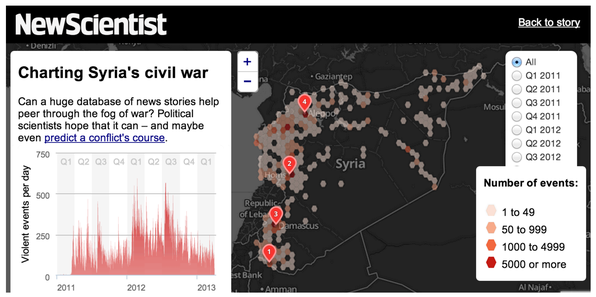
Learning to Extract International Relations from News Text

Presenter: Brendan O'Connor, Carnegie Mellon University

Joint work with Brandon Stewart (political scientist at Harvard) and Noah Smith (CMU).

Forthcoming at ACL 2013. See more: 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 for the Syria conflict, which extracts events from news using a knowledge engineering approach. 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. This 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

Data

6.5 million news articles, 1987-2008

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

Preprocess with **syntactic parsing** and **named entity identification**

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

Textual event tuples:
Every pair of countries has time-series of verb events.

Learn a **Bayesian latent variable model**, which simultaneously discovers:

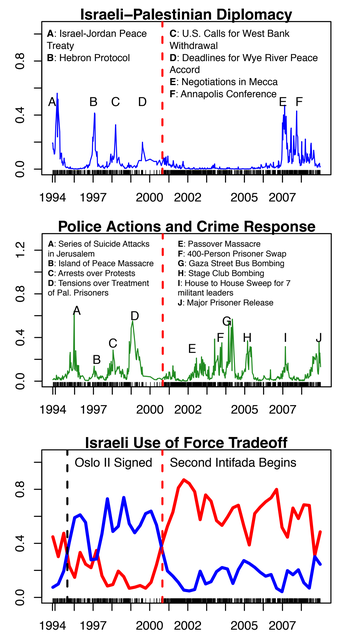
"diplomacy": arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to, praise

"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

Event types (ϕ):
An event type is a (soft) cluster of verbs.
Above: example clusters discovered by our model.

Dyadic relations (θ):
Every pair of countries has time-series of event type probabilities.



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

Qualitative evaluation: Case study. Israeli-Palestinian relations. These are inferences from our model.
Left: shows event class probability time-series.
Right: Verbs for the event class.

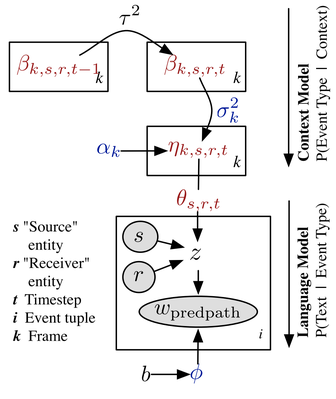
Model

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

This causes event type's verb clusters to reflect real-world co-occurrences, which are often semantically meaningful.

Mathematically, this is encoded as a logistic normal admixture model (i.e. a type of "topic model").

Training is with blocked Gibbs sampling (a Markov Chain Monte Carlo algorithm).



Context model (smoothed frames):

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

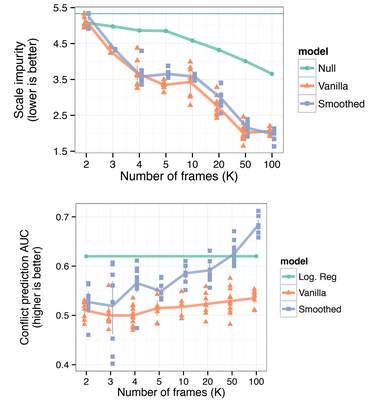
Language model:

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

Quantitative evaluations

Does the automatic ontology match one designed by experts?
Compare verb clusters to manually defined ones in previous work (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.



Conclusions

Our novel method simultaneously (1) extracts a database of political events, (2) infers latent sociopolitical context, and (3) organizes insightful summaries of 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.