Comparison Metrics for Large Scale Political Event Data Sets

Philip A. Schrodt

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

Paper presented at the New Directions in Text as Data
New York University, 16-17 October 2015

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

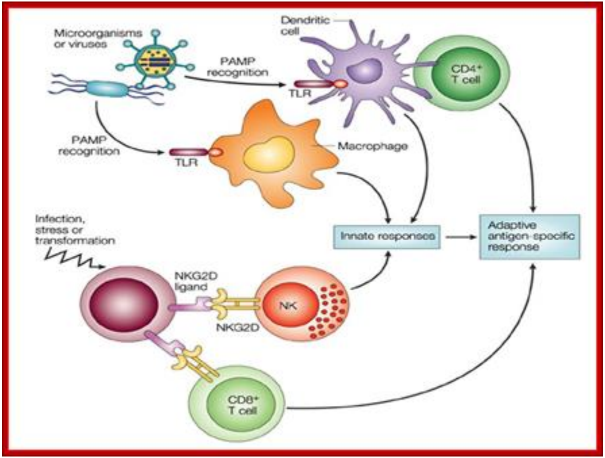
Outline

- 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
- Example 3: Generate data using simple pattern matching and “bag of words” methods
- Next steps

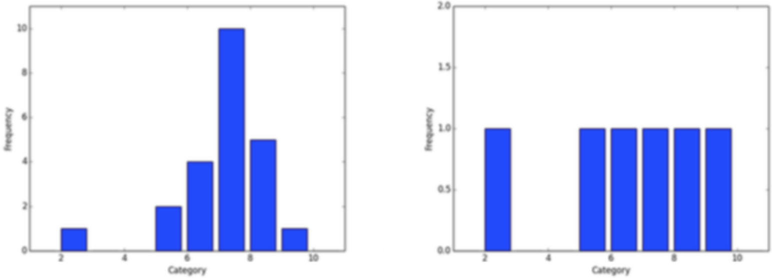
Humans use multiple sources to create narratives

- 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

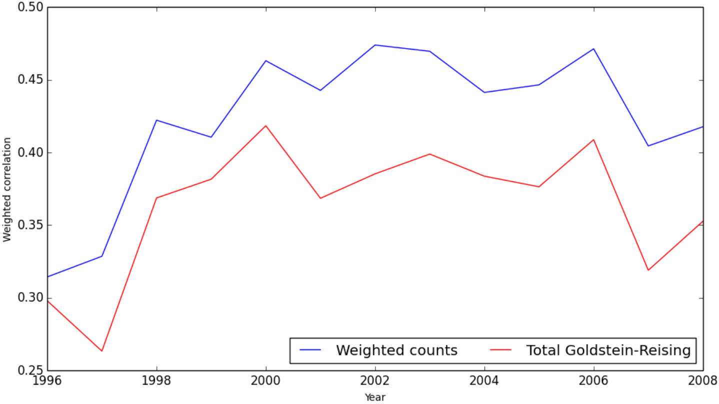
Weighted correlation between two data sets

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

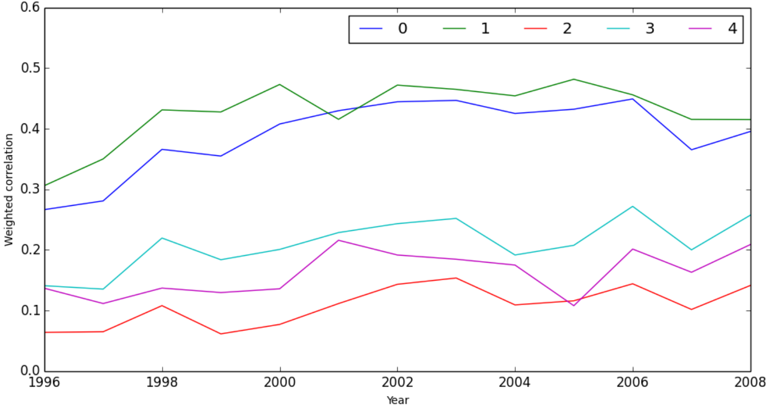
where

- 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

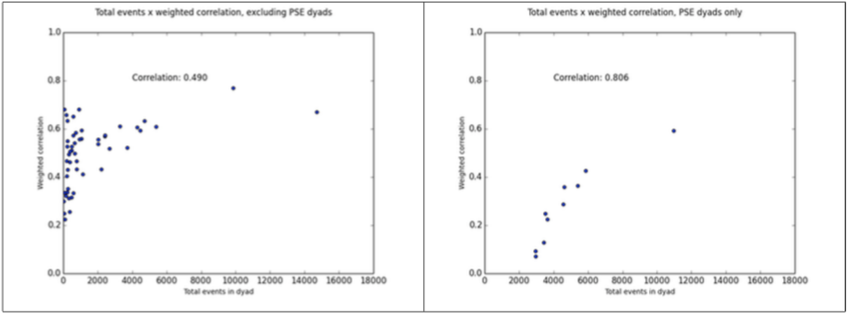
ISR-LBN (9871) 0.7684	ISR-PSE (39655) 0.7554	JOR-TUR (75) 0.6798	EGY-SYR (924) 0.6798
ISR-USA (14722) 0.6689	JOR-FRA (188) 0.6567	SYR-JOR (591) 0.6503	EGY-TUR (251) 0.6327
EGY-USA (4727) 0.6318	LBN-USA (3300) 0.6096	ISR-EGY (5399) 0.608	SYR-USA (4301) 0.6054
ISR-GBR (1075) 0.5929	ISR-IGO (4480) 0.5923	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

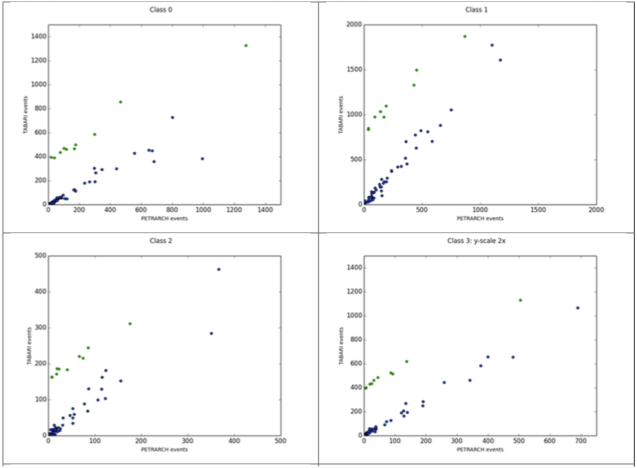
TABARI vs PETRARCH: High frequency dyads generally have higher correlations

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.



Experimenting with minimal "bag of words" approaches

- PETRARCH AFP and Reuters Levant data is the reference set
- Actors and agents: simply look for the patterns found in generic dictionaries
- Events: use support vector machines on lede-sentence texts to classify these into pentacodes
  - Experiment 1: train on 400 cases, test on remainder
  - Experiment 2: train on first half of cases, test on remainder

Pattern-based recognition of actors and agents

Table 8: AFP results [N = 86,450]

alpha-3	70271	81.29%
alpha-6	30367	35.13%
none found	5968	6.90%
order-3	11705	27.08%
order-6	0	0.00%

Table 9: Reuters results [N = 31,256]

alpha-3	25538	81.71%
alpha-6	10212	32.67%
none found	2314	7.40%
order-3	4184	26.78%
order-6	0	0.00%

SVM event classification: 400 training cases for each category

Table 11: Test set: AFP remaining cases
Correct: 33.48%

	0	1	2	3	4	5	True cases	Category accuracy
0	3860	1183	1555	1476	1496	2538	12108 (10.50%)	31.88%
1	2558	5466	2325	1631	1855	2623	16458 (14.27%)	33.21%
2	413	276	1316	404	438	713	3560 (3.09%)	36.97%
3	816	383	562	2554	667	1037	6019 (5.22%)	42.43%
4	696	431	948	945	2948	1824	7792 (6.76%)	37.83%
5	3682	2232	4231	3065	4958	10723	28891 (25.06%)	37.12%
6	5699	3661	5905	5460	8020	11737	40482 (35.11%)	28.99%

SVM event classification: 50% training cases for AFP

Table 14: Training set: AFP 2005-2009
Correct: 63.12%

	0	1	2	3	4	5	Category accuracy
0	5120	1424	161	397	426	1585	56.18%
1	893	10656	149	216	236	1124	80.28%
2	344	537	923	105	228	719	32.32%
3	557	447	72	2561	255	696	55.82%
4	403	498	108	303	3479	1323	56.90%
5	1225	1732	259	613	1202	8473	62.74%

Table 15: Test set: AFP 2010-2014
Correct: 53.4%

	0	1	2	3	4	5	True cases	Category accuracy
0	1950	633	93	239	209	1012	4136 (11.27%)	47.15%
1	426	3371	98	168	131	699	4893 (13.33%)	68.89%
2	183	196	207	50	94	374	1104 (3.01%)	18.75%
3	265	251	29	954	143	365	2007 (5.47%)	47.53%
4	212	260	58	192	945	613	2280 (6.21%)	41.45%
5	856	769	246	550	839	6046	9306 (25.35%)	64.97%
6	1830	2058	353	994	1625	6119	12979 (35.36%)	47.15%

SVM event classification: 50% training cases for Reuters

Table 16: Training set: Reuters 2005-2009
Correct: 76.85%

	0	1	2	3	4	5	Category accuracy
0	2051	259	47	80	132	307	71.31
1	177	2650	15	30	28	134	87.34
2	55	53	775	10	34	79	77.04
3	114	61	15	820	54	103	70.27
4	106	48	24	42	1581	196	79.17
5	345	234	55	88	246	2651	73.25

Table 17: Test set: Reuters 2010-2014
Correct: 45%

	0	1	2	3	4	5	True cases	Category accuracy
0	1098	491	163	147	178	583	2660 (11.14%)	41.28%
1	418	1274	130	74	142	441	2479 (10.38%)	51.39%
2	155	127	175	55	74	249	835 (3.50%)	20.96%
3	206	127	63	257	98	235	986 (4.13%)	26.06%
4	147	107	76	95	459	327	1211 (5.07%)	37.90%
5	997	866	549	314	776	4370	7872 (32.97%)	55.51%
6	1480	1114	483	506	1132	3120	7835 (32.81%)	39.82%

OEDA NSF RIDIR Project

- Sustained support for the Phoenix real-time data
- Long time-frame data sets based on Lexis-Nexis
- Open-access gold standard cases
- Coding systems in Spanish and Arabic, possibly extended to French and Chinese
- Further improvements in automated geolocation
- Automated dictionary development tools
- Extend CAMEO and standardize sub-state actor codes: canonical CAMEO is too complicated, but ICEWS substate actors are too simple
- Develop event-specific coding modules, starting with protests

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`