Modernizing Political Event Data

Patrick T. Brandt [1] Latifur Khan [2] Javier Osorio [3]

[1]UT Dallas (pbrandt@utdallas.edu), [2]UT Dallas (lkhan@utdallas.edu), [3]Arizona (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 RIDIR Grant No. SBE-SMA-1539302 and XSEDE Jetstream at TACC / Indiana University through allocation SES170012.
- 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

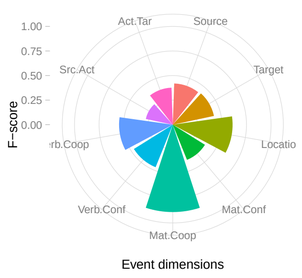
Event Data

- Machine-coded event data are a description of:
  - **Who** = Source of the event.
  - **Did/Said** = material/verbal interactions of conflict/cooperation.
  - **To Whom** = Target or the object of the action.
  - **Where** = The geo-location of the event.

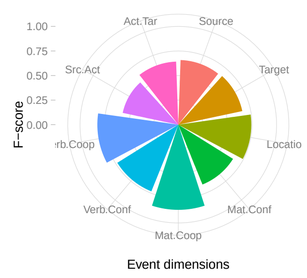
Human limitations

- Humans can easily process complex information.
  - Parse information from sentences.
  - Discard duplicate 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:
  - **Variance**: Trained humans are 70-75% reliable.
  - **Volume**: Global daily processing requirement 10K-13K 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 (300+ 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. Convert the verb into one of the 200 categories in the Conflict and Mediation Event Ontology (CAMEO). We will also extend these categories.

Software development

Technology developments are available at these sites:

- `http://eventdata.utdallas.edu`
- `https://github.com/openeventdata`
- `http://openeventdata.org`

UTD Real-time Phoenix: Real-time events at UT-Dallas.

OEDA: Provides infrastructure for research.

RePAIR: Real-Time Political Actor Recommendation.

PRoFiLE: PRimary Focus Location Extraction tool.

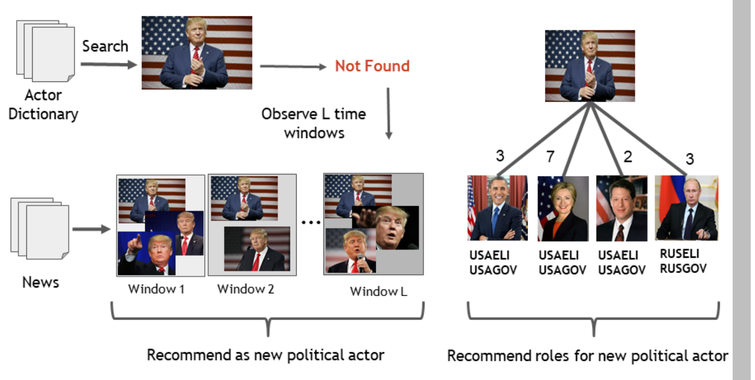
UD-PETRARCH: Language-agnostic event data coder.

Two-Ravens: Statistical analysis of event data.

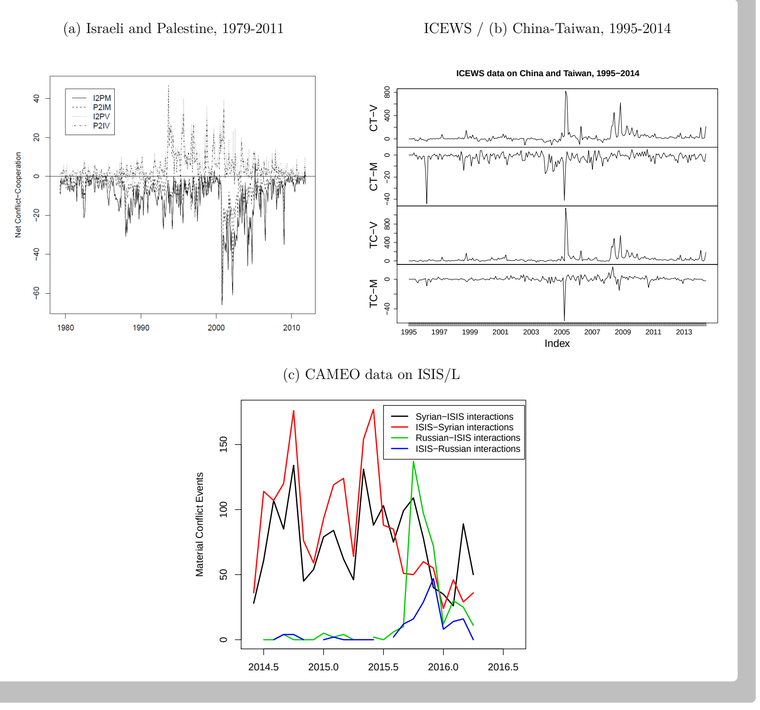
UTDEventData R package: Accessing event data in R.

Actor detection

- RePAIR: Recommend Political Actors In Real-time
- Challenge: Actors can have multiple alias or change role over time.
- Contribution: real-time recommendation of:
  - Possible new actors.
  - Related roles.
- Users can validate using an app: `http://1-dot-utd-actors-roles.appspot.com/`



Event data examples



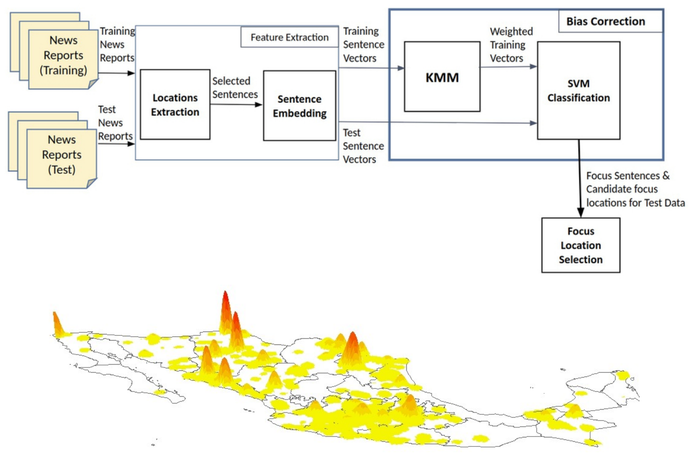
PETRARCH + Universal Dependencies

Political event coding using universal dependencies (UD):

- Develops event coder for three languages: English, Spanish, Arabic.
- Relies on dependency parse tree 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.
  - Dissimilarity of writing styles, linguistic content, etc. between news agencies' may cause bias in training and test data.
- Contributions:
  - Distinguishing primary focus location.
  - Utilizing bias correction.



Phoenix Data Examples

1200194_v0.2.0	20150708	2015	7	8	EGYGOV	EGY	GOV		USALEG	USA
1200195_v0.2.0	20150708	2015	7	8	FRACVL	FRA	CVL		FRA	FRA
1200196_v0.2.0	20150707	2015	7	7	NGOHLHIRC	NGO		HLH;IRC	NGOHLHIRC	NGO
1200197_v0.2.0	20150708	2015	7	8	IRQELIGOV	IRQ		ELI;GOV	IND	IND
1200198_v0.2.0	20150708	2015	7	8	PAKGOVMED	PAK	GOV	MED	PAK	PAK
1200199_v0.2.0	20150707	2015	7	7	SAU	SAU			SYR	SYR

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 (iv) *UI developers*. Email us if you are interested in collaborations.

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.
- JavaScript Object Notation (JSON) output format or read with our R package.
- Real-time is geo-located using Mordecai.
- Datasets available:
  1 Phoenix real time data, October 2017–present.
  2 ICEWS 1995–September 2016.
  3 Cline Center Phoenix New York Times, 1945–2005.
  4 Cline Center Phoenix BBC SWB, 1979–2015.
  5 Cline Center Phoenix FBIS, 1995–2004.

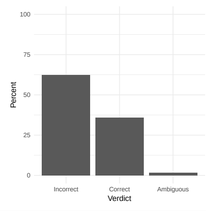
Spanish and Arabic developments

CAMEO Verb Translation Application (VTA)

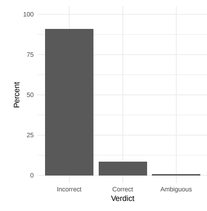
- `github.com/openeventdata/synset_validator`
- Relies on WordNet technology.
- Suggests synsets and translation.
- Takes advantage of years of CAMEO dictionary development.
- Facilitates systematic translation and enables escalation.

Human verdicts on

English synsets



Spanish translations



Interactive data uses

TwoRavens: Explore and extract event data `eventdata.2ravens.org`

- Subset events subject to constraints.
- Aggregate by date and source/actor relationships.
- Visually explore statistical properties.

Data Access Library in R

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

Data Table	Timeline	Information
Phoenix RT	Oct. 2017 – Today	OEDA
ICEWS	1995 – Sep.2016	ICEWS Dataverse
Cline Phoenix NYT	1945 – 2005	Cline Center
Cline Phoenix FBIS	1945 – 2005	Cline Center
Cline Phoenix SWB	1979 – 2015	Cline Center

- Allows event data aggregation for merging other data.
- `github.com/KateHyoung/UTDEventData`

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 massive data.

Cross-disciplinary collaboration

Political Science	Computer Science
Benjamin Bagozzi, Delaware	Latifur Khan, UT-Dallas
Patrick T. Brandt, UT-Dallas	Vincent Ng, UT-Dallas
John Freeman, Minnesota	**Students**
Andy Halterman, MIT	Graduate students 12+
Jennifer Holmes, UT-Dallas	Undergraduate students 20+
Jill Irvine, Oklahoma	
Vito D'Orazio, UT-Dallas	
Javier Osorio, Arizona	
Philip Schrodt, Parus Analytics	