed to* highlight bad social and economic conditions *in* Slovakia [...]

Czech News Agency, 2 March 2001

ACTOR	trade unions
DATE	02.03.2001
ACTION FORM	blockade
ISSUE	welfare
LOCATION	Slovakia
...	...

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]
  - (CONTEND|COMPETE|OPPOSE|PROTEST|CONTEST|...) AGAINST DISCRIMINATION OF ⇒ **Engage in diplomatic cooperation**
  - (CONTEND|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 *why*s and *how many*s.

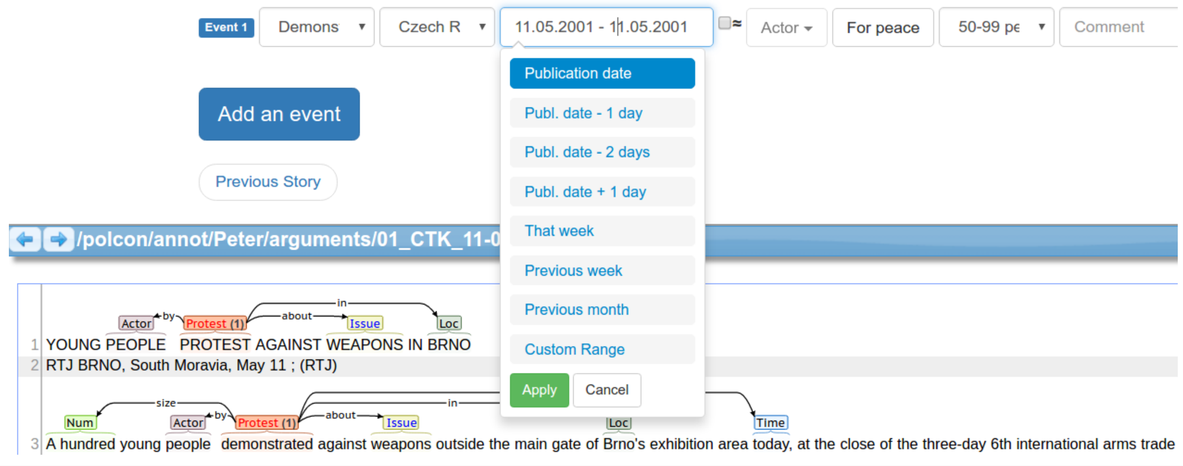
Our corpus

- 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 *participle modifying 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.cl.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 `brat` [10] to annotate at the token level. We embed `brat` in a simple web form-like interface that supports document-level coding.



Intermediate results

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

Table 1: Average pairwise F1-scores for exact match computed for 4 annotators. *Size and time are predominantly multi-word annotations. The average number of e.g. event anchors is 59.0, date 23.5.

Component	F1-score	
	μ	σ
event anchor	0.827	0.028
anchor (docs)	0.864	0.145
actors	0.897	0.028
size	0.724*	0.046
location	0.872	0.026
date	0.758*	0.056

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

[1] John Beieler, Patrick T Brandt, Andrew Halterman, Philip A Schrodt, and Erin M Simpson. Generating political event data in near real time. *Computational Social Science*, page 98, 2016.

[2] Deborah J Gerner, Philip A Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr. Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. *International Studies Association, New Orleans*, 2002.

[3] David Graff, Linguistic Data Consortium, et al. *English Gigaword Third Edition LDC2007T07*. 2007.

[4] Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. Events are not simple: Identity, non-identity, and quasi-identity. In *NAACL HLT*, volume 2013, page 21, 2013.

[5] Gary King and Will Lowe. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57(03):617–642, 2003.

[6] Slava Mikhaylov, Michael Laver, and Kenneth R Benoit. Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, 20(1):78–91, 2012.

[7] Sean P O'Brien. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1):87–104, 2010.

[8] Philip A Schrodt. Automated production of high-volume, real-time political event data. In *APSA 2010 Annual Meeting Paper*, 2010.

[9] Philip A Schrodt and Blake Hall. Twenty years of the Kansas event data system project. *The political methodologist*, 14(1):2–8, 2006.

[10] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. BRAT: a web-based tool for NLP-assisted text annotation. pages 102–107. Association for Computational Linguistics, 2012.

[11] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. ACE 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 2006.

Contact

University of Zurich[UZH]

Institute of Computational Linguistics

http://www.cl.uzh.ch/

Peter Makarov
Binzmühlestrasse 14
8051 Zurich
makarov@cl.uzh.ch