Comparison Metrics for Large Scale Political Event Data Sets

Philip A. Schrodt

Parus Analytics
Charlottesville, Virginia, USA
schrodt735@gmail.com

Paper presented at the European Political Science Association meetings, Vienna, 25 June 2015

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

Outline

- EL:DIABLO/Phoenix open-source system
- Why multiple sources are not necessarily a good thing
- A comparison metric for event data sets
- Example 1: BBC single-source data set vs ICEWS multi-source
- Example 2: shallow (TABARI)vs full (PETRARCH) parsing for the KEDS Levant data
- Next steps

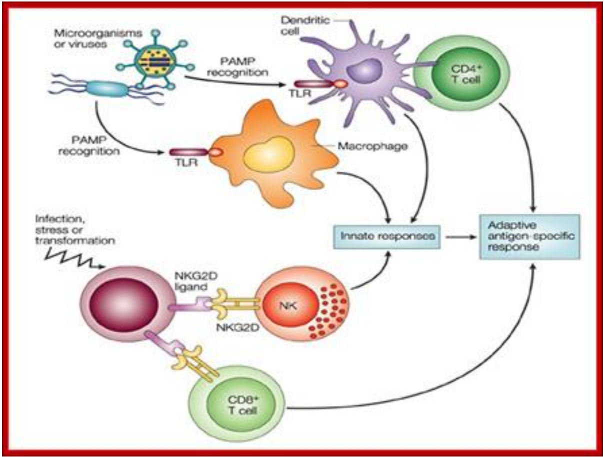
EL:DIABLO
Event Location: Dataset in a Box, Linux Option

- Open source: `https://openeventdata.github.io`
- Full modular open-source pipeline to produce daily event data from web sources. `http://phoenixdata.org`
- Scraper from white-list of RSS feeds and web pages
- Event coding from any of several coders: TABARI, PETRARCH, others
- Geolocation: new "Mordecai" open source geolocater from Caerus Associates
- "One-A-Day" deduplication keeping URLs of all duplicates
- Designed for implementation in inexpensive Linux cloud systems
- Supported by Open Event Data Alliance `http://openeventdata.org`

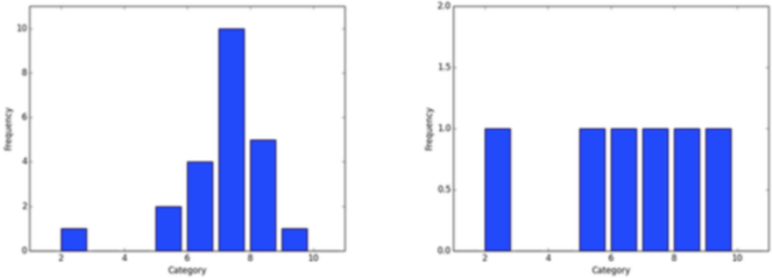
Humans use multiple sources to create narratives

- Redundant information is automatically discarded
- Sources are assessed for reliability and validity
- Obscure sources can be used to "connect the dots"
- Episodic processing in humans provides a pleasant dopamine hit when you put together a "median narrative": this is why people read novels and watch movies.

Machines latch on to anything that looks like an event



This must be filtered



(a) What was generated (b) What remains

Figure 2: Effect of One-A-Day filtering

Implications of one-a-day filtering

- Expected number of correct codes from a single incident increases exponentially but is asymptotic to 1
- Expected number of incorrect codings increases linearly and is bounded only by the number of distinct codes

Tension in two approaches to using machines [Isaacson]

- “Artificial intelligence” [Turing, McCarthy]: figure out how to get machines to think like humans
- “Computers are tools” [Hopper, Jobs]: Design systems to optimally *complement* human capabilities

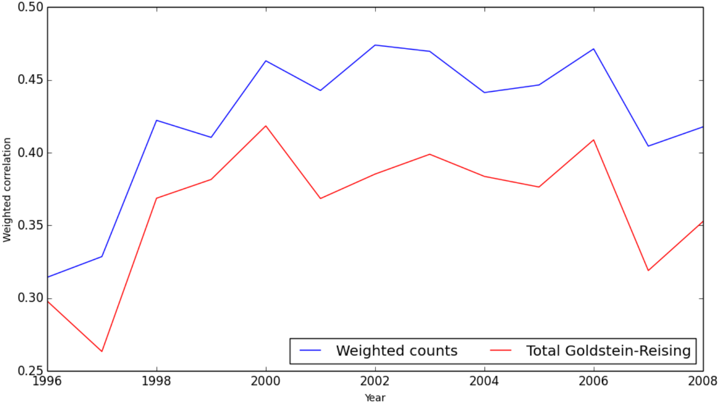
Weighted correlation between two data sets

$$wtcorr = \sum_{i=1}^{A-1} \sum_{j=i}^{A} \frac{n_{i,j}}{N} r_{i,j} \quad (1)$$

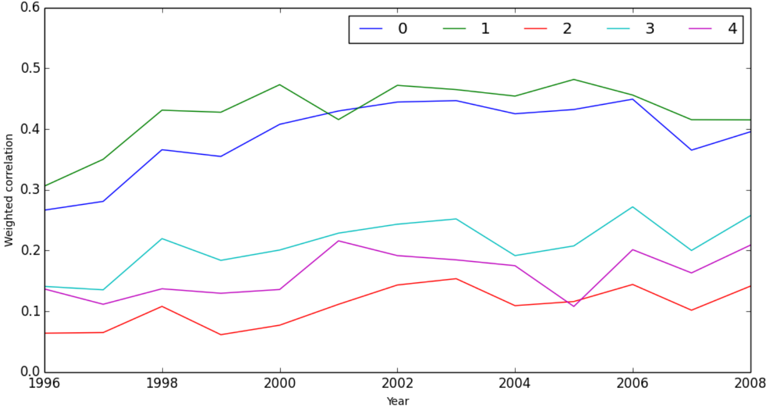
where

- A = number of actors;
- $n_{i,j}$ = number of events involving dyad i,j
- N = total number of events in the two data sets which involve the undirected dyads in A x A
- $r_{i,j}$ = correlation on various measures: counts and Goldstein-Reising scores

BBC vs. ICEWS: Correlations over time: total counts and Goldstein-Reising totals



Correlations over time: pentacode counts



Dyads with highest correlations

Table 1: Fifty dyads with highest average correlation on total counts

RUS-CHN 0.76	CHN-ZAF 0.72	CHN-EGY 0.67	CHN-PAK 0.66	CHN-DEU 0.66
CHN-SYR 0.66	CHN-HRV 0.65	CHN-JPN 0.64	RUS-JPN 0.63	UKR-HRV 0.63
RUS-IRN 0.61	CHN-FRA 0.60	CHN-ROU 0.60	CHN-IND 0.59	CZE-HRV 0.59
CHN-GBR 0.59	CHN-MEX 0.59	RUS-PSE 0.59	CHN-LKA 0.59	CHN-VNM 0.59
HRV-ROU 0.58	CHN-PSE 0.58	RUS-IND 0.58	RUS-DEU 0.57	TUR-POL 0.57
CHN-TUR 0.57	IRN-PAK 0.56	CHN-IRN 0.56	IRN-TUR 0.56	RUS-VNM 0.56
IRN-SYR 0.56	CHN-BRA 0.55	CHN-ESP 0.55	RUS-GBR 0.55	TUR-UKR 0.55
DEU-ROU 0.54	USA-CHN 0.54	RUS-CAN 0.54	CHN-AUS 0.54	RUS-EGY 0.54
CHN-ARG 0.54	RUS-ISR 0.54	TUR-ROU 0.54	RUS-SYR 0.54	RUS-POL 0.54
UKR-SVK 0.54	TUR-GEO 0.53	RUS-ROU 0.53	PSE-PAK 0.53	RUS-KOR 0.53

Dyads with lowest correlations

Table 2: Fifty dyads with lowest average correlation on total counts

MEX-SAU -0.0090	AUS-ITA -0.0086	GBR-VEN -0.0060	ISR-BGD -0.0060	AFG-SYR -0.0050
BRA-POL -0.0047	AFG-LKA -0.0045	SAU-NZL -0.0043	AUS-CZE -0.0042	CZE-LKA -0.0038
IDN-AZE -0.0037	ITA-NZL -0.0031	PRK-SAU -0.0030	IRQ-ZWE -0.0030	IND-ARG -0.0029
NPL-CAN -0.0028	PHL-LKA -0.0028	BRA-ITA -0.0027	VNM-SAU -0.0025	ESP-MYS -0.0025
NGA-LBN -0.0025	NGA-ITA -0.0025	PHL-ARG -0.0024	PSE-GEO -0.0024	IRN-NPL -0.0023
AZE-MYS -0.0022	GEO-SYR -0.0022	EGY-MEX -0.0022	BGD-SYR -0.0021	CAN-NZL -0.0020
TWN-EGY -0.0020	PRK-KEN -0.0019	COL-BGD -0.0018	PRK-LBN -0.0018	EGY-VEN -0.0018
CZE-VEN -0.0016	KOR-GEO -0.0016	KOR-VEN -0.0015	TUR-VEN -0.0015	NGA-VNM -0.0015
PHL-KEN -0.0015	SVK-SAU -0.0015	AFG-BRA -0.0015	SVK-ZWE -0.0015	AFG-VEN -0.0015
GEO-SAU -0.0015	KOR-ZWE -0.0015	SYR-ARG -0.0015	PSE-MEX -0.0014	ZAF-NZL -0.0014

TABARI vs PETRARCH

Table 3: Twenty dyads with highest weighted average correlation

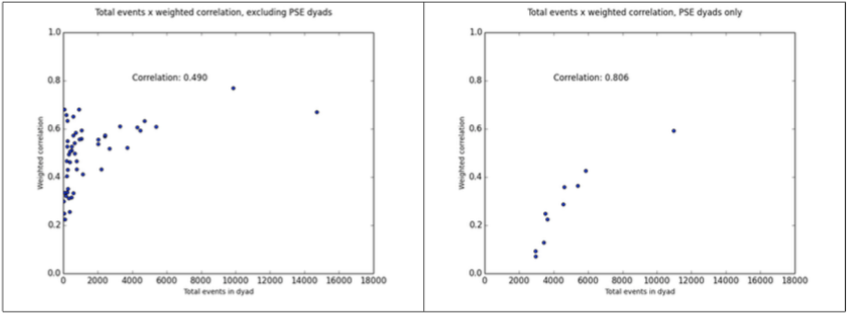
ISR-LBN (9871) 0.7684	ISR-PSE (39655) 0.7554	JOR-TUR (75) 0.6798	EGY-SYR (924) 0.6798
ISR-USA (14722) 0.6689	JOR-FRA (188) 0.6567	SYR-JOR (591) 0.6503	EGY-TUR (251) 0.6327
EGY-USA (4727) 0.6318	LBN-USA (3300) 0.6096	ISR-EGY (5399) 0.608	SYR-USA (4301) 0.6054
ISR-GBR (1075) 0.5929	ISR-IGO (4480) 0.5923	PSE-USA (10980) 0.5914	EGY-JOR (737) 0.583
JOR-USA (2435) 0.5724	EGY-FRA (594) 0.5718	ISR-JOR (2424) 0.5682	ISR-FRA (1068) 0.558

Table 4: Twenty dyads with lowest weighted average correlation

LBN-DEU (219) 0.403	PSE-IGO (5414) 0.3631	PSE-JOR (4632) 0.3577	USA-DEU (282) 0.3505
IGO-TUR (243) 0.3361	FRA-GBR (90) 0.3343	ISR-DEU (599) 0.3326	LBN-JOR (166) 0.321
USA-FRA (492) 0.3146	IGO-GBR (335) 0.3111	TUR-DEU (38) 0.2983	PSE-LBN (4574) 0.2861
IGO-FRA (384) 0.2549	LBN-TUR (61) 0.248	PSE-FRA (3532) 0.2473	PSE-SYR (3654) 0.2237
IGO-DEU (106) 0.2235	PSE-GBR (3445) 0.1275	PSE-TUR (2964) 0.0919	PSE-DEU (2973) 0.0701

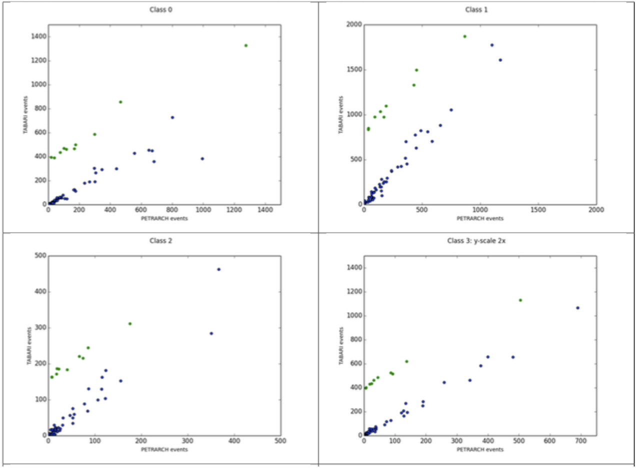
TABARI vs PETRARCH: High frequency dyads generally have higher correlations

Table 5: Total counts by weighted correlation by dyad.



TABARI vs PETRARCH: Palestine is an outlier

Table 7: Total counts by dyad, excluding ISR-PSE. Green markers are dyads involving PSE; blue are all other dyads.



What is to be done: Part 1

- Open-access gold standard cases, then use the estimated classification matrices for statistical adjustments
- Systematically assess the trade-offs in multiple-source data, or create more sophisticated filters
- Evaluate the utility of multiple-data-set methods such as multiple systems estimation
- Systematic assessment of the native language versus machine translation issue
- Extend CAMEO and standardize sub-state actor codes: canonical CAMEO is too complicated, but ICEWS substate actors are too simple

What is to be done: Part 2

- Automated verb phrase recognition and extraction: this will also be required to extend CAMEO. Entity identification, in contrast, is largely a solved problem (ICEWS: 100,000 actors in dictionary)
- Establish a user-friendly open-source collaboration platform for dictionary development
- Systematically explore aggregation methods: ICEWS has 10,742 aggregations, which is too many
- Solve—or at least improve upon—the open source geocoding issue
- Develop event-specific coding modules

Thank you

Email:
schrodt735@gmail.com

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

Data: `http://phoenixdata.org`

Software: `https://openeventdata.github.io/`

Papers:
`http://eventdata.parusanalytics.com/papers.html`