Modernizing Political Event Data
Patrick T. Brandt 1, Latifur Khan 2, Javier Osorio 3
1UT Dallas (brandt@utdallas.edu), 2UT Dallas (lkhan@utdallas.edu), 3Arizona (josorio1@email.arizona.edu)

Existing Project Overview

Introduction
- We develop technology and methodology to detect, understand, and predict intra- and inter-state conflict around the globe.
- Funding from NSF HBCU Trust Grant No. SSE-SMA-1559092 and SSEDEI.A, Challenge to TACC / Indiana University through allocation SSE17012
- Generates daily computer-coded data about conflict from news reports in English, Spanish, and Arabic.
- Have access to historical news reports back to 1945.

Motivation
- Researchers and practitioners need high quality, timely, and event data at global scale.
  - No data set currently provides accurate structured data on political and social events around the globe, with historical coverage, geographic-location, drawn from multiple languages, and freely available in near-real time.
- There is need to modernize event coding by moving beyond the specialized skills required for generating, visualizing, and analyzing event data.
- Working with computer scientists, we built a platform to code across multiple languages, topics, and issue areas.
- Allows researchers to capture richer details from textual data sources and derive sharper analysis from the data.

Coding Events

Event Data
- Machine-coded event data are a description of:
  - **Who** = Source of the event.
  - **Src.Act**: Source at the action.
  - **Act.Tar**: Target of the action.
  - **Where** = Geolocation of the event.

Human limitations
- Humans can easily process complex information.
  - Parse information from sentences.
  - Describe implicit information.
  - Process source validity and rarity of information.
  - Disambiguation of actors and geography is easy.
- But doing it consistently is hard for humans.
- This is a real big data problem, which has the three V’s:
  - Volume: Global daily processing requirement 10B-10K reports.
  - Velocity: We need near-real-time data and analysis.
  - Accuracy of human coding:
    - Initial
    - Experienced

How we code event data

1. Scrape news reports from the web (100+ sources) or from a database of them (200+ sources).
2. Parse the sentences (usually first 4-5 of them) using NLP that applies Part of Speech tagging.
3. Using the Part of Speech tags and co-referencing, record the Source, Verb and Target (and the actual names and verbs), the event date, and location.
4. Count the verb into one of the 200 categories in the Conflict and Mediation Event Ontology (CAMBO). We will also extend these categories.

Cognitive development

Technology development is available at these sites:

- [http://eventdata.utdallas.edu](http://eventdata.utdallas.edu)
- [https://github.com/open eventdata](https://github.com/open eventdata)
- [http://openeventdata.org](http://openeventdata.org)

Online access to event datasets

HTTP protocol-based REST-API for accessing the data:
- Supported on NSF’s XSEDE / TACC JetStream Cloud.
- Personalized API key based authentication mechanism.
- Java APIs: Object. Notation (JSON) output format or read with our C package.
- Real-time is geo-located using Moreovai.
- Datasets available:
  - Phoenix real time data, October 2017: present.

Spanish and Arabic developments

CAMBIO Translating Application (volt)
- github.com/open eventdata / synthet_validator
- Requires WordNet technology.
- Supports synonym and translation.
- Takes advantage of years of CAMBIO-dictionary development.
- Facilitates systematic translation and enables escalation.

Interactive data uses

TwoRavens: Explore and extract event data
- eventdata/zravons.org
  - Subset events subject to constraints.
  - Aggregate by date and source/actor relationships.
  - Visually explores statistical properties.

Data Access Library in R

- Extracting event data from the UT-Dallas via R.
- Can subset data by country and time ranges.

Policy implications

Bridge the gap between practical policy decision making and analysis, and the technical and sophisticated generation of big data on conflict at global scale.

Monitoring: Near-real time global data on conflict.
Expansion: Use of local sources in multiple languages.
Forecasting: Precise estimation of possible scenarios and early warning systems.
Detailed data: High-quality information.
Learning: Identify causes and consequences of policy interventions.
Public good: Lower technological and skill costs to access data.

Cross-disciplinary collaboration

### Seeking Collaborators

Willing to partner with those interested in (i) forecasting, (ii) computational linguistics, (iii) event extraction and classification methods, (iv) big data mining, and (v) UI developers. Email us if you are interested in collaborations.

Event data examples

PETRARCH + Universal Dependencies

Political event coding using universal dependencies (UD)

- Develop event code for three languages: English, Spanish, Arabic.
- Relies on dependency parses with universal tags.
- Uses modular structure for easy extensions.

Geolocation tool: PROFILE

- Challenges:
  - Different candidate locations exist within each report.
  - Unavailability of suitable labeled instances.
  - Disambiguation of writing styles, language contrast, etc., between news agencies’ may cause bias in training and test data.
- Contributions:
  - Distinguishing primary focus location.
  - Utilizing bias correction.

Software development

Technology developments are available at these sites:

- [http://eventdata.utdallas.edu](http://eventdata.utdallas.edu)
- [https://github.com/open eventdata](https://github.com/open eventdata)
- [http://openeventdata.org](http://openeventdata.org)

UTD Real-time Phoenix: Real-time events at UT-Dallas.
ODIA: Provides infrastructure for research.
RePAIR: Real-Time Political Actor Recognition.
PRFILER: Plloric Focus Location Extraction tool.
UD-PROFILE: Language-agnostic event data coding.
UTDEventData R package: Accessing event data in R.