Learning to Extract International Relations from Political Context

Brendan O'Connor
Carnegie Mellon University

Brandon Stewart
Harvard University

Noah A. Smith
Carnegie Mellon University, nasmith@cs.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/lti
Part of the Computer Sciences Commons

Published In
Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 1094-1104.

This Conference Proceeding is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Language Technologies Institute by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.
Learning to Extract International Relations from Political Context

Brendan O’Connor  Brandon M. Stewart  Noah A. Smith
School of Computer Science  Department of Government  School of Computer Science
Carnegie Mellon University  Harvard University  Carnegie Mellon University
Pittsburgh, PA 15213, USA  Cambridge, MA 02139, USA  Pittsburgh, PA 15213, USA
brenocon@cs.cmu.edu bstewart@fas.harvard.edu 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-

1Available from the Penn State Event Data Project: http://eventdata.psu.edu/
2For arbitrary reasons this portion of the data is much smaller (we only parse the first five sentences of each arti-
ford CoreNLP system, 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 \(E\), \(t\) is a timestep (i.e., a 7-day period) derived from the article’s published date, and \(w_{\text{predpath}}\) 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 \((\text{PAK, IND, 678, accuse } \text{dobj } \langle \text{accuse} \rangle)\). (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, 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 prep.with and fight prep.alongside (“talk with R,” “fight alongside R”) but many interesting multiword constructions also result, such as reject det.allegation pos. (“rejected R’s allegation”) or verb chains as in offer xcomp help 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 \((s, r, w)\) tuple \((\text{USA, CUB, want xcomp seize dobj } \langle \text{want} \rangle)\) 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 xcomp 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, paraaxis, 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. 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 \(\ell_1\)-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

footnotes:


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 \((s, r, t)\).

![Diagram of the model](image)

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

- **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 \( w_{\text{predpath}}^i \sim \text{Mult}(\phi_{z(i)}) \).

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 \( \alpha_k \) and variability \( \sigma_k^2 \).
- For each \((s, r, t)\),
  - Draw \( \phi_{k,s,r,t} \sim N(\alpha_k, \sigma_k^2) \) 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 \( \alpha_k \) parameter can learn that some frames tend to be more likely than others. The variance parameters \( \sigma_k^2 \) 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-occurrences, 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 \( \beta_{k,s,r,t} \) 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 \((\alpha_k, \sigma_k^2)\) are same as the vanilla model. This model assumes a random walk process on \( \beta \), a variable which exists even for contexts that contain no events. Thus inferences about \( \eta \) 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 \( \beta \) are for all event tuples besides \( i \).

The language model sampler sequentially updates every \( z^{(i)} \) (and implicitly \( \phi \) via collapsing) in the manner of Griffiths and Steyvers (2004):

\[
p(z^{(i)} | \theta, w^{(i)}, b) \propto \theta_{s,t,z}(n_{w,z} + b/V)/(n_z + b),
\]

where counts \( n \) are for all event tuples besides \( i \).

For the context model, \( \alpha \) is conjugate resampled as a normal mean. The random walk variables \( \beta \) 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 \( \eta \) observation; the Kalman filter implementation is appropriately modified to handle this.

The \( \eta \) 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 \( \eta^* \) 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 \( \tau^2, \sigma_k \) 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. 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

\[
E[(g(w_i) - g(w_j)) \mid z_i = z_j, w_i \neq w_j] \tag{1}
\]

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:

\[
\frac{1}{N} \sum_{w, v \in M} \sum_{w \neq v} n_{w,k} n_{v,k} |g(w) - g(v)| \tag{2}
\]

where \( n \) refers to the averaged Gibbs samples’ counts of event tuples having frame \( k \) and a particular verb path,\(^6\) 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 \( \sim1000 \) 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 possessives. This approach can not match them.

\(^8\)Derivation in supplementary appendix.

\(^9\)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).
The verb-path logistic regression performs strongly at AUC 0.62; it outperforms all of

\text{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.

Using the R glmnet package (Friedman et al., 2010).

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

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.
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
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 concurrent 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 Lee-taru, 2013). 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-occurrence-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](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


