PLOVER Event Data Ontology

Philip A. Schrodt

Parus Analytics LLC
Charlottesville, VA
http://philipschrodt.org
http://eventdata.parusanalytics.com

Peace Research Institute, Oslo
15-16 August 2023

Event Data Ontologies used in US Government Research

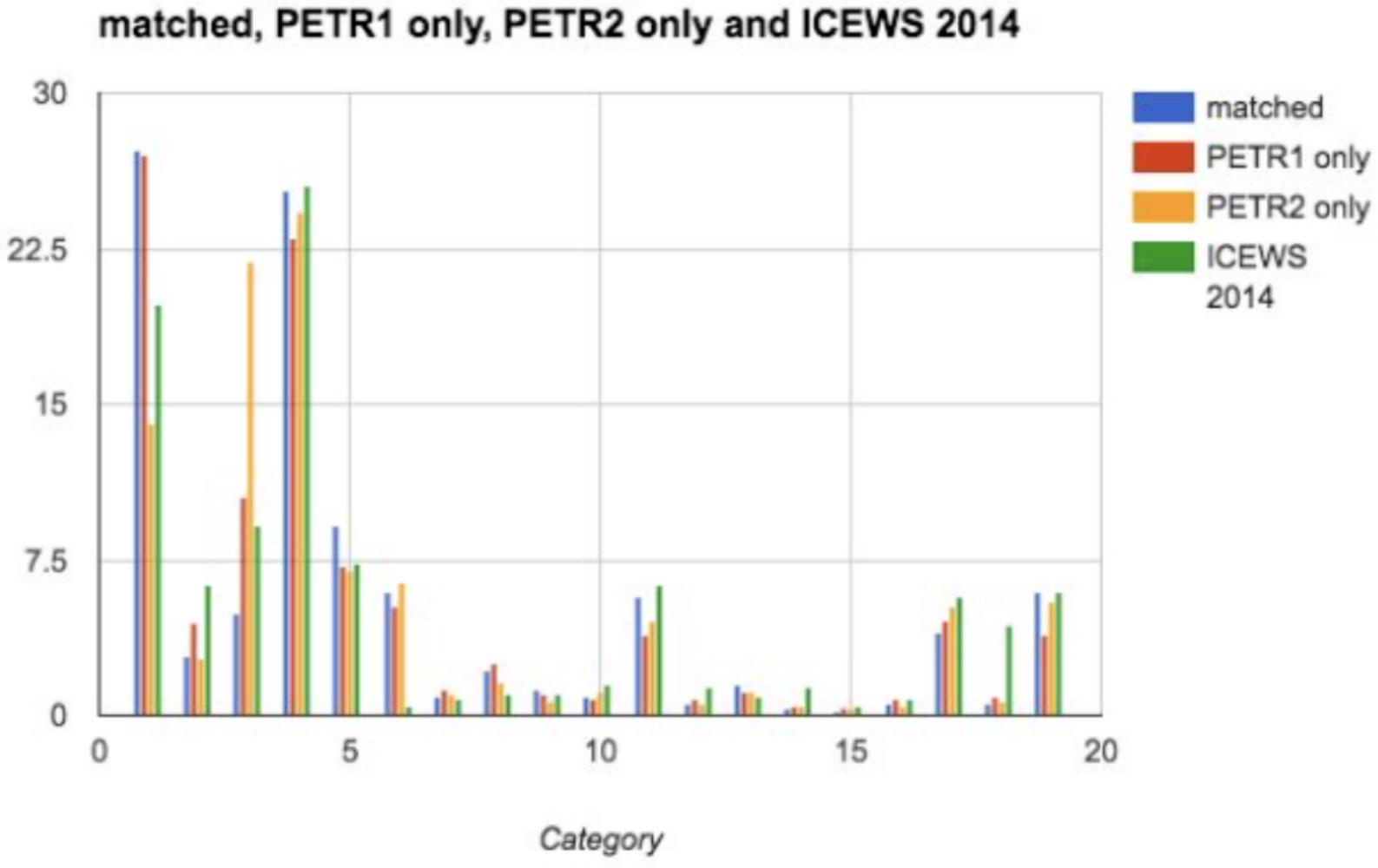
- WEIS (McClelland 1965) 1970-2005. Human coded from *New York Times* by, variously, students at UC Santa Barbara, CACI for Reagan National Security Council, and cadets at U.S. Naval Academy. Machine coded by KEDS program for Middle East from Reuters.
- CAMEO (Gerner et al, 2002) 2009 (ICEWS) to present. Released with regular updates on Dataverse 2015-2023; global coding from Factiva by TABARI, then SERIF.
- PLOVER: initially developed in a government, academic, and NGO collaboration ca. 2018, then subsequently developed by Political Instability Task Force for POLECAT dataset

ICEWS event data

`https://asecondmouse.wordpress.com/2015/03/30/seven-observations-on-the-newly-released-icews-data/`

- Produced by Lockheed, BBN, and PITF
- Data released on open-access Dataverse and covers 1996-March-2023
- Government version includes source texts
- Raytheon/BBN [proprietary] Serif/ACCENT coder
- Global coverage but events are still disproportionately from Asia
- BBN has extensively refined the CAMEO specification and coding manual is on Dataverse
- Includes a subset of the actor dictionaries but not verb dictionaries
- Geolocated, though with quite a few errors

Distribution of CAMEO events



PLOVER

Political Language Ontology for Verifiable Event Records

Event, Actor and Data Interchange Specification

Open Event Data Alliance

http://openeventdata.org/

http://ploverdata.org/

DRAFT Version: 0.6b2

March 2017

PLOVER objectives

GitHub: `https://github.com/openeventdata/PLOVER`

- Only the 2-digit event "cue categories" have been retained from CAMEO. However, PLOVER takes the general purpose "events" of CAMEO (as well as the earlier WEIS, IDEA and COPDAB ontologies) and splits these into "$event - mode - context$" which generally corresponds to "$what - how - why$."

- Standard optional fields have been defined for some categories, and the "target" is optional in some categories.

- A set of standardized names for line-delimited JSON (`http://www.json.org/`) records are specified for both the core event data fields and for extended information such as geolocation and extracted texts;

PLOVER events

- AGREE
- CONSULT
- SUPPORT
- CONCEDE
- COOPERATE
- AID
- RETREAT
- REQUEST
- ACCUSE
- REJECT
- THREATEN
- PROTEST
- SANCTION
- MOBILIZE
- COERCE
- ASSAULT

Examples of PLOVER modes

RETREAT

- withdraw (from territory)
- release (captives)
- return (property)
- disarm
- ceasefire
- access (allow third party access)
- resign

COERCE

- seize
- restrict
- ban
- censor
- martial law
- arrest
- deport
- withhold

PLOVER contexts (partial list)

- military
- diplomatic
- executive
- legal
- intelligence
- legislative
- political institutions
- pro-democracy
- pro-autocracy
- economic
- reparations

"New generation event coder"

- Use the massive neural network "transformer models" that have been developed by Google, Amazon, Facebook, and Microsoft and are largely open-sourced
- Use training examples (250 to 500 per category) rather than dictionaries
- Observations are 512-token texts rather than single sentences (this is consistent with Google BERT family of models).
- Approaches for components of PLOVER
  - Events: transformer classification models
  - Mode and context: support vector machine models
  - Actors and locations: transformer "question answering" models linking to open databases such as Wikipedia and Geonames
  - Miscellaneous fine tuning (e.g. compound actors): dependency parsing

POLECAT event data set

- Replaces the PITF Dataverse ICEWS as of April-2023.
- As with ICEWS, texts are available for US government users but due to intellectual property issues, not on Dataverse
- Currently coded back to 2018, though further backcoding is constrained by
  - Cost of Factiva source texts
  - Computing time: currently about 16 hours per week of data

Further developments for PLOVER/POLECAT

- We started this in summer-2021, about 18 months before the proliferation of very large language models such as GPT-3, GPT-4 and LLaMA: BERT models are a couple orders of magnitude smaller. However, it is not clear how to most effectively use the larger models
- Need to have a canonical definition of PLOVER using training cases that do not have intellectual property constraints, either through "fair use" or synthetical cases: US and European legal situations may be quite different here, with EU much more permissive for research development
- High quality training cases appear to be very important in these models, so we need better fine tuning.
- Find the right balance between LLM and parser/dictionary methods: POLECAT produces too many false positivers.

- We should not have “one data set or model to rule them all”
- Follow the approach of hurricane and snowstorm forecasters who triangulate results of multiple independently developed data sets and models which have different assumptions and strengths

Thank you

Email:
schrodt735@gmail.com

Slides:
`http://eventdata.parusanalytics.com/presentations.html`

Links to data and software: `http://philipschrodt.org`

Blog with lots of extended commentary on event data:
`http://asecondmouse.org`