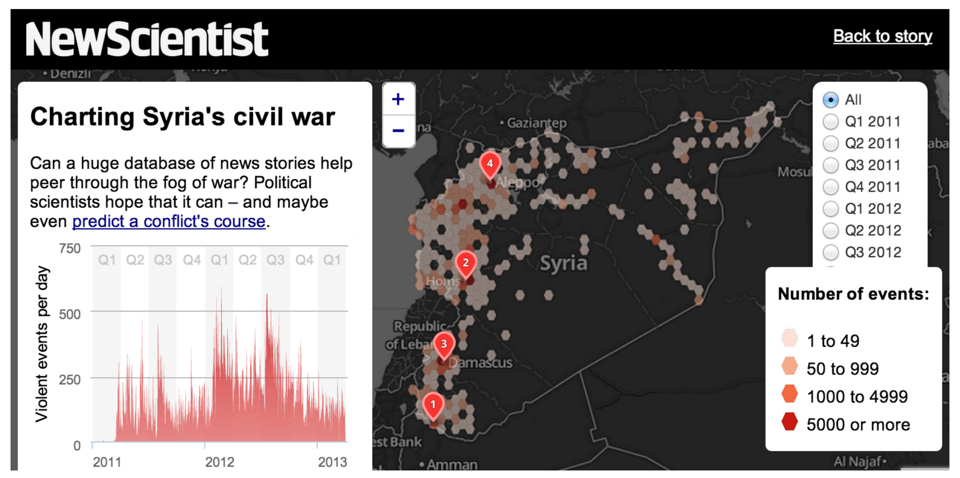
Learning to Extract International Relations from News Text

Brendan O'Connor,[†] Brandon M. Stewart,[‡] Noah A. Smith[†]

[†] School of Computer Science, Carnegie Mellon University [‡] Government Department, Harvard University

More information: http://brenocon.com/irevents



Event data in international relations

What are the causes of war and peace? Do democracies engage in fewer wars? Why do some crises spiral into conflict, but others are resolved peacefully? Can we forecast future conflicts?

To help answer these questions, political scientists use *event data*: historical datasets of friendly and hostile interactions between countries, as reported in news articles. How can we extract this structured information, from millions of news articles?

Left: visualization of GDELT data (subsetted to the Syria conflict). The core of GDELT's event extraction is rule-based (the TABARI software package). http://gdelt.utdallas.edu

Previous work: knowledge engineering

Besides manual coding (which is too labor-intensive at scale), previous work in political science uses a knowledge engineering approach: a manually defined ontology of event types and 15,000 textual patterns to identify events. This took decades of knowledge engineering to construct. it is very difficult to maintain and must be completely rebuilt for new domains (e.g. domestic politics, commercial news, literature...)

We seek to automate some of this process: from the textual data, is it possible to automatically learn the semantic event types, and extract meaningful real-world political dynamics?

Our approach: learning both event types and political dynamics

Text Data

6.5 million news articles, 1987-2008

"Pakistan promptly accused India"
[AP, 1/1/2000]

Preprocess with:

1. Syntactic Parsing
Stanford Parser/Dependencies. Predicate as dependency path between verb arguments. Only use main verbs of sentences.

2. Named Entity Identification
Noun phrases that match lexicon of country names from previous work.

This pipeline is designed to be high precision, low recall.

Event Tuples

Timestep (week):	268
Source (~subject):	PAK
Receiver (~object):	IND
Predicate path (~verb):	**accuse**(*subj=Src, dobj=Rec*)

Every pair of countries has time-series of verb events (based on article timestamps).

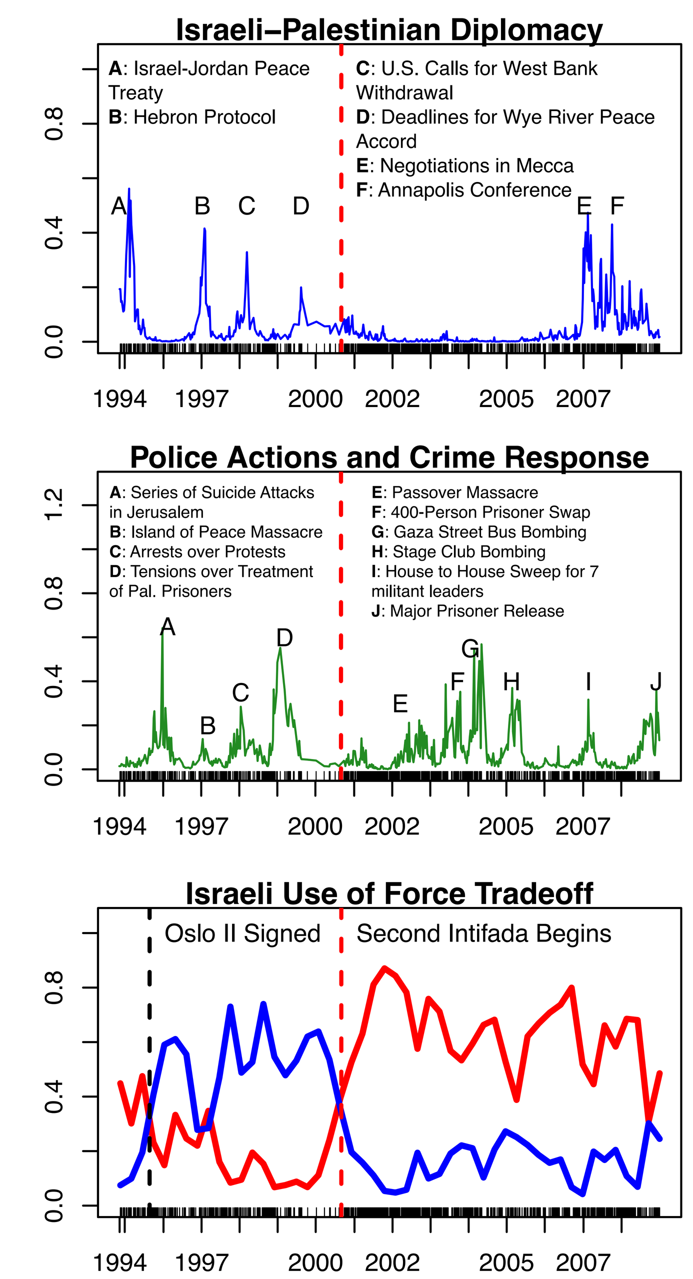
Learn a
Bayesian latent variable model

Model Inferences

Event types (ϕ):

An event type is a (soft) cluster of verbs.
Below: example clusters discovered by our model.

"diplomacy"	arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to, praise
"verbal conflict"	accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say←ccomp come from, say ←ccomp, suspect, slam, accuse government ←poss, accuse agency ←poss, criticize, identify
"material conflict"	kill in, have troops in, die in, be in, wound in, have soldier in, hold in, kill in attack in, remain in, detain in, have in, capture in, stay in, about ←pobj troops in, kill, have troops ←partmod station in, station in, injure in, invade, shoot in



meet with, sign with, praise, say with, arrive in, host, tell, welcome, join, thank

accuse, criticize, reject, tell, hand to, warn, ask, detain, release, order'

kill, fire at, enter, kill, attack, raid, strike, move, pound, bomb

impose, seal, capture, seize, arrest, ease, close, deport, close, release

Dyadic relations (θ):

Every pair of countries has time-series of event type probabilities.

Left: event type probability time-series (θ).
Right: Verbs for the event class (ϕ).

Qualitative evaluation: Case study

The model's inferences about Israeli-Palestinian relations correspond to important events in the historical record.

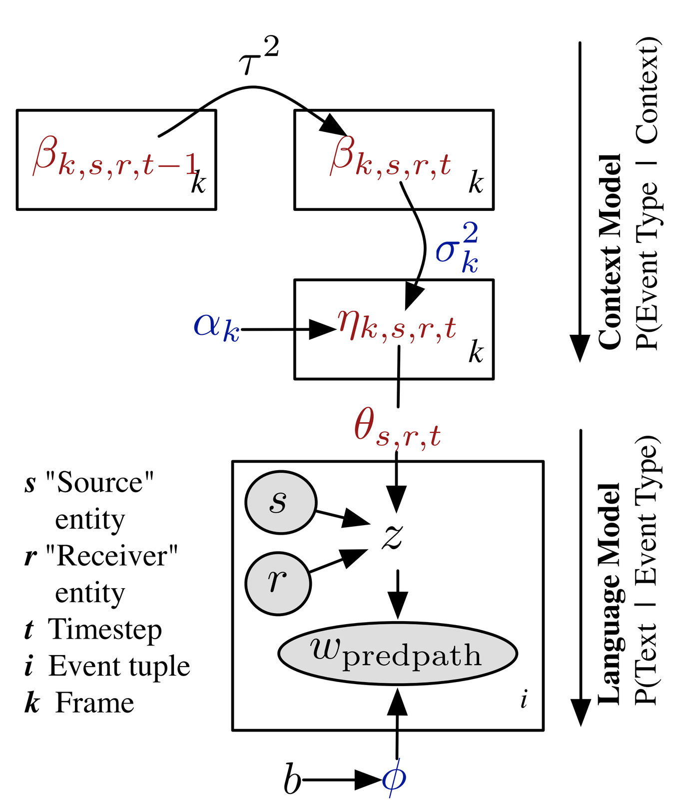
Model

The key assumption is
dyadic and temporal coherence,
that a pair of countries tends to have similar event types during one time period (and nearby time periods).

This causes event type's verb clusters to reflect real-world co-occurrences, which are often semantically meaningful. Thus **social context drives semantic learning**.

This is encoded as a logistic normal admixture model (i.e. a type of "topic model", for dependency paths in a particular time-dyad slice).

Training is with blocked Gibbs sampling (MCMC).



s "Source" entity
r "Receiver" entity
t Timestep
i Event tuple
k Frame

Context model (smoothed frames):

$$\tau^2 \sim \text{InvGamma}$$
$$\sigma_k^2 \sim \text{InvGamma}$$
$$\alpha_k \sim \text{Normal}$$
$$\beta_{s,r,1,k} \sim N(0, 100)$$
$$\beta_{s,r,t>1,k} \sim N(\beta_{k,s,r,t-1}, \tau^2)$$
$$\eta_{s,r,t,k} \sim N(\alpha_k + \beta_{k,s,r,t}, \sigma_k^2)$$
$$\theta_{s,r,t,*} = \text{Softmax}(\eta_{s,r,t,*})$$

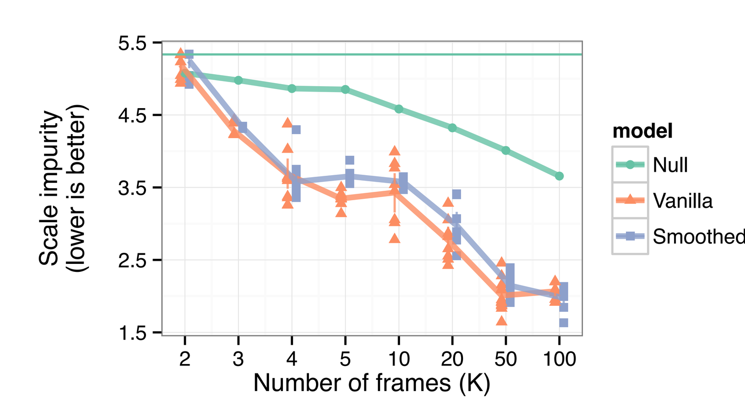
Language model:

$$b \sim \text{ImproperUniform}$$
$$\phi_k \sim \text{Dir}(b/V)$$
$$z \sim \theta_{s,r,t}$$
$$w \sim \phi_z$$

Quantitative evaluations

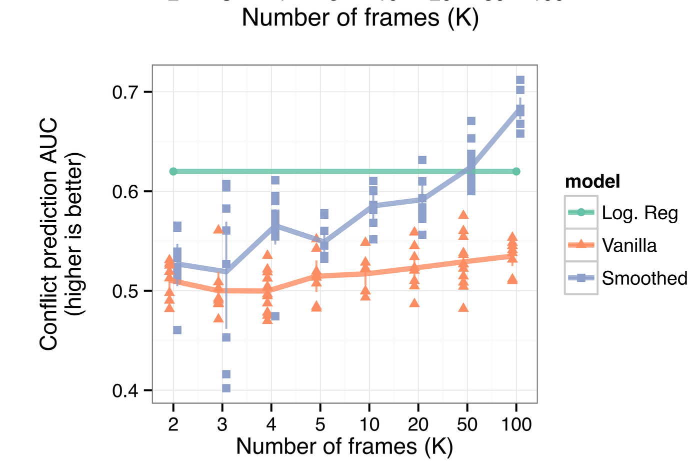
Does the learned ontology match one designed by experts?

Compare verb clusters to manually defined ones in previous work (rule patterns from TABARI).



Does the model predict conflict?

Use the model's inferred political dynamics to predict whether a conflict is happening between countries, as defined by the Militarized Interstate Dispute dataset from political science.



Conclusions

Our method simultaneously
(1) extracts a database of political events
(2) infers latent sociopolitical context
(3) organizes insightful summaries of this large and high-dimensional textual data.

Next steps include semi-supervised methods to exploit previously built knowledge bases, which will greatly help political science researchers, the incorporation of temporal and location textual analysis, and discovery of new actors and their properties.

More generally, ***event data analysis*** from political science is an interesting and exciting application area of NLP. It combines traditional concerns in *text mining* with *information extraction* and *semantics*. Numerous techniques and approaches are possible.