Learning to Extract International Relations from Political Context

Brendan O'Connor
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
brenocon@cs.cmu.edu

Brandon M. Stewart
Department of Government
Harvard University
Cambridge, MA 02139, USA
bstewart@fas.harvard.edu

Noah A. Smith
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
nasmith@cs.cmu.edu

Abstract

We describe a new probabilistic model for extracting events between major political actors from news corpora. Our unsupervised model brings together familiar components in natural language processing (like parsers and topic models) with contextual political information—temporal and dyad dependence—to infer latent event classes. We quantitatively evaluate the model's performance on political science benchmarks: recovering expert-assigned event class valences, and detecting real-world conflict. We also conduct a small case study based on our model's inferences.

A supplementary appendix, and replication software/data are available online, at: **http://brenocon.com/irevents**

[This paper is forthcoming in Proceedings of ACL 2013; Sofia, Bulgaria.]

1 Introduction

The digitization of large news corpora has provided an unparalleled opportunity for the systematic study of international relations. Since the mid-1960s political scientists have used political *events data*, records of public micro-level interactions between major political actors of the form "someone does something to someone else" as reported in the open press (Schrodt, 2012), to study the patterns of interactions between political actors and how they evolve over time. Scaling this data effort to modern corpora presents an information extraction challenge: can a structured collection of accurate, politically relevant events between major political actors be extracted automatically and efficiently? And can they be grouped into meaningful event types with a low-dimensional structure useful for further analysis?

We present an unsupervised approach to event extraction, in which political structure and linguistic evidence are combined. A political context model of the relationship between a pair of political actors imposes a prior distribution over types of linguistic events. Our probabilistic model infers latent frames, each a distribution over textual expressions of a kind of event, as well as a representation of the relationship between each political actor pair at each point in time. We use syntactic preprocessing and a logistic normal topic model, including latent temporal smoothing on the political context prior.

We apply the model in a series of comparisons to benchmark datasets in political science. First, we compare the automatically learned verb classes to a pre-existing ontology and hand-crafted verb patterns from TABARI,[1] an open-source and widely used rule-based event extraction system for this domain. Second, we demonstrate correlation to a database of real-world international conflict events, the Militarized Interstate Dispute (MID) dataset (Jones *et al.*, 1996). Third, we qualitatively examine a prominent case not included in the MID dataset, Israeli-Palestinian relations, and compare the recovered trends to the historical record.

We outline the data used for event discovery (§2), describe our model (§3), inference (§4), evaluation (§5), and comment on related work (§6).

2 Data

The model we describe in §3 is learned from a corpus of 6.5 million newswire articles from the English Gigaword 4th edition (1994–2008, Parker *et al.*, 2009). We also supplement it with a sample of data from the *New York Times* Annotated Corpus (1987–2007, Sandhaus, 2008).[2] The Stan-

[1] Available from the Penn State Event Data Project: http://eventdata.psu.edu/

[2] For arbitrary reasons this portion of the data is much smaller (we only parse the first five sentences of each arti-

ford CoreNLP system,[3] under default settings, was used to POS-tag and parse the articles, to eventually produce event tuples of the form

$$\langle s, r, t, w_{\text{predpath}} \rangle$$

where s and r denote "source" and "receiver" arguments, which are political actor entities in a predefined set $\mathcal{E}$, t is a timestep (i.e., a 7-day period) derived from the article's published date, and w_{predpath} is a textual predicate expressed as a dependency path that typically includes a verb (we use the terms "predicate-path" and "verb-path" interchangeably). For example, on January 1, 2000, the AP reported "Pakistan promptly accused India," from which our preprocessing extracts the tuple $\langle$PAK, IND, 678, *accuse* $\xleftarrow{\text{dobj}}$ $\rangle$. (The path excludes the first source-side arc.) Entities and verb paths are identified through the following sets of rules.

Named entity recognition and resolution is done deterministically by finding instances of country names from the *CountryInfo.txt* dictionary from TABARI,[4] which contains proper noun and adjectival forms for countries and administrative units. We supplement these with a few entries for international organizations from another dictionary provided by the same project, and clean up a few ambiguous names, resulting in a final actor dictionary of 235 entities and 2,500 names.

Whenever a name is found, we identify its entity's mention as the minimal noun phrase that contains it; if the name is an adjectival or noun-noun compound modifier, we traverse any such *amod* and *nn* dependencies to the noun phrase head. Thus *NATO bombing*, *British view*, and *Palestinian militant* resolve to the entity codes IGONAT, GBR, and PSE respectively.

We are interested in identifying actions initiated by agents of one country targeted towards another, and hence concentrate on verbs, analyzing the "CCprocessed" version of the Stanford Dependencies (de Marneffe and Manning, 2008). Verb paths are identified by looking at the shortest dependency path between two mentions in a sentence. If one of the mentions is immediately dominated by a *nsubj* or *agent* relation, we consider that the Source actor, and the other mention is the Receiver. The most common cases are simple direct objects and prepositional arguments like *talk* $\xleftarrow{\text{prep\_with}}$ and *fight* $\xleftarrow{\text{prep\_alongside}}$ ("talk with *R*," "fight alongside *R*") but many interesting multiword constructions also result, such as *reject* $\xleftarrow{\text{dobj}}$ *allegation* $\xleftarrow{\text{poss}}$ ("rejected *R*'s allegation") or verb chains as in *offer* $\xleftarrow{\text{xcomp}}$ *help* $\xleftarrow{\text{dobj}}$ ("offer to help *R*").

We wish to focus on instances of directly reported events, so attempt to remove factively complicated cases such as indirect reporting and hypotheticals by discarding all predicate paths for which any verb on the path has an off-path governing verb with a non-*conj* relation. (For example, the verb at the root of a sentence always survives this filter.) Without this filter, the $\langle s, r, w \rangle$ tuple $\langle$USA, CUB, *want* $\xleftarrow{\text{xcomp}}$ *seize* $\xleftarrow{\text{dobj}}$ $\rangle$ is extracted from the sentence "Parliament Speaker Ricardo Alarcon said the United States wants to seize Cuba and take over its lands"; the filter removes it since *wants* is dominated by an off-path verb through *say* $\xleftarrow{\text{ccomp}}$ *wants*. The filter was iteratively developed by inspecting dozens of output examples and their labelings under successive changes to the rules.

Finally, only paths length 4 or less are allowed, the final dependency relation for the receiver may not be *nsubj* or *agent*, and the path may not contain any of the dependency relations *conj*, *parataxis*, *det*, or *dep*. We use lemmatized word forms in defining the paths.

Several document filters are applied before tuple extraction. Deduplication removes 8.5% of articles.[5] For topic filtering, we apply a series of keyword filters to remove sports and finance news, and also apply a text classifier for diplomatic and military news, trained on several hundred manually labeled news articles (using ℓ_1-regularized logistic regression with unigram and bigram features). Other filters remove non-textual junk and non-standard punctuation likely to cause parse errors.

For experiments we remove tuples where the source and receiver entities are the same, and restrict to tuples with dyads that occur at least 500 times, and predicate paths that occur at least 10

cle, while Gigaword has all sentences parsed), resulting in less than 2% as many tuples as from the Gigaword data.

[3] http://nlp.stanford.edu/software/corenlp.shtml

[4] http://eventdata.psu.edu/software.dir/dictionaries.html.

[5] We use a simple form of shingling (ch. 3, Rajaraman and Ullman, 2011): represent a document signature as its $J = 5$ lowercased bigrams with the lowest hash values, and reject a document with a signature that has been seen before within the same month. J was manually tuned, as it affects the precision/recall tradeoff.

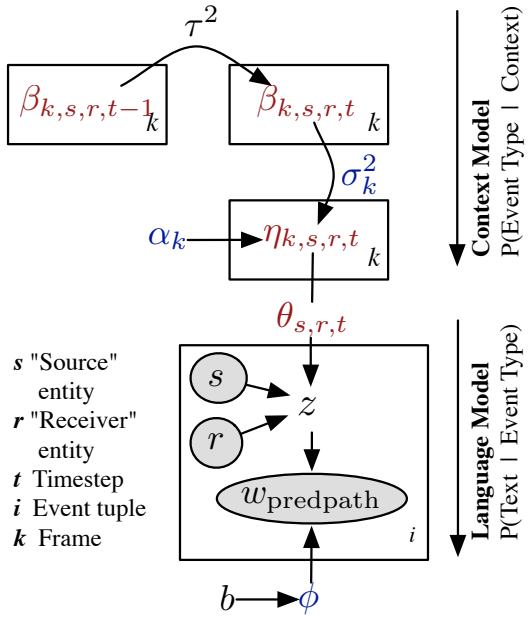

Figure 1: Directed probabilistic diagram of the model for one (s, r, t) dyad-time context, for the smoothed model.

times. This yields 365,623 event tuples from 235,830 documents, for 421 dyads and 10,457 unique predicate paths. We define timesteps to be 7-day periods, resulting in 1,149 discrete timesteps (1987 through 2008, though the vast majority of data starts in 1994).

3 Model

We design two models to learn linguistic event classes over predicate paths by conditioning on real-world contextual information about international politics, $p(w_{\text{predpath}} \mid s, r, t)$, leveraging the fact there tends to be dyadic and temporal coherence in international relations: the types of actions that are likely to occur between nations tend to be similar within the same dyad, and usually their distribution changes smoothly over time.

Our model decomposes into two submodels: a Context submodel, which encodes how political context affects the probability distribution over event types, and a Language submodel, for how those events are manifested as textual predicate paths (Figure 1). The overall generative process is as follows. We color global parameters for a frame blue, and local context parameters red, and use the term "frame" as a synonym for "event type." The fixed hyperparameter K denotes the number of frames.

- The **context model** generates a frame prior $\theta_{s,r,t}$ for every context (s, r, t).
- **Language model:**
  - Draw lexical sparsity parameter b from a diffuse prior (see §4).
  - For each frame k, draw a multinomial distribution of dependency paths, $\phi_k \sim \text{Dir}(b/V)$ (where V is the number of dependency path types).
  - For each (s, r, t), for every event tuple i in that context,
    - Sample its frame $z^{(i)} \sim \text{Mult}(\theta_{s,r,t})$.
    - Sample its predicate realization $w^{(i)}_{\text{predpath}} \sim \text{Mult}(\phi_{z^{(i)}})$.

Thus the language model is very similar to a topic model's generation of token topics and wordtypes.

We use structured logistic normal distributions to represent contextual effects. The simplest is the **vanilla** (**v**) context model,

- For each frame k, draw global parameters from diffuse priors: prevalence α_k and variability σ_k^2.
- For each (s, r, t),
  - Draw $\eta_{k,s,r,t} \sim N(\alpha_k, \sigma_k^2)$ for each frame k.
  - Apply a softmax transform,

$$\theta_{k,s,r,t} = \frac{\exp \eta_{k,s,r,t}}{\sum_{k'=1}^{K} \exp \eta_{k',s,r,t}}$$

Thus the vector $\eta_{*,s,r,t}$ encodes the relative log-odds of the different frames for events appearing in the context (s, r, t). This simple logistic normal prior is, in terms of topic models, analogous to the asymmetric Dirichlet prior version of LDA in Wallach *et al.* (2009), since the α_k parameter can learn that some frames tend to be more likely than others. The variance parameters σ_k^2 control admixture sparsity, and are analogous to a Dirichlet's concentration parameter.

Smoothing Frames Across Time

The vanilla model is capable of inducing frames through dependency path co-occurences, when multiple events occur in a given context. However, many dyad-time slices are very sparse; for example, most dyads (all but 18) have events in fewer than half the time slices in the dataset. One solution is to increase the bucket size (e.g., to months); however, previous work in political science has demonstrated that answering questions of interest about reciprocity dynamics requires recovering the events at weekly or even daily granularity (Shellman, 2004), and in any case wide

buckets help only so much for dyads with fewer events or less media attention. Therefore we propose a **smoothed frames** (**SF**) model, in which the frame distribution for a given dyad comes from a latent parameter $\beta_{*,s,r,t}$ that smoothly varies over time. For each (s, r), draw the first timestep's values as $\beta_{k,s,r,1} \sim N(0, 100)$, and for each context $(s, r, t > 1)$,

- Draw $\beta_{k,s,r,t} \sim N(\beta_{k,s,r,t-1},\ \tau^2)$
- Draw $\eta_{k,s,r,t} \sim N(\alpha_k + \beta_{k,s,r,t},\ \sigma_k^2)$

Other parameters (α_k, σ_k^2) are same as the vanilla model. This model assumes a random walk process on β, a variable which exists even for contexts that contain no events. Thus inferences about η will be smoothed according to event data at nearby timesteps. This is an instance of a linear Gaussian state-space model (also known as a linear dynamical system or dynamic linear model), and is a convenient formulation because it has well-known exact inference algorithms. Dynamic linear models have been used elsewhere in machine learning and political science to allow latent topic frequencies (Blei and Lafferty, 2006; Quinn *et al.*, 2010) and ideological positions (Martin and Quinn, 2002) to smoothly change over time, and thus share statistical strength between timesteps.

4 Inference

After randomly initializing all $\eta_{k,s,r,t}$, inference is performed by a blocked Gibbs sampler, alternating resamplings for three major groups of variables: the language model (z,ϕ), context model (α, γ, β, p), and the η, θ variables, which bottleneck between the submodels.

The **language model** sampler sequentially updates every $z^{(i)}$ (and implicitly ϕ via collapsing) in the manner of Griffiths and Steyvers (2004): $p(z^{(i)}|\theta, w^{(i)}, b) \propto \theta_{s,r,t,z}(n_{w,z} + b/V)/(n_z + b)$, where counts n are for all event tuples besides i.

For the **context model**, α is conjugate resampled as a normal mean. The random walk variables β are sampled with the forward-filtering-backward-sampling algorithm (FFBS; Harrison and West, 1997; Carter and Kohn, 1994); there is one slight modification of the standard dynamic linear model that the zero-count weeks have no η observation; the Kalman filter implementation is appropriately modified to handle this.

The η update step is challenging since it is a nonconjugate prior to the z counts. Logistic normal distributions were introduced to text modeling by Blei and Lafferty (2007), who developed a variational approximation; however, we find that experimenting with different models is easier in the Gibbs sampling framework. While Gibbs sampling for logistic normal priors is possible using auxiliary variable methods (Mimno *et al.*, 2008; Holmes and Held, 2006; Polson *et al.*, 2012), it can be slow to converge. We opt for the more computationally efficient approach of Zeger and Karim (1991) and Hoff (2003), using a Laplace approximation to $p(\eta \mid \bar{\eta}, \Sigma, z)$, which is a mode-centered Gaussian having inverse covariance equal to the unnormalized log-posterior's negative Hessian (§8.4 in Murphy, 2012). We find the mode with the linear-time Newton algorithm from Eisenstein *et al.* (2011), and sample in linear time by only using the Hessian's diagonal as the inverse covariance (i.e., an axis-aligned normal), since a full multivariate normal sample requires a cubic-time-to-compute Cholesky root of the covariance matrix. This η^* sample is a proposal for a Metropolis-within-Gibbs step, which is moved to according to the standard Metropolis-Hastings acceptance rule. Acceptance rates differ by K, ranging approximately from 30% ($K = 100$) to nearly 100% (small K).

Finally, we use diffuse priors on all global parameters, conjugate resampling variances τ^2, σ_k once per iteration, and slice sampling (Neal, 2003) the Dirichlet concentration b every 100 iterations. Automatically learning these was extremely convenient for model-fitting; the only hyperparameter we set manually was K. It also allowed us to monitor the convergence of dispersion parameters to help debug and assess MCMC mixing. For other modeling and implementation details, see the online appendix and software.

5 Experiments

We fit the two models on the dataset described in §2, varying the number of frames K, with 8 or more separate runs for each setting. Posteriors are saved and averaged from 11 Gibbs samples (every 100 iterations from 9,000 to 10,000) for analysis.

We present intrinsic (§5.1) and extrinsic (§5.2) quantitative evaluations, and a qualitative case study (§5.4).

5.1 Lexical Scale Impurity

In the international relations literature, much of the analysis of text-based events data makes use of a unidimensional conflict to cooperation scale. A popular event ontology in this domain, CAMEO, consists of around 300 different event types, each given an expert-assigned scale in the range from -10 to $+10$ (Gerner *et al.*, 2002), derived from a judgement collection experiment in Goldstein (1992). The TABARI pattern-based event extraction program comes with a list of almost 16,000 manually engineered verb patterns, each assigned to one CAMEO event type.

It is interesting to consider the extent to which our unsupervised model is able to recover the expert-designed ontology. Given that many of the categories are very fine-grained (e.g. "Express intent to de-escalate military engagement"), we elect to measure model quality as *lexical scale purity*: whether all the predicate paths within one automatically learned frame tend to have similar gold-standard scale scores. (This measures cluster cohesiveness against a one-dimensional continuous scale, instead of measuring cluster cohesiveness against a gold-standard clustering as in VI, Rand index, or purity.) To calculate this, we construct a mapping between our corpus-derived verb path vocabulary and the TABARI verb patterns, many of which contain one to several word stems that are intended to be matched in surface order. Many of our dependency paths, when traversed from the source to receiver direction, also follow surface order, due to English's SVO word order.[6] Therefore we convert each path to a word sequence and match against the TABARI lexicon—plus a few modifications for differences in infinitives and stemming—and find 528 dependency path matches. We assign each path w a gold-standard scale $g(w)$ by resolving through its matching pattern's CAMEO code.

We formalize *lexical scale impurity* as the average absolute difference of scale values between two predicate paths under the same frame. Specifically, we want a token-level posterior expectation

$$\mathbb{E}(|g(w_i) - g(w_j)| \mid z_i = z_j, w_i \neq w_j) \quad (1)$$

which is taken over pairs of path instances (i, j) where both paths w_i, w_j are in M, the set of verb paths that were matched between the lexicons. This can be reformulated at the type level as:[7]

$$\frac{1}{N} \sum_k \sum_{\substack{w,v \in M \\ w \neq v}} n_{w,k}\, n_{v,k}\, |g(w) - g(v)| \quad (2)$$

where n refers to the averaged Gibbs samples' counts of event tuples having frame k and a particular verb path,[8] and N is the number of token comparisons (i.e. the same sum, but with a 1 replacing the distance). The worst possible impurity is upper bounded at 20 ($= \max(g(w)) - \min(g(w))$) and the best possible is 0. We also compute a randomized null hypothesis to see how low impurity can be by chance: each of $\sim$1000 simulations randomly assigns each path in M to one of K frames (all its instances are exclusively assigned to that frame), and computes the impurity. On average the impurity is same at all K, but variance increases with K (since small clusters might by chance get a highly similar paths in them), necessitating this null hypothesis analysis. We report the 5th percentile over simulations.

5.2 Conflict Detection

Political events data has shown considerable promise as a tool for crisis early warning systems (O'Brien, 2010; Brandt *et al.*, 2011). While conflict forecasting is a potential application of our model, we conduct a simpler prediction task to validate whether the model is learning something useful: based on news text, tell whether or not an armed conflict is *currently* happening. For a gold standard, we use the Militarized Interstate Dispute (MID) dataset (Jones *et al.*, 1996; Ghosn *et al.*, 2004), which documents historical international disputes. While not without critics, the MID data is the most prominent dataset in the field of international relations. We use the Dyadic MIDs, each of which ranks hostility levels between pairs of actors on a five point scale over a date interval; we define conflict to be the top two categories "Use of Force" (4) and "War" (5). We convert the data into a variable $y_{s,r,t}$, the highest hostility level reached by actor s directed towards receiver r in the dispute that overlaps with our 7-day interval t, and want to predict the binary indicator

[6] There are plenty of exceptions where a Source-to-Receiver path traversal can have a right-to-left move, such as dependency edges for posessives. This approach can not match them.

[7] Derivation in supplementary appendix.

[8] Results are nearly identical whether we use counts averaged across samples (thus giving posterior marginals), or simply use counts from a single sample (i.e., iteration 10,000).

$\mathbf{1}\{y_{s,r,t} \geq 4\}$. For the illustrative examples (USA to Iraq, and the Israel-Palestine example below) we use results from a smaller but more internally comparable dataset consisting of the 2 million Associated Press articles within the Gigaword corpus.

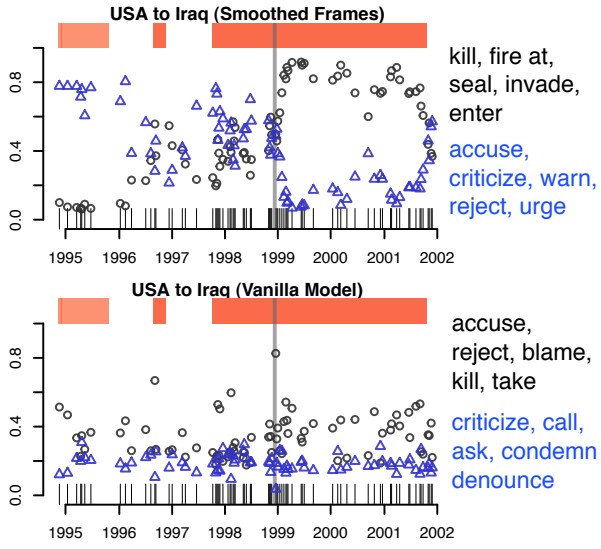

Figure 2: The USA→Iraq directed dyad, analyzed by smoothed (above) and vanilla (below) models, showing (1) gold-standard MID values (red intervals along top), (2) weeks with non-zero event counts (vertical lines along x-axis), (3) posterior $E[\theta_{k,\text{USA},\text{IRQ},t}]$ inferences for two frames chosen from two different $K = 5$ models, and (4) most common verb paths for each frame (right). Frames corresponding to material and verbal conflict were chosen for display. Vertical line indicates Operation Desert Fox (see §5.2).

For an example of the MID data, see Figure 2, which depicts three disputes between the US and Iraq in this time period. The MID labels are marked in red.

The first dispute is a "display of force" (level 3), cataloguing the U.S. response to a series of troop movements along the border with Kuwait. The third dispute (10/7/1997 to 10/10/2001) begins with increasing Iraqi violations of the no-fly zone, resulting in U.S. and U.K. retaliation, reaching a high intensity with Operation Desert Fox, a four-day bombing campaign from December 16 to 19, 1998—which is not shown in MID. These cases highlight MID's limitations—while it is well regarded in the political science literature, its coarse level of aggregation can fail to capture variation in conflict intensity.

Figure 2 also shows model inferences. Our smoothed model captures some of these phenomena here, showing clear trends for two relevant frames, including a dramatic change in December 1998. The vanilla model has a harder time, since it cannot combine evidence between different timesteps.

The MID dataset overlaps with our data for 470 weeks, from 1993 through 2001. After excluding dyads with actors that the MID data does not intend to include—Kosovo, Tibet, Palestine, and international organizations—we have 267 directed dyads for evaluation, 117 of which have at least one dispute in the MID data. (Dyads with no dispute in the MID data, such as Germany-France, are assumed to have $y = 0$ throughout the time period.) About 7% of the dyad-time contexts have a dispute under these definitions.

We split the dataset by time, training on the first half of the data and testing on the second half, and measure area under the receiver operating characteristic curve (AUC).[9] For each model, we train an ℓ_1-regularized logistic regression[10] with the K elements of $\theta_{*,s,r,t}$ as input features, tuning the regularization parameter within the training set (by splitting it in half again) to optimize held-out likelihood. We weight instances to balance positive and negative examples. Training is on all individual θ samples at once (thus accounting for posterior uncertainty in learning), and final predicted probabilities are averaged from individual probabilities from each θ test set sample, thus propagating posterior uncertainty into the predictions. We also create a baseline ℓ_1-regularized logistic regression that uses normalized dependency path counts as the features (10,457 features). For both the baseline and vanilla model, contexts with no events are given a feature vector of all zeros.[11] (We also explored an alternative evaluation setup, to hold out by dyad; however, the performance variance is quite high between different random dyad splits.)

5.3 Results

Results are shown in Figure 3.[12]

The verb-path logistic regression performs strongly at AUC 0.62; it outperforms all of

[9] AUC can be interpreted as follows: given a positive and negative example, what is the probability that the classifier's confidences order them correctly? Random noise or predicting all the same class both give AUC 0.5.

[10] Using the R *glmnet* package (Friedman *et al.*, 2010).

[11] For the vanilla model, this performed better than linear interpolation (about 0.03 AUC), and with less variance between runs.

[12] Due to an implementation bug, the model put the vast majority of the probability mass only on $K - 1$ frames, so these settings might be better thought of as $K = 1, 2, 3, 4, 9, \ldots$; see the appendix for details.

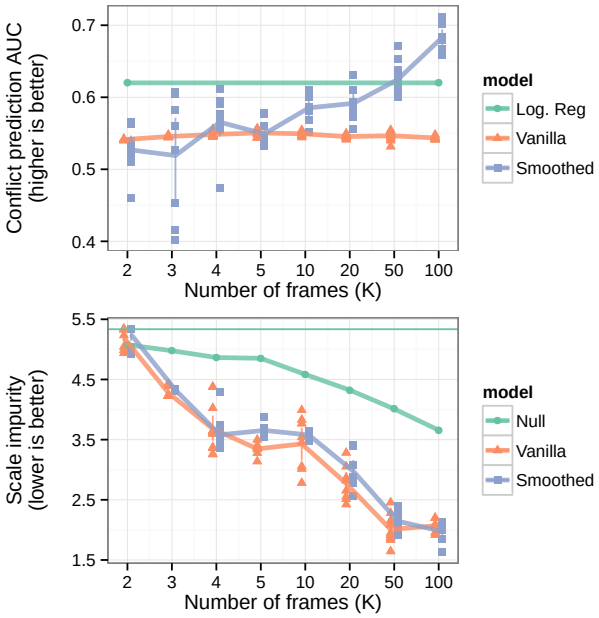

Figure 3: Evaluation results. Each point indicates one model run. Lines show the average per K, with vertical lines indicating the 95% bootstrapped interval. **Top:** Conflict detection AUC for different models (§5.2). Green line is the verb-path logistic regression baseline. **Bottom:** Lexical scale impurity (§5.1). Top green line indicates the simple random baseline $E(|g(w_i) - g(w_j)|) = 5.33$; the second green line is from the random assignment baseline.

the vanilla frame models. This is an example of individual lexical features outperforming a topic model for predictive task, because the topic model's dimension reduction obscures important indicators from individual words. Similarly, Gerrish and Blei (2011) found that word-based regression outperformed a customized topic model when predicting Congressional bill passage, and Eisenstein *et al.* (2010) found word-based regression outperformed Supervised LDA for geolocation,[13] and we have noticed this phenomenon for other text-based prediction problems.

However, adding smoothing to the model substantially increases performance, and in fact outperforms the verb-path regression at $K = 100$. It is unclear why the vanilla model fails to increase performance in K. Note also, the vanilla model exhibits very little variability in prediction performance between model runs, in comparison to the smoothed model which is much more variable (presumably due to the higher number of parameters in the model); at small values of K, the smoothed model can perform poorly. It would also be interesting to analyze the smoothed model with higher values of K and find where it peaks.

We view the conflict detection task only as one of several validations, and thus turn to lexical evaluation of the induced frames. For lexical scale purity (bottom of Figure 3), the models perform about the same, with the smoothed model a little bit worse at some values of K (though sometimes with better stability of the fits—opposite of the conflict detection task). This suggests that semantic coherence does not benefit from the longer-range temporal dependencies.

In general, performance improves with higher K, but not beyond $K = 50$. This suggests the model reaches a limit for how fine-grained of semantics it can learn.

5.4 Case study

Here we qualitatively examine the narrative story between the dyad with the highest frequency of events in our dataset, the Israeli-Palestinian relationship, finding qualitative agreement with other case studies of this conflict (Brandt *et al.*, 2012; Goldstein *et al.*, 2001; Schrodt and Gerner, 2004). (The MID dataset does not include this conflict because the Palestinians are not considered a state actor.) Using the Associated Press subset, we plot the highest incidence frames from one run of the $K = 20$ smoothed frame models, for the two directed dyads, and highlight some of the interesting relationships.

Figure 4(a) shows that tradeoffs in the use of military vs. police action by Israel towards the Palestinians tracks with major historical events. The first period in the data where police actions ('impose, seal, capture, seize, arrest') exceed military actions ('kill, fire, enter, attack, raid') is with the signing of the "Interim Agreement on the West Bank and the Gaza Strip," also known as the Oslo II agreement. This balance persists until the abrupt breakdown in relations that followed the unsuccessful Camp David Summit in July of 2000, which generally marks the starting point of the wave of violence known as the Second Intifada.

In Figure 4(b) we show that our model produces a frame which captures the legal aftermath of particular events ('accuse, criticize,' but also 'detain, release, extradite, charge'). Each of the major spikes in the data coincides with a particular event which either involves the investigation of a particular attack or series of attacks (as in A,B,E) or a discussion about prisoner swaps or mass arrests (as in events D, F, J).

Our model also picks up positive diplomatic

[13] In the latter, a problem-specific topic model did best.

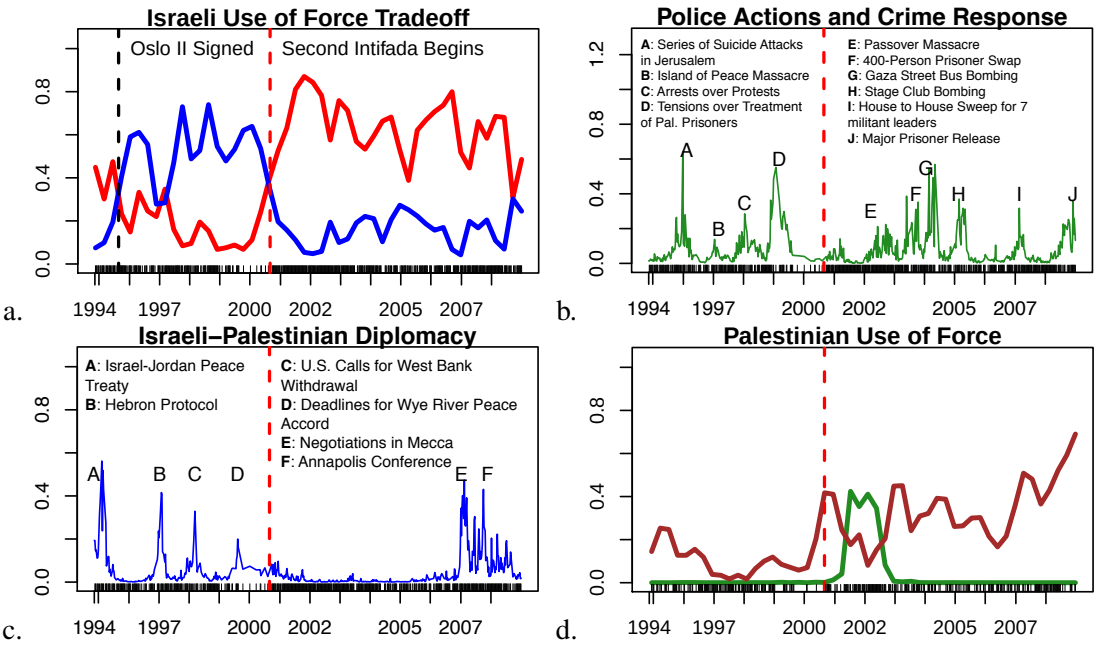

Figure 4: For Israel-Palestinian directed dyads, plots of $E[\theta]$ (proportion of weekly events in a frame) over time, annotated with historical events. (a): Words are 'kill, fire at, enter, kill, attack, raid, strike, move, pound, bomb' and 'impose, seal, capture, seize, arrest, ease, close, deport, close, release' (b): 'accuse, criticize, reject, tell, hand to, warn, ask, detain, release, order' (c): 'meet with, sign with, praise, say with, arrive in, host, tell, welcome, join, thank' (d): again the same 'kill, fire at' frame in (a), plus the erroneous frame (see text) 'include, join, fly to, have relation with, protest to, call, include bomber $\xleftarrow{\text{appos}}$ informer for'. Figures (b) and (c) use linear interpolation for zero-count weeks (thus relying exclusively on the model for smoothing); (a) and (d) apply a lowess smoother. (a-c) are for the ISR→PSE direction; (d) is PSE→ISR.

events, as seen in Figure 4(c), a frame describing Israeli diplomatic actions towards Palestine ('meet with, sign with, praise, say with, arrive in'). Not only do the spikes coincide with major peace treaties and negotiations, but the model correctly characterizes the relative lack of positively valenced action from the beginning of the Second Intifada until its end around 2005–2006.

In Figure 4(d) we show the relevant frames depicting use of force from the Palestinians towards the Israelis (brown trend line). At first, the drop in the use of force frame immediately following the start of the Second Intifada seems inconsistent with the historical record. However, there is a concucrrent rise in a different frame driven by the word 'include', which actually appears here due to an NLP error compounded with an artifact of the data source. A casualties report article, containing variants of the text "The Palestinian figure includes... 13 Israeli Arabs...", is repeated 27 times over two years. "Palestinian figure" is erroneously identified as the PSE entity, and several noun phrases in a list are identified as separate receivers. This issue causes 39 of all 86 PSE→ISR events during this period to use the word 'include', accounting for the rise in that frame. (This highlights how better natural language processing could help the model, and the dangers of false positives for this type of data analysis, especially in small-sample drilldowns.) Discounting this erroneous inference, the results are consistent with heightened violence during this period.

We conclude the frame extractions for the Israeli-Palestinian case are consistent with the historical record over the period of study.

6 Related Work

6.1 Events Data in Political Science

Projects using hand-collected events data represent some of the earliest efforts in the statistical study of international relations, dating back to the 1960s (Rummel, 1968; Azar and Sloan, 1975; McClelland, 1970). Beginning in the mid-1980s, political scientists began experimenting with automated rule-based extraction systems (Schrodt and Gerner, 1994). These efforts culminated in the open-source program, TABARI, which uses pattern matching from extensive hand-developed phrase dictionaries, combined with basic part of speech tagging (Schrodt, 2001); a rough analogue in the information extraction literature might be

the rule-based, finite-state FASTUS system for MUC IE (Hobbs *et al.*, 1997), though TABARI is restricted to single sentence analysis. Later proprietary work has apparently incorporated more extensive NLP (e.g., sentence parsing) though few details are available (King and Lowe, 2003). The most recent published work we know of, by Boschee *et al.* (2013), uses a proprietary parsing and coreference system (BBN SERIF, Ramshaw *et al.*, 2011), and directly compares to TABARI, finding significantly higher accuracy. The original TABARI system is still actively being developed, including just-released work on a new 200 million event dataset, GDELT (Schrodt and Leetaru, 2013).[14] All these systems crucially rely on hand-built pattern dictionaries.

It is extremely labor intensive to develop these dictionaries. Schrodt (2006) estimates 4,000 trained person-hours were required to create dictionaries of political actors in the Middle East, and the phrase dictionary took dramatically longer; the comments in TABARI's phrase dictionary indicate some of its 15,789 entries were created as early as 1991. Ideally, any new events data solution would incorporate the extensive work already completed by political scientists in this area while minimizing the need for further dictionary development. In this work we use the actor dictionaries, and hope to incorporate the verb patterns in future work.

6.2 Events in Natural Language Processing

Political event extraction from news has also received considerable attention within natural language processing in part due to government-funded challenges such as MUC-3 and MUC-4 (Lehnert, 1994), which focused on the extraction of terrorist events, as well as the more recent ACE program. The work in this paper is inspired by unsupervised approaches that seek to discover types of relations and events, instead of assuming them to be pre-specified; this includes research under various headings such as template/frame/event learning (Cheung *et al.*, 2013; Modi *et al.*, 2012; Chambers and Jurafsky, 2011; Li *et al.*, 2010; Bejan, 2008), script learning (Regneri *et al.*, 2010; Chambers and Jurafsky, 2009), relation learning (Yao *et al.*, 2011), open information extraction (Banko *et al.*, 2007; Carlson *et al.*, 2010), verb caseframe learning (Rooth *et al.*, 1999; Gildea, 2002; Grenager and Manning, 2006; Lang and Lapata, 2010; Ó Séaghdha, 2010; Titov and Klementiev, 2012), and a version of frame learning called "unsupervised semantic parsing" (Titov and Klementiev, 2011; Poon and Domingos, 2009). Unlike much of the previous literature, we do not learn latent roles/slots. Event extraction is also a large literature, including supervised systems targeting problems similar to MUC and political events (Piskorski and Atkinson, 2011; Piskorski *et al.*, 2011; Sanfilippo *et al.*, 2008).

One can also see this work as a relational extension of co-occurence-based methods such as Gerrish (2013; ch. 4), Diesner and Carley (2005), Chang *et al.* (2009), or Newman *et al.* (2006), which perform bag-of-words-style analysis of text fragments containing co-occurring entities. (Gerrish also analyzed the international relations domain, using supervised bag-of-words regression to assess the expressed valence between a pair of actors in a news paragraph, using the predictions as observations in a latent temporal model, and compared to MID.) We instead use parsing to get a much more focused and interpretable representation of the relationship between textually co-occurring entities; namely, that they are the source and target of an action event. This is more in line with work in relation extraction on biomedical scientific articles (Friedman *et al.*, 2001; Rzhetsky *et al.*, 2004) which uses parsing to extracting a network of how different entities, like drugs or proteins, interact.

7 Conclusion

Large-scale information extraction can dramatically enhance the study of political behavior. Here we present a novel unsupervised approach to an important data collection effort in the social sciences. We see international relations as a rich and practically useful domain for the development of text analysis methods that jointly infer events, relations, and sociopolitical context. There are numerous areas for future work, such as: using verb dictionaries as semi-supervised seeds or priors; interactive learning between political science researchers and unsupervised algorithms; building low-dimensional scaling, or hierarchical structure, into the model; and learning the actor lists to handle changing real-world situations and new domains. In particular, adding more supervision to the model will be crucial to improve semantic

[14] http://eventdata.psu.edu/data.dir/GDELT.html

quality and make it useful for researchers.

Acknowledgments

Thanks to Justin Betteridge for providing the parsed Gigaword corpus, Erin Baggott for help in developing the document filter, and the anonymous reviewers for helpful comments. This research was supported in part by NSF grant IIS-1211277, and was made possible through the use of computing resources made available by the Pittsburgh Supercomputing Center. Brandon Stewart gratefully acknowledges funding from an NSF Graduate Research Fellowship.

References

Azar, E. E. and Sloan, T. (1975). Dimensions of interactions. Technical report, University Center of International Studies, University of Pittsburgh, Pittsburgh.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open Information Extraction from the Web. *IJCAI*.

Bejan, C. A. (2008). Unsupervised discovery of event scenarios from texts. In *Proceedings of the 21st Florida Artificial Intelligence Research Society International Conference (FLAIRS), Coconut Grove, FL, USA*.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of ICML*.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics*, **1**(1), 17–35.

Boschee, E., Natarajan, P., and Weischedel, R. (2013). Automatic extraction of events from open source text for predictive forecasting. *Handbook of Computational Approaches to Counterterrorism*, page 51.

Brandt, P. T., Freeman, J. R., and Schrodt, P. A. (2011). Real time, time series forecasting of inter-and intra-state political conflict. *Conflict Management and Peace Science*, **28**(1), 41–64.

Brandt, P. T., Freeman, J. R., Lin, T.-m., and Schrodt, P. A. (2012). A Bayesian time series approach to the comparison of conflict dynamics. In *APSA 2012 Annual Meeting Paper*.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 1306–1313.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, **81**(3), 541–553.

Chambers, N. and Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of ACL-IJCNLP*. Association for Computational Linguistics.

Chambers, N. and Jurafsky, D. (2011). Template-based information extraction without the templates. In *Proceedings of ACL*.

Chang, J., Boyd-Graber, J., and Blei, D. M. (2009). Connections between the lines: augmenting social networks with text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178. ACM.

Cheung, J. C. K., Poon, H., and Vanderwende, L. (2013). Probabilistic frame induction. In *Proceedings of NAACL*. arXiv preprint arXiv:1302.4813.

de Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Stanford University.

Diesner, J. and Carley, K. M. (2005). Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. In *Causal mapping for information systems and technology research*, pages 81–108. Harrisburg, PA: Idea Group Publishing.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277—1287.

Eisenstein, J., Ahmed, A., and Xing, E. (2011). Sparse additive generative models of text. In *Proceedings of ICML*, pages 1041–1048.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**(suppl 1), S74–S82.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1).

Gerner, D. J., Schrodt, P. A., Yilmaz, O., and Abu-Jabr, R. (2002). The Creation of CAMEO (Conflict and Mediation Event Observations): An Event Data Framework for a Post Cold War World. *Annual Meeting of the American Political Science Association*.

Gerrish, S. M. (2013). *Applications of Latent Variable Models in Modeling Influence and Decision Making*. Ph.D. thesis, Princeton University.

Gerrish, S. M. and Blei, D. M. (2011). Predicting legislative roll calls from text. In *Proceedings of ICML*.

Ghosn, F., Palmer, G., and Bremer, S. A. (2004). The MID3 data set, 1993–2001: Procedures, coding rules, and description. *Conflict Management and Peace Science*, **21**(2), 133–154.

Gildea, D. (2002). Probabilistic models of verb-argument structure. In *Proceedings of COLING*.

Goldstein, J. S. (1992). A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution*, **36**, 369–385.

Goldstein, J. S., Pevehouse, J. C., Gerner, D. J., and Telhami, S. (2001). Reciprocity, triangularity, and cooperation in the middle east, 1979-97. *Journal of Conflict Resolution*, **45**(5), 594–620.

Grenager, T. and Manning, C. D. (2006). Unsupervised discovery of a statistical verb lexicon. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, page 18.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *PNAS*, **101**(suppl. 1), 5228–5235.

Harrison, J. and West, M. (1997). *Bayesian forecasting and dynamic models*. Springer Verlag, New York.

Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., and Tyson, M. (1997). FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. *Finite-State Language Processing*, page 383.

Hoff, P. D. (2003). Nonparametric modeling of hierarchically exchangeable data. *University of Washington Statistics Department, Technical Report*, **421**.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, **1**(1), 145–168.

Jones, D., Bremer, S., and Singer, J. (1996). Militarized interstate disputes, 1816–1992: Rationale, coding rules, and empirical patterns. *Conflict Management and Peace Science*, **15**(2), 163–213.

King, G. and Lowe, W. (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, **57**(3), 617–642.

Lang, J. and Lapata, M. (2010). Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 939–947. Association for Computational Linguistics.

Lehnert, W. G. (1994). Cognition, computers, and car bombs: How Yale prepared me for the 1990s. In *Beliefs, Reasoning, and Decision-Making. Psycho-Logic in Honor of Bob Abelson*, pages 143–173, Hillsdale, NJ, Hove, UK. Erlbaum. http://ciir.cs.umass.edu/pubfiles/cognition3.pdf.

Li, H., Li, X., Ji, H., and Marton, Y. (2010). Domain-independent novel event discovery and semi-automatic event annotation. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Sendai, Japan, November*.

Martin, A. D. and Quinn, K. M. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis*, **10**(2), 134–153.

McClelland, C. (1970). Some effects on theory from the international event analysis movement. Mimeo, University of Southern California.

Mimno, D., Wallach, H., and McCallum, A. (2008). Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS Workshop on Analyzing Graphs*.

Modi, A., Titov, I., and Klementiev, A. (2012). Unsupervised induction of frame-semantic representations. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 1–7. Association for Computational Linguistics.

Murphy, K. P. (2012). *Machine Learning: a Probabilistic Perspective*. MIT Press.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, pages 705–741.

Newman, D., Chemudugunta, C., and Smyth, P. (2006). Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 680–686. ACM.

Ó Séaghdha, D. (2010). Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444. Association for Computational Linguistics.

O'Brien, S. P. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, **12**(1), 87–104.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2009). English Gigaword Fourth Edition. *Linguistic Data Consortium*. LDC2009T13.

Piskorski, J. and Atkinson, M. (2011). Frontex real-time news event extraction framework. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 749–752. ACM.

Piskorski, J., Tanev, H., Atkinson, M., van der Goot, E., and Zavarella, V. (2011). Online news event extraction for global crisis surveillance. *Transactions on computational collective intelligence V*, pages 182–212.

Polson, N. G., Scott, J. G., and Windle, J. (2012). Bayesian inference for logistic models using Polya-Gamma latent variables. *arXiv preprint arXiv:1205.0310*.

Poon, H. and Domingos, P. (2009). Unsupervised semantic parsing. In *Proceedings of EMNLP*, pages 1–10. Association for Computational Linguistics.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, **54**(1), 209228.

Rajaraman, A. and Ullman, J. D. (2011). Mining of massive datasets. Cambridge University Press; http://infolab.stanford.edu/~ullman/mmds.html.

Ramshaw, L., Boschee, E., Freedman, M., MacBride, J., Weischedel, R., , and Zamanian, A. (2011). SERIF language processing effective trainable language understanding. *Handbook of Natural Language Processing and Machine Translation*, pages 636–644.

Regneri, M., Koller, A., and Pinkal, M. (2010). Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 979–988.

Rooth, M., Riezler, S., Prescher, D., Carroll, G., and Beil, F. (1999). Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, page 104111.

Rummel, R. (1968). The Dimensionality of Nations project.

Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboué, P. A., Weng, W., Wilbur, W. J., Hatzivassiloglou, V., and Friedman, C. (2004). GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, **37**(1), 43–53.

Sandhaus, E. (2008). The New York Times Annotated Corpus. *Linguistic Data Consortium*. LDC2008T19.

Sanfilippo, A., Franklin, L., Tratz, S., Danielson, G., Mileson, N., Riensche, R., and McGrath, L. (2008). Automating frame analysis. *Social computing, behavioral modeling, and prediction*, pages 239–248.

Schrodt, P. (2012). Precedents, progress, and prospects in political event data. *International Interactions*, **38**(4), 546–569.

Schrodt, P. and Leetaru, K. (2013). GDELT: Global data on events, location and tone, 1979-2012. In *International Studies Association Conference*.

Schrodt, P. A. (2001). Automated coding of international event data using sparse parsing techniques. *International Studies Association Conference*.

Schrodt, P. A. (2006). Twenty Years of the Kansas Event Data System Project. *Political Methodologist*.

Schrodt, P. A. and Gerner, D. J. (1994). Validity assessment of a machine-coded event data set for the Middle East, 1982-1992. *American Journal of Political Science*.

Schrodt, P. A. and Gerner, D. J. (2004). An event data analysis of third-party mediation in the middle east and balkans. *Journal of Conflict Resolution*, **48**(3), 310–330.

Shellman, S. M. (2004). Time series intervals and statistical inference: The effects of temporal aggregation on event data analysis. *Political Analysis*, **12**(1), 97–104.

Titov, I. and Klementiev, A. (2011). A Bayesian model for unsupervised semantic parsing. In *Proceedings of ACL*.

Titov, I. and Klementiev, A. (2012). A Bayesian approach to unsupervised semantic role induction. *Proceedings of EACL*.

Wallach, H., Mimno, D., and McCallum, A. (2009). Rethinking lda: Why priors matter. *Advances in Neural Information Processing Systems*, **22**, 1973–1981.

Yao, L., Haghighi, A., Riedel, S., and McCallum, A. (2011). Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1456–1466. Association for Computational Linguistics.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**(413), 79–86.