Automatically Extracting and Aggregating Events From Text

Andy Halterman*

22 October 2019

Abstract

Measuring what political actors do in the world at the core of empirical social science, but existing automated methods to extract actions and behavior from text are highly specific, innaccurate, or expensive to build. This paper introduces a method for automatically extracting political events from text that uses syntactic information provided by natural language processing tools and neural networks trained on a diverse set of hand-labeled text. The method treats event extraction as a "slot filling" task, identifying the words in text that report who is doing what to whom, where and when, as reported by whom? In contrast to previous methods, this method not require hand-constructed dictionaries or pre-specified ontologies. To aggregate extracted actions into useful clusters, I introduce a new short text clustering algorithm that uses word embeddings to provide prior information. I illustrate the method by extracting one million events reported in State Department annual human rights reports and find that the types of abuses and specificity of reporting have changed over time.

1 Introduction

Much of the quantitative data used by social scientists consists of descriptions of political events that are produced by hand from newspaper reports, encyclopedias, NGO reports, and government documents (Table 1). Our existing and growing set of automated text analysis methods, however, are built primarily for summarizing documents, not for extracting information from them. The methods that do exist for extracting "who did what to whom" from text in political science and computer science have often require enormous up-front effort to customize systems to new event types (e.g. Raytheon BBN Technologies 2015; Norris, Schrodt, and Beieler 2017), they extract events that are irrelevant to political science (Gildea and Jurafsky 2002; Carreras and Màrquez 2005; Palmer, Gildea, and Xue 2010), or they have little to no ability to inductively learn event types from text.

*Ph.D. Candidate, Department of Political Science, Massachusetts Institute of Technology. ahalt@mit.edu

The primary contribution of this paper is in introducing a new automated technique to produce event-type data from text that does not require dictionaries, can inductively learn event types from text, and can be applied to a wide range of documents. The method makes it possible for researchers to extract event information and then usefully aggregate similar events from text. The method consists two steps: a "slot filling" step that extracts the words from a sentence that correspond to the actors, actions, and other information around an event, and a second "aggregation" step that groups the extracted phrases into categories for analysis. Previous methods have required large up-front human labor to build dictionary-based event recognition models and required rigid, pre-specified ontologies of actors and events. My new techniques allow researchers to learn events inductively from text using a single model that generalizes across domains without the need for retraining.

Dataset	Citation	Citation Count	Text Sources
MIDS	Jones, Bremer, and Singer (1996)	2171	diplomatic sources, histories, newspapers
CIRI human rights	Cingranelli and Richards (2004)	55 (?)	State Dept. reports
GTD	LaFree and Dugan (2007)	455	newswire, newspaper, gov. documents
Archigos	Goemans, Gleditsch, and Chiozza (2009)	670	Encyclopedias, newspapers
ACLED	Raleigh et al. (2010)	845	news text and humanitarian reporting
coups	Powell and Thyne (2011)	317	NYT and other text sources
SCAD	Salehyan et al. (2012)	270	AP and AFP
SIPRI arms transfers			commercial publications, newspapers, gov. publications
UCDP intrastate conflict	Sundberg, Eck, and Kreutz (2012)	174	newspapers
SPEED "civil strife"	Nardulli, Althaus, and Hayes (2015)	26	NYT, BBC Monitoring, FBIS
regime type	Geddes, Wright, and Frantz (2014)	542	News reports, published literature

Table 1: Many standard datasets in comparative politics and international relations are derived from text sources. Producing and updating them is often a multi-year, multi-annotator undertaking.

The first technique I introduce is a method for recognizing the spans of text associated with each of the "slots" of a political event. I propose a new set of standard slots that generalize across event types, consisting of actors who do an action, the action itself, the

political entity receiving the action, and the means or instrument involved in the action, along with slots for the reported cause or reason for the event, any reporter or source attribution in the text, and the date and location of the event. To fill these slots in practice, I introduce a technique that combines a rule-based system that uses the grammatical information of the sentence to identify spans of text that potentially correspond to each slot, and machine learning models trained on labeled spans to determine the correct event slot for each span of words.

The second technique performs the aggregation step, learning which actions belong together using a new text clustering algorithm. Rather than relying on dictionaries or supervised models to categorize events, it instead learns event types inductively. Because extracted action spans are often quite short, traditional topic models do not perform well. Instead, I uses pretrained embeddings to provide prior knowledge on word similarity and an iterative model to cluster very short phrases into useable classes of events. I show that this model outperforms standard topic modeling approaches in a simulation setup, and produces qualitatively good results on real text.

Finally, I demonstrate the utility of these new techniques to answer substantive questions in political science by returning to the ongoing debate on whether respect for human rights has improved over time. I produce new, disaggregated data on the specific acts of human rights abuses reported by the State Department in their monitoring documents over time. I offer clear evidence that the contents of reporting are changing over time, and suggestive evidence that the threshold for inclusion are changing as well. This application demonstrates the importance of creating new, tailored data answer open questions in political science.

2 Producing events from text

Political *events* are structured representations of political behavior, consisting of information on actions, the involved actors, and information on the manner, place, and time of the actions. Which actors or actions are "political" or relevant will depend on the specific research question at hand. Many methods for extracting events from text exist in political science and computer science and are discussed at length below. To aid exposition of the method, it is useful to conceptually decompose event extraction into two separate steps. The first step is a "slot filling" task that involves identifying the short pieces of text that correspond to various attributes of the event, such as the actor performing an action, the event's location, or any reporting source mentioned in the text. For example, slot filling would involve recognizing "met" as the word describing an action in the sentence "Obama met with Merkel in Berlin", and "Obama" as the word describing an actor involved in the meeting.

The second step consists of aggregating similar entities together in a category for later analysis. While slot filling would recognize that "detain" and "arrest" are both descriptions of actions, the aggregation step handles resolving both of them to the same type of action. Aggregation can be done in a supervised way, with spans assigned to clusters using a

trained machine learning model or, more commonly, with hand-created dictionaries. It can also be done in an unsupervised way, with categories learned inductively from the collection of spans. How spans are resolved to categories will depend on the specific research question being asked.

2.1 An ontology of event slots

Before an algorithm can identify the event slots in text, the set of slots first needs to be defined. I propose an ontology of event slots that builds on the strengths of existing approaches in political science and linguistics to accurately record information from events in a way that is standardized across different kinds of political behavior.

A wide body of literature on slot filling and "semantic role labeling" exists in computer science and natural language processing, attempting to create systems that can faithfully reflect the tremendous variety of human language and human behavior. Early semantic role labeling approaches have highly variable "slots" or "frame elements" that differ by the recognized event type. FrameNet (Baker, Fillmore, and Lowe 1998), for instance, specifies around 1,000 linguistic "frames." Many of these slots are specific to the event type: a "cook food" event, for instance, might have a slot for "source of heat". Many event types, however, have slots that are roughly comparable: a "crime" event's "victim" slot is roughly comparable to a "hire" event's "worker". Slots can only be filled once the type of event has been recognized, making automated approaches to slot filling difficult. For political scientists, some of these frames involve potentially political actions such as a "revenge" frame, specifying the injured party, the victim, and the manner of revenge. Other frames are less interesting to political scientists: a "clothing" frame includes roles for garment, material, color descriptors, and wearer.

This approach suffers from several drawbacks for applied information extraction work. First, these themes must be laboriously constructed by expert linguists, and their great level of specificity is aimed more at linguistic correctness than at practical usefulness (for example, great care is taken to distinguish bank deposits from alluvial silt deposits, or a "killer" role in murder from the "perpetrator" role in a kidnapping). Second, as with all hand constructed dictionary methods, it faces problems of low recall (Pavlick et al. 2015). Finally, the specificity of the slots makes the system difficult to train. A "victim" of a crime and a "beneficiary" of a gift both receive an action in some sense but FrameNet treats them as completely different entities.

Building on theoretical insights by Dowty (1991) on "proto-agents" and "proto-patients", Palmer, Gildea, and Kingsbury (2005) developed a much more general approach to the task, replacing specific frame elements with more general, numbered arguments. These numbered arguments often correspond to the "agents" committing an action, "instruments" used in committing the act, and "patients" recieving the action, but the meaning of each numbered argument varies by the specific verb, making it difficult to use in an applied political science setting.

Existing event data methods in political science assume a small number of slots that all events are expected to have: a "source" actor, an action, and a "target" actor (Gerner et al.

2002).[1] These slots are unsurpringly better suited to studying political questions, but still face shortcomings. Not every political event involves a recipient or target of the action, though the ontology requires that they do.

I propose an ontology of slots that consists of eight possible slots comprising an event.

1. An "actor" slot that the contains the actor doing the action. Grammatically, this slot will usually consist of subject nouns. In natural language processing, this slot is usually referred to as the "sender" or "agent".
2. An "action" slot that contains a description of the action that took place. Grammatically, this slot will contain at least one verb, but they also contain adverbs, adjectives, and other modifiers of the verb.
3. A "recipient" slot that contains information about the actor receiving the action. Grammatically, this lot will involve direct objects object of prepositions or indirect objects. In natural language processing this is referred to as the "receiver" or "patient" and in some earlier political science approaches (e.g. Gerner et al. (2002)), the "target".
4. An "instrument" or "means" slot, comprising the objects used by the actor in performing the action. For instance, the italicized objects in the following sentences are instruments or means: deliver *aid*, fire *mortars*, disperse using *tear gas*. Grammatically, these are reported in direct objects, prepositional phrases, and indirect objects. These grammatical roles are the same as where the action's recipient is alo reported,
5. A "reason/cause" slot for the contextual information that is often reported alongside events in political text. For instance, the italicized span in "arrested two people *for participating in last week's protests*" does not provide information about the event itself, but rather context for the event.
6. A date slot, with information on when the events took place.
7. A location slot, with information on where the events took place.
8. A "reporter" slot, with information on what source reported that the occurrence of the event.

At a minimum, an event must have an action and an actor or recipient, but other slots are optional and sentences reporting all eight pieces of information will be uncommon. Conceptualizing events in this way has several advantages over existing approach in political science. First, it decomposes the previous "event" slot into more granular "action" and "instrument" slots. Actions and instruments are grammatically quite distinct, and splitting them up will help automated systems to fill these slots from real sentences. Providing both an "instrument" and "recipient" slot also clarifies the challenge of distinguishing (in)direct objects that recieve actions and those that are involved in the commission of actions. This "direct object" problem is discussed at length below. Finally, "reason/cause" and "reporter" slots are useful for separating out parts of the sentence that provide important contextual information for the event, but should not themselves be coded as separate events.

[1] Many systems conceptually include a location slot, but techniques for properly filling location slots are only just emerging. See Halterman (2019).

2.2 Existing techniques for slot filling

Given a definition of event slots, the next task becomes creating a system to fill these slots from real sentences. I review the existing approaches to slot filling in computer science and political science and argue that the for political events, the major outstanding obstacle is the "direct object" problem, of determining when objects of verbs are "instruments" of the action, and when they are "recipients" of the action.

One existing approach from natural language processing to the span labeling task is PredPatt (Rudinger and Van Durme 2014; White et al. 2016), which uses deterministic rules on a universal dependency parse to label the arguments of an event. A rule-based system eliminates the need for training data. PredPatt will not work for the political event extraction task task without modification, however. First, important information about the role of actors is lost in prepositional phrases. As they put it,

> " 'Mary stuffed envelopes with coupons' and 'Mary stuffed envelopes with John' have identical dependency structures, yet 'coupons' and 'John' are (hopefully for John) taking on different semantic roles" (Rudinger and Van Durme 2014, 57).

Distinguishing between political actors and other objects is crucial for political science applications. Recipients and instruments cannot be distinguished on the basis of grammar alone. Instead, identifying which nouns are recipients and which nouns are instruments requires substantive knowledge to distinguish them. This requirement for substantive knowledge explains why the computer science literature has not yet produced a useful political event extraction system. PredPatt also cannot overcome the PropBank problem of not producing labels (e.g. agent, patient) instead of PropBank's more generic numbered arguments.

In political science, the dominant approach to event coding relies on dictionary methods. In early systmes, spans of text matching a list of actions are assigned to action slots and spans matching entries in an actor dictionary were coded as actors (e.g. Schrodt, Davis, and Weddle 1994; Schrodt 2009; Boschee et al. 2015). Later systems (e.g. Norris, Schrodt, and Beieler 2017) use grammatical information about the sentence in conjunction with dictionary information to perform the slot filling task. Dictionary-based methods require enormous up-front investment, have very low recall between 5% and 35% (Makarov 2018; Althaus, Peyton, and Shalmon 2018), and are difficult to extend to new event types. Moreover, systems that depend on dictionaries cannot be used to learn new event types inductively from text. An approach at the intersection of the two has been to directly classify sentences into one of four broad categories of events (cooperative and conflictual, verbal and material), but without an attempt to extract spans (Beieler 2016). This approach is only suitable though for extremely coarse event types.

A promising hybrid approach comes from O'Connor, Stewart, and Smith (2013), which uses dictionaries to identify actors (countries, in their case) and grammatical information from the dependency tree to fill the "action" slot linking the two actors. A modified topic model that accounts for temporal dependency in dyadic relationships learns events inductively. This approach still depends on pre-built dictionaries, however.

Similarly, Van Atteveldt et al. (2017) use handwritten rules and the produced dependency parse of the sentence to segment the sentence into actor, predicate, and "source." This approach is very similar to that of PredPatt (Rudinger and Van Durme 2014; White et al. 2016). While innovative in its use of dependency parses to model actions with sentences, this model has a major limitation. It combines actions and the recipients of actions together into a single predicate span. Doing so makes it impossible to distinguish which words are political entities receiving actions. They use as an example sentence, "Hospital officials in Gaza said that 390 people were killed by Israeli fighter planes." Their method returns "390 people [were] killed" as a single predicate span, rather than separating out "killed" as an action and "390 people" as a target (or, in my terminology, recipient) of that action.

2.3 Slot filling algorithm

I introduce a slot filling algorithm that uses information from the dependency parse and a neural net classifier to distinguish between instruments of action and recipients of actions. Any approach to filling the slots I specify must use both syntax (the grammar of the sentence) and semantics (the meanings of words). Figure 1 illustrates why.

A purely syntactic representation of a sentence cannot distinguish between, for instance, a direct object being an instrument of an action ("missiles") and a direct object being an actor receiving the action ("Tillerson"). In contrast, semantic analysis of words provides information about whether words are likely to describe people, actions, weapons, locations, and so on, but cannot link these words together into the meaningful relations encoded in text. My model uses both syntactic and semantic information to fill an event's slots.

My model proceeds in three steps (Figure 2.3 presents an overview of the steps). First, it performs a grammatical dependency parse of sentence. Next, it uses hand-specified rules and the dependency parse to segment the sentence into rough spans. Finally it uses a machine learning model to determine whether objects are instruments/means or recipients of the action and separate models to locate reporter and reason spans.

First, it uses the automatically-recognized dependency structure of a sentence (Honnibal and Montani 2017). Dependency parses encode the grammatical relationships between words in a sentence in a directed tree. For example, a verb ("fired") could be connected to its subject noun ("Trump") and its direct object ("Tillerson"). I generate a set of deterministic rules on this tree to produce candidate spans for the actor, action, recipient, instrument, location, date, and reporter slots for each event.[2]

Algorithm 1

def *children*:

[2]*Note: the model to detect "reason/cause" spans is still under development. More labeled data is required to produce a model with good accuracy.

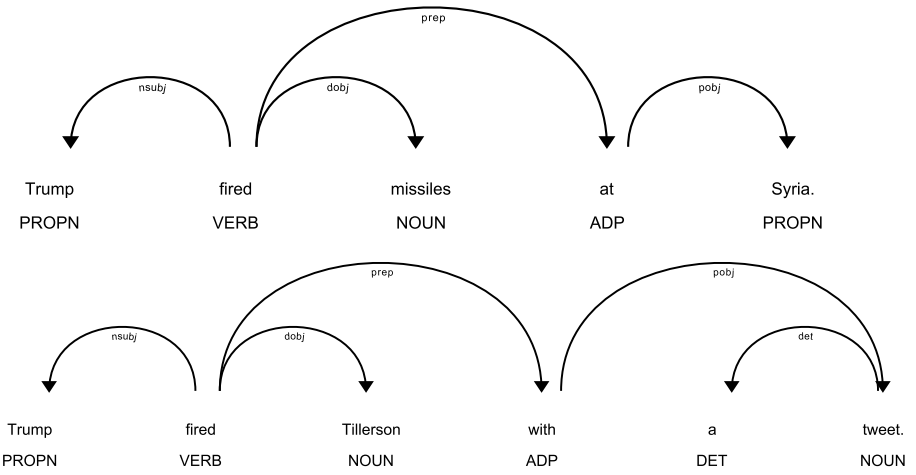

Figure 1: A dependency parse representation of two sentences. Dependency parses can be read as directed trees, beginning with a "root" verb (here, "fired") and having labeled paths (e.g. "nsubj") from parent to child nodes (e.g. "Trump", "missiles"). The all-capitals labels below each word are the words' part-of-speech tags. (Note the error made by the automated parsing system in labeling "Tillerson" as a noun instead of a proper noun.) Part-of-speech tags are invariant to the grammar of the sentence: "missiles" is a noun, but across sentences it could play the role of subject noun, direct object, dative object, or object of a preposition. The sentences are nearly identical in their grammatical structure, but the grammatical parts of the sentences correspond to different slots, illustrating the need for semantic information about the words as well. In the first example, the direct object in the sentence plays the role of an "instrument" of the action, while in the second sentence, the direct object plays the role of the recipient of the action.

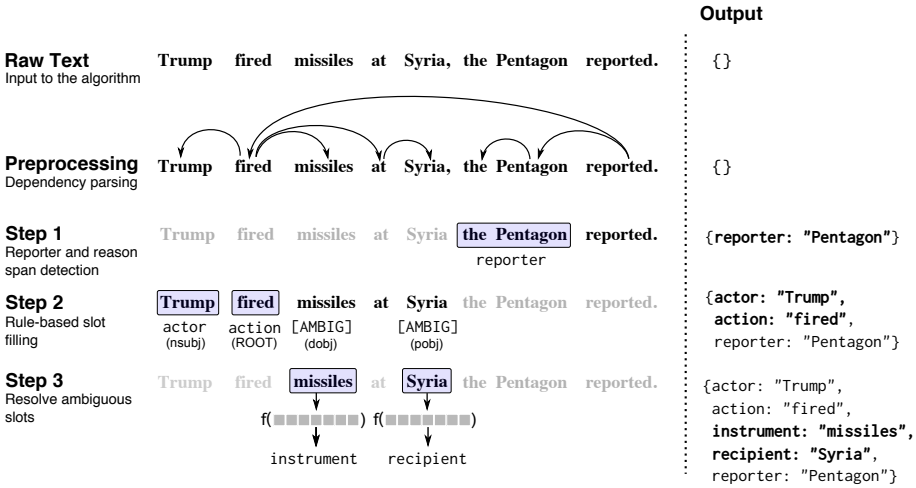

Figure 2: Algorithm for extracting politically relevant event spans from text. The output from each step is reported in the right column. Sentences are preprocessed with a dependency parser, then "reporter" and "reason" spans are detected with a word-level convolutional neural network (Step 1). "Actors" and "actions" are detected using rules applied to the dependency parse of the remaining text, along with spans that could either be instruments or recipients of the action (Step 2). A neural net classifier is applied to these ambiguous spans' embeddings to determine their role (Step 3). Dates and locations are detected using NER (not shown in this example).

a word's children are the nodes immediately "below" it on the dependency tree. (E.g., in Figure 1, "tweet" is a child of "with".)

def *ancestor*:
All words upstream of the word in the dependency parse in the path to the root verb.

def *predicate subtree*:
traverse all branches of tree, with the exception of words that are marks (`mark`) or adverbial clauses (`advcl`), or words that are labeled as reporters, or words whose ancestor subject noun (`nsubj`) is different from the subject noun of interest.

input: a subject noun (`nsubj` relation)
oututs: ⟨source actor, action, recipient ⟩

1. define the actor as all the subtree[3] of the subject noun.
2. define candidate recipient spans as the subtrees of all direct objects, objects of prepositions, dative objects, and in the passive case
3. actual recipients are candidate recipient spans where the recipient labeler function returned a high predicted probability of them being recipients, rather than instruments/means.
4. the predicate is the action itself combined with the instrument/means spans (grammatically, a pruned subtree of the subject noun's parent verb, with detected recipi-

[3]The subtree is recursively all children of that word and the children of its children, etc.

ents removed).

Note: This model is for sentences in the active voice. A slightly modified version handles passive sentences.

The syntactic information provided by the dependency parse cannot on its own fully resolve each span, however. The (grammatically identical) sentence "Trump fired missiles" uses "missiles" in a different semantic role from "Tillerson", despite their identical grammar. "Tillerson" is a recipient of the action, while "missiles" is an "instrument" of the action and belongs in the action slot alongside "fired". To resolve these issues, I train a convolutional neural network (CNN) classifier on a new set of labeled data to classify noun phrases either as recipients or instruments of actions. The model operates on the words' pretrained embeddings, meaning that it can easily classify new words it did not see during training, and the CNN can account for word order over the short spans without the computational cost of a recurrent neural network. Specifically, I create a dataset of "candidate" actors, consisting of spans of text that syntactically may be actors, but semantically may be instruments of actions. I manually labeled 2,000 of these spans, drawn from newspaper, newswire, Wikipedia, and government reports to give good cross-domain performance. The convolutional neural network that I fit uses pretrained embeddings as inputs. Each convolution in the network is applied to a window of three words at once, meaning that the model can learn trigram information. The model stacks several convolutional layers to learn wider relationships between words. The model achieves 81% accuracy and 83% F1. In production, phrases that are recognized as recipients are then removed from the predicate and placed in the recipient slot. See Figure .

I also train a "reporter" model that recognizes phrases with a sentence that provide a source attribution for the event, such as "…, Amnesty International reported." These phrases are then removed from the sentence, preventing them from being coded as extra events, and allowing them to be added as metadata to the extracted events. The reporter recognition task is similar to named entity recognition tasks, so I use a multilayer convolutional neural network that uses pretrained word embeddings and that performs well on named entity recognition tasks (Honnibal and Montani 2017). The reporter model achieves 78% accuracy.

Van Atteveldt et al. (2017) develop a model for recognizing "sources" [reporters] that uses a set of hand-specified rules.[4] The advantage of using a machine learning model over a rule-based system is that machine learning models often higher recall on actual production text. Information on dates and locations is easily extracted using off-the-shelf named entity recognition. A more sophisticated approach to linking actions and the most specific locations where they are reported to occur is described in Halterman (2019) and could be easily incorporated into the algorithm.

[*Note to the Kim Research Group: I'm in middle of collecting a large set of hand-labeled*

[4] I use the term "reporter" instead of "source" to avoid confusion with the terminology used in the standard political science ontology, CAMEO, where "source" is often used where I use the term "actor". (Gerner et al. 2002)

sentences in order to measure overall accuracy on sentences. Collecting annotated sentences is quite slow, however, so I don't yet have enough to provide overall accuracy numbers.]

2.4 Aggregating actions

Once spans are extracted from text, they still require more work before they can be analyzed. Specifically, spans of text that describe the same action need to grouped together.[5] In some cases, researchers will only be interested in a small, pre-identified set of behaviors, in which case they can use supervised learning techniques to identify the subset of actions they would like to label. In other cases, however, researchers may want to inductively learn clusters of actions from the text, either as exploratory research or to test hypotheses about the behaviors that actors engage in.[6]

Unsupervised text analysis, specifically latent Dirichlet allocation (Blei, Ng, and Jordan 2003) and its variants (Blei and Lafferty 2007; Roberts et al. 2013) is a mainstay of empirical political science. LDA, however, is unsuited for an application like this one because this application requires clustering extremely short spans of text, including those as short as a single word. I introduce a method to perform very short document clustering that draws on word embeddings to provide prior information about word meaning and a latent variable interpretation of a document embedding technique.

Latent Dirichlet allocation can be interpreted as a probabilistic factoring of a matrix of counts of words in each document into a matrix of topic proportions per document ($\boldsymbol{\theta}$) and a distribution over words for each topic ($\boldsymbol{\beta}$) (Buntine 2002; Hoffman, Bach, and Blei 2010). Specifically, by marginalizing out the topic indicator z_i, the probability of a word is given by

$$p(w_{di}|\theta_d, \boldsymbol{\beta}) = \sum_k \theta_d \boldsymbol{\beta}. \tag{1}$$

In situations such as this one, where "documents" are in fact short spans that can be as short as a single word, the matrix of word–document counts will be extremely sparse. In situations where documents have only one word, their representations in the document–word count matrix will be completely orthogonal, making it difficult to learn a low-rank approximation using LDA. More heuristically, LDA uses the co-occurance of related words in long documents to learn high-quality topics. In very short spans, synonyms are very unlikely to co-occur: the use of a word almost precludes the use of a close synonym in a span of 1–10 words.

[5] I focus here on actions, rather than actors or recievers because the appropriate groupings of political actors are generally easy to specific a priori than the best groupings of actions, and because in practice, grouping actors is fairly straightforward using dictionary methods.

[6] A hybrid approach combines unsupervised clustering with the small number of human analyst decisions. For example, Ritter et al. (2015) propose a weakly supervised model for recognizing events, that require analysts to only specify small number of positive documents of interest. A semi-supervised approach to learning event categories is promising but is left for future work.

An alterative technique is to use pretrained word embeddings to provide prior information on the similarity of words. Word embeddings are a technique for learning dense, low-dimensional vector representations of words based on their context in large corpora or text. Specifically, word2vec (Mikolov et al. 2013), one of the common word embedding algorithms, is a factoring of the (shifted) pointwise mutual information (PPMI) matrix of words and contexts (Levy and Goldberg 2014). Specificially, word2vec's skip-gram negative sampling technique is factorizing an implicit matrix $\boldsymbol{M}$, which Levy and Goldberg (2014) show is related to PPMI:

$$\boldsymbol{M}_{ij}^{SGNS} = w_i \cdot c_j = \log\left(\frac{\#(w,c)\cdot|D|}{\#(w)\cdot\#(c)}\right) - \log k, \tag{2}$$

where $\#(w)$ indicates the count of word w, c represents a context window of surrounding words, and k is a fixed scalar. Written this way, the relationship becomes apparant between LDA's objectives and word2vec (and similar embedding algorithms). While LDA is factoring the counts of words in documents, word2vec is factoring the counts of words in contexts, accounting for the marginal probabilities of words and contexts, and logged.

word2vec is already commonly used as a replacement for LDA is political science applications (see, e.g. Kornilova, Argyle, and Eidelman 2018; Spirling and Rodriguez 2019; Rheault and Cochrane 2019; Lauretig 2019 or recent conference programs from Text as Data or PolMeth.) Words that are used in similar contexts in a corpus will have similar vector representations, allowing researchers to learn how words are being used, for example, by different parties or in a way that changes over time. The technique I propose here does not learn new embeddings. Instead, it uses embeddings that have been pretrained on a large corpus of text to provide prior information about the (similar) meanings of words: very roughly, pretrained embeddings provide an approximation of $\boldsymbol{\beta}$, the word–topic distribution. The model, like a human reader, thus comes to a corpus with a sense that "arrest" is more similar to "detain" than to "France."[7] The next step is to learn an equivalent to θ_d, the topic present in each document.

2.5 Representing documents

Word embeddings provide a representation of *words*, but do not provide an obvious technique for representing *documents*. A standard technique is to produce a document vector by averaging the embeddings of all words in a document, and sometimes by appending the elementwise maximum to that vector (Goldberg 2017). Simple averaging treats all words as equally informative and causes documents to appear more alike as they increase in length. A more sophisticated approach is to learn a document embedding alongside word embeddings ("doc2vec") (Le and Mikolov 2014), but this approach requires a sepa-

[7] In this sense, word embeddings are being used here as a form of transfer learning, in which a representation learned on one corpus or task is applied to another to improve performance. Recent improvements in transfer learning for natural language processing are producing rapid improvements in the field. See Howard and Ruder (2018); Peters et al. (2018); Devlin et al. (2018). Ruder (2018) provides a non-technical overview.

rate step to learn paragraph embeddings for new documents at test time and the resulting document vectors are not easily interpretable.

Instead, I adopt a sentence embedding model proposed by Arora, Liang, and Ma (2017). Their model is theoretically motivated, simple to implement, and achieves very good performance on sentence classification tasks, beating even sophisticated supervised sentence classification models. In previous work Arora et al. (2016) show what embeddings models such as word2vec and GloVe (Pennington, Socher, and Manning 2014) can be interpreted in a generative model: words in a span of text are emitted with a probability given by the word's distance from a latent "discourse vector", which conducts a random walk through embedding space as words are emitted. They offer a simple interpretation of this discourse vector: "Its coordinates represent what is being talked about." (Arora et al. 2016, 387).

Arora, Liang, and Ma (2017) propose a sentence embedding technique the approximates the maximum likelihood estimates of this discourse vector. Specifically, each sentence embedding is initially represented using a smoothed, weighted elementwise mean of its constituent words' embeddings:

$$\tilde{v}_i = \frac{1}{|d|} \sum_{w \in d} \frac{a}{a + p(w)} v_w, \tag{3}$$

where v_w is the pretrained word embedding of word w, $p(w)$ be the empirical frequency of word w in a large corpus, and $a = 0.0001$ is a smoothing hyperparameter. This weighting approximates the standard tf-idf weighting scheme in traditional text analysis and information retrieval. Next, the sentence vectors then have a "common component" removed, in which the first singular vector of all the vectors in the corpus are removed from each: $v_i = \tilde{v}_i - uu^T \tilde{v}_i$, where u is the first singular vector of the matrix X of all $\tilde{v}_i$. This sentence embedding techniques has two nice properties: it generates a fixed size embedding for a short document, in a way that is theoretically motivated and preserves information in the document as well as more sophisticated task-specific representations.

I modify the original SIF sentence embedding model to improve its applicability to this specific domain. Specifically, I vary word weights by their part-of-speech, in addition to by their word frequency. Because I focus on the specific domain of actions in the clustering algorithm, and because verbs are generally the most important component, I give their embeddings full weight regardless of their empirical frequency. Auxiliary words, digits, and proper nouns, in contrast, are reduced in importance to have the embedding over-weight rare but uninteresting words and to avoid overfitting downstream.

2.6 Clustering document spans

Clustering is a process for learning a useful, low-dimensional representation of data by placing "similar" units closer than "different" units in some space. The decisions involved

in designing clustering involve picking definitions of distance and space that produce clustering results that are useful to researchers in some way, similar to the decisions involved in designing clustering algorithms for international relations (Zhukov and Stewart 2013).

I opt for k-means clustering to learn clusters of similar spans. k-means clustering hard assigns points to the closest cluster center. These k cluster centroids are iteratively moved to reduce the within-cluster variance, which implicitly minimizes the squared Euclidean distance between points and their cluster centroids. This clustering algorithm is appropriate here for two reasons. First, because the spans being clustered are so short, it is reasonable to treat them as belonging to a single cluster, rather than the mixture of clusters that a model like LDA or a Gaussian mixture model assumes. Second, a Gaussian mixture model would be prefereable if we believed that the points in embedding space were actually distributed as a mixture of Gaussians. Arora et al. (2016), however, find that latent word vectors are approximately uniformly distributed in vector space. The embedding dimensions are also approximately the same scale, making it less important to have a GMM's elliptical cluster shapes. This makes it appropriate to use k-means, which comes without distributional assumptions.

2.6.1 Evaluating cluster quality

To evaluate the topic models, I first assess their accuracy on a set of synthetic spans created by a known data generating process. While evaluation of methods against simulated data is a standard technique in most quantitative methodology, it is rarely applied in in topic modeling (though see Boyd-Graber and Blei 2009). I manually specify verbs, direct objects, and adjectives corresponding to eight political topics (see Appendix A). Each document is generated by sampling a topic indicator, then a single corresponding verb from that topic's set of verbs, and 0-4 other words for that topic. For example, one topic in the simulation contains "meeting" words, while another includes aid-related words including the verbs "deliver" and "provide" and other predicate words "humanitarian", "water", "food", and "aid". Generated documents also include draws from a set of "junk" terms consisting of conjunctions and prepositions that are shared across all topics.

I compare the performance of k-means clustering on SIF embeddings with a standard LDA model across two corpus sizes, performing 100 simulations for each condition (Figure 3). I match learned clusters with the DGP clusters by assigning each learned cluster to the true cluster that maximizes estimated accuracy, allowing for multiple clusters to be assigned to the same true cluster. An optimal "Hungarian" algorithm for assigning 1-to-1 matches produces significantly worse performance for all methods (Kuhn 1955). Even when LDA is run with a hyperparameter that encourages it to find a single topic per span, the SIF embedding/k-means approach is better able to recover topics from the data generating process.

Techniques for conducting structured, subjective evaluations of topic model quality are the subject of ongoing research (Chang et al. 2009; Demszky et al. 2019; Spirling and Rodriguez 2019), but as an unsupervised technique, the quality of topic models is best evaluated through their usefulness on on a substantive question of interest.

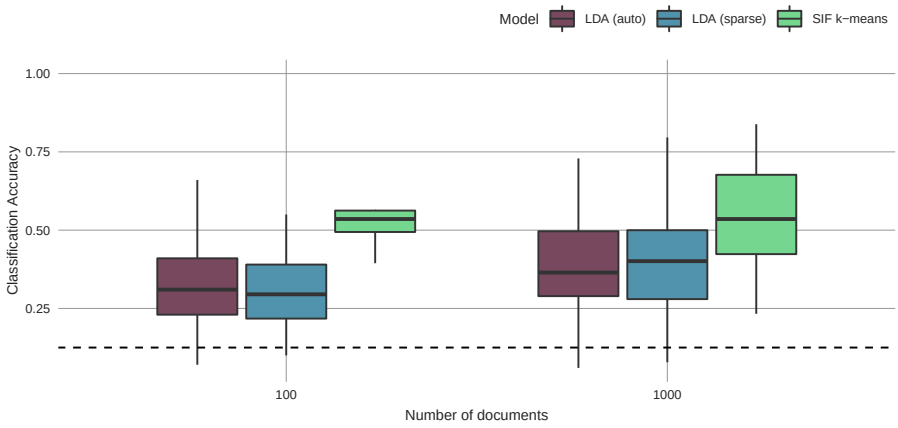

Figure 3: Comparing LDA and the short span clustering model's ability to recover known topic labels on synthetic data across 100 simulations. The new model using the SIF embedding technique with k-means outperforms both versions of LDA. The auto LDA model uses a self-tuning hyperparameter for the expected number of topics per document. The sparse model is set to find one topic per document, which matches the data generating process. The dashed line indicates expected random performance.

3 Changing respect and changing reporting for global human rights

An ongoing debate in international relations and comparative politics concerns whether respect for human rights has changed over time. Many observers expect, on anecdotal or qualitative grounds, that the global human rights situation has improved since the 1970s. In contrast, the major datasets of respect for human rights, including the CIRI Human Rights Dataset (Cingranelli and Richards 2004) and the Political Terror Scale (Wood and Gibney 2010) dataset show a fairly constant level of human rights violations over the past four decades.

Fariss (2014; 2018) argues that this counterinuitive finding is the product of changes in how human rights violations are reported. As NGOs gain greater access and human rights observers have better information, a greater proportion of human rights violations will be recorded than in the past. If the probability of detecting human rights violations is increasing faster than the overall rate of actual violations is decreasing, we will observe an apparent increase in human rights violations. Similarly, as the human rights record improves in different countries, human rights activists are likely to change the focus of their activism to other, less egregious violations.

Fariss (2014) models this change using a dynamic IRT model, using incidents of genocide as a perfectly observed anchoring observation to estimate the probability of incident reporting. He distinctions between what he calls "event" and "standards"-based report-

ing, with "events" like genocide being more accurately measured than the "standards" that the State Department and Amnesty International measure because the definition of events changes less than the definition of standards and because data on events is updated retrospectively as better information becomes available.

Fariss' paper makes an important contribution to the debate in positing the existence and mechanisms of changing reporting standards. The model that it uses, however, rests on several major assumptions, the greatest of which is that all state repression, from arbitrary arrest to genocide, exists along a single latent space, meaning that values can be compared across them. Instead, we might believe that genocide is simply different from other violations of human rights, violating the assumption of the unidimensional latent variable. D. Cingranelli and Filippov (2018a) and D. Cingranelli and Filippov (2018b) dispute this finding, largely on objections to Fariss's IRT model.

Rather than relying on the same limited set of country-year ratings to measure human rights respect and the changing standard of human rights violations, I instead generate new data on respect for human rights by returning to the original State department text used to create the country year ratings. Other researchers (Greene, Park, and Colaresi 2019) have also begun looking directly at the text, but in ways that do not preserve the relationships between actors and actions in the text. This allows us to produce fine-grained data on actions and the ability to link those actions to government actors.

I applied both steps of my new method to the State Department's annual country human rights reports from 1977 until 1999, when the format of the documents changed. From this text, the event extraction model produced 1.02 million events. Because this debate is over government respect for human rights, I then subset the events to only those in which the extracted actor span contained terms in a list of terms that I specified. This list included all country names and demonyms, along with terms describing government officials, such as "soldier", "authorities", "police", or "government". Approximately one quarter of the total events, 243,449, had actor spans that included these words. The date I produce is thus a compromise between between what Fariss calls standards-based reporting and event reporting. Rather than producing a single country or score as in the standard approach I produce a set of disaggregated events. Unlike codings of genocide, however, these machine extracted events are not updated retroactively as better data becomes available.

I then fit the SIF/k-means clustering algorithm to these extracted spans. I fit the model using $k = 60$ clusters, after experimenting with several values of k. Many of the topics are quite specific and contain only a small number of spans. A small number of topics together contain the majority of spans, which may be better modeled by an even larger number of topics.[8]

[8]It was only after applying the method to this corpus of text that I modified the SIF embedding method to decrease the weight on proper nouns and digits, since several of the clusters seemed to have only the presence of names and years as common traits.

3.1 Empirical Results

As Fariss observes, the total amount of reporting, measured by the number of words, has increased over time. Figure 4 shows that the number of reported events is gone up as well, from approximately 2,500 per year to around 25,000 per year. On its own, this figure offers some suggestive evidence that the standard of reporting has changed. We may believe that human rights practices are stagnant or perhaps even slightly worsening, but we do not believe that human rights violations have become an order of magnitude more common.

More interestingly, the data suggests that the nature of reporting is changing, becoming more focused on specific events over time. After normalizing by the length of documents, the number of events reported as increased. From 1979 to 1999, the number of events has gone from 10 to 16 events per 1,000 words. This indicates at least higher specificity in the content of the reports. Interestingly, however, the proportion of events with government actors remains steady between 22% to 25% over the period.

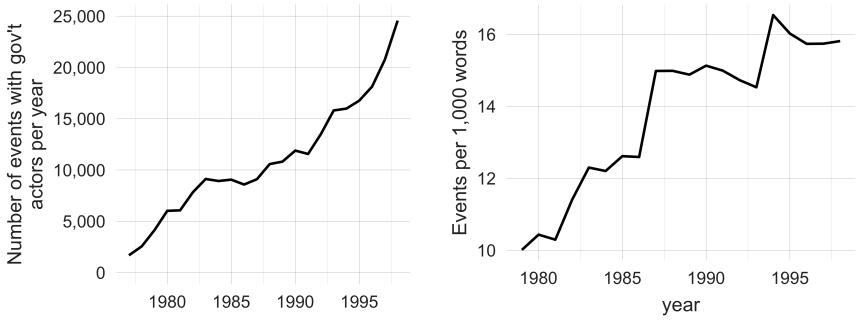

Figure 4: The number of extracted events from the State Department annual human rights reports with government actors (left) and the number of events per 1000 words per year. Events are extracted using the method introduced in the paper. Events are limited to those with an extacted actor that matches a government keyword (e.g. "police" or the name of a country). The results indicate both an overall increase in reporting and an increasing density and specificity of reporting.

Viewing aggregate trends of the number of events only provides marginally more information than the raw count of words, however. The real strength of the new method comes in learning different events types and decomposing the set of events into more specific event types. When we decompose the total line into a proportion of each event type we see variation in which event type are occupying a larger proportion of total events (Figure 5).

If we then focus only on the topics whose shares are decreasing (Table 2), we again see suggestive evidence for changing standard of reporting. Inspecting documents sampled from these clusters indicates that three of these clusters described not human rights abuses but other events or information about these countries. Topic 16 seems to focus on descriptions of the countries economic system. Topic 47 includes in large part positive reports of

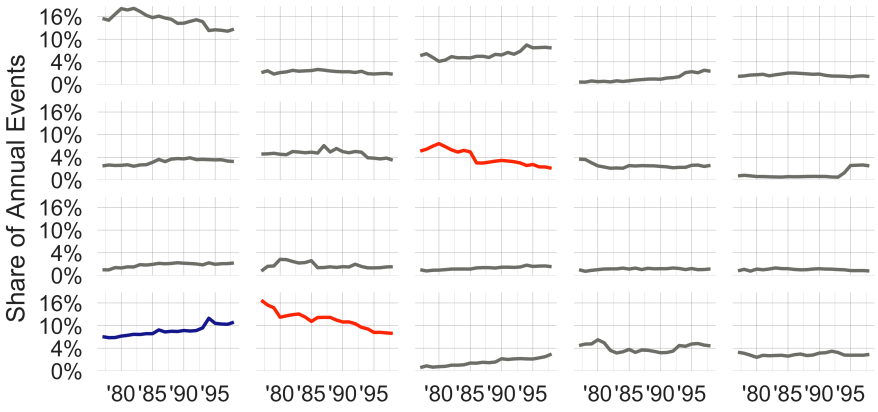

Figure 5: The share of extracted government events per learned cluster from State Department human rights reporting. Events are extracted using the new slot filling method and are clustered using the SIF embedding/k-means method introduced here. Even as the total number of events is increasing, the share coming from each event cluster is changing.

the countries respect for rights. Topic 23 is included for contrast and represents a more typical event type, involving arrests and detentions. The share of this event type has increased over time. Decreases in positive or "status update" events offers some evidence that the standard of reporting has changed.

Topic 16	"made internal economic reform and market stabilization"
	"pursued a successful export-oriented agricultural growth strategy"
	"own and run banking and insurance, air and rail transport, public utilities, and key industries"
Topic 47	'is a constitutional monarchy'
	'have lively and free, -, multiparty political systems'
	'is a multiparty democracy with mandatory universal suffrage'
	'is a representative democracy'
Topic 23	'provides for detention for an indefinite period without trial in national security cases'
	'not hold without a hearing before a magistrate'
	'when apprehend the accused during commission of a crime'
	'arrest persons without', 'picked up and held on suspicion of robbery.'
	'remained in Zomba Central Prison'

Table 2: Table to test captions and labels

Examining the human rights reporting corpus demonstrates some of the advantages of this method. First, we can directly measure what actors are *doing*, as opposed to just the words used in the documents, as previous methods are limited to (Greene, Park, and Colaresi 2019). This allows us to more precisely measure the outcome of interest: human rights abuses. Second, because the new technique does not rely on dictionaries or prespecified categories of behavior, we can inductively learn the types of behavior that gov-

ernment actors are reportly engaged in. Doing so allows us to find types of reports that we might not have expected beforehand, such as the set of actions involving economic reforms.

4 Conclusion

This paper introduces two new techniques that together allow researchers to inductively learn political events from text. A slot filling model uses grammatical information and new machine learning models to identify the parts of a sentence corresponding to different "slots" in an event. It does so with much finer resolution than previous grammar-based event extraction models, and with far greater coverage than dictionary based methods. A second model takes these short spans and aggregates them into useful categories for further analysis. It overcomes the short document problem by using prior information in the form of word embeddings, a theoretically motivated document embedding scheme, and k-means clustering to learn useful aggregations of events. This model potentially has broader applicability beyond event extraction. It could be useful in other situations where very short documents need to be clustered.

I then apply the model to an open question in international politics, about whether respect for human rights as improved overtime. I produce new disaggregated data on human rights related events with government actors and offer some evidence for the arguments that the standard of reporting has changed over time. While the volume of human rights reporting has increased greatly over time, specific kinds of rights violations have changed in their overall proportion of reporting. Because the model is completely general it can be applied to a wide range of questions in political science, anywhere information on the behavior of actors is important.

As in the rest of science, the availability of new data is often the precipitating cause of new research and improved understanding. Automating some production of structured data from text would allow more project-specific creation of data, leading to better measurement strategies that use better data that is customized to the question at hand, and ultimately, improved understanding of the world.

5 References

Althaus, Scott L, Buddy Peyton, and Dan A Shalmon. 2018. "Spatial and Temporal Dynamics of Boko Haram Activity in 6 Event Data Pipelines." *APSA Mini Conference on Modern Event Data Development and Analysis*.

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. "A Latent Variable Model Approach to Pmi-Based Word Embeddings." *Transactions of the Association for Computational Linguistics* 4. MIT Press: 385–99.

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. 2017. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings." *ICLR*.

Baker, Collin F, Charles J Fillmore, and John B Lowe. 1998. "The Berkeley FrameNet Project." In *Proceedings of the 17th International Conference on Computational Linguistics-Volume 1*, 86–90. Association for Computational Linguistics.

Beieler, John. 2016. "Generating Politically-Relevant Event Data." *CoRR*. http://arxiv.org/abs/1609.06239.

Blei, David M, and John D Lafferty. 2007. "A Correlated Topic Model of Science." *The Annals of Applied Statistics* 1 (1): 17–35.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. "ICEWS Coded Event Data." *Harvard Dataverse* 12.

Boyd-Graber, Jordan L, and David M Blei. 2009. "Syntactic Topic Models." In *Advances in Neural Information Processing Systems*, 185–92.

Buntine, Wray. 2002. "Variational Extensions to Em and Multinomial Pca." In *European Conference on Machine Learning*, 23–34. Springer.

Carreras, Xavier, and Lluís Màrquez. 2005. "Introduction to the Conll-2005 Shared Task: Semantic Role Labeling." In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, 152–64. Association for Computational Linguistics.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Advances in Neural Information Processing Systems*, 288–96.

Cingranelli, David L, and David L Richards. 2004. "CIRI Human Rights Dataset." http://www.humanrightsdata.com.

Cingranelli, David, and Mikhail Filippov. 2018a. "Are Human Rights Practices Improving?" *American Political Science Review* 112 (4). Cambridge University Press: 1083–9.

———. 2018b. "Problems of Model Specification and Improper Data Extrapolation." *British Journal of Political Science* 48 (1). Cambridge University Press: 273–74.

Demszky, Dorottya, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse

Shapiro, and Dan Jurafsky. 2019. "Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings." *17th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv Preprint arXiv:1810.04805*.

Dowty, David. 1991. "Thematic Proto-Roles and Argument Selection." *Language* 67 (3). Linguistic Society of America: 547–619.

Fariss, Christopher J. 2014. "Respect for Human Rights Has Improved over Time: Modeling the Changing Standard of Accountability." *American Political Science Review* 108 (2). Cambridge University Press: 297–318.

———. 2018. "Are Things Really Getting Better? How to Validate Latent Variable Models of Human Rights." *British Journal of Political Science* 48 (1). Cambridge University Press: 275–82.

Geddes, Barbara, Joseph Wright, and Erica Frantz. 2014. "Autocratic Breakdown and Regime Transitions: A New Data Set." *Perspectives on Politics* 12 (02): 313–31.

Gerner, Deborah J., Philip A Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr. 2002. "Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions." *International Studies Association, New Orleans*.

Gildea, Daniel, and Daniel Jurafsky. 2002. "Automatic Labeling of Semantic Roles." *Computational Linguistics* 28 (3). MIT Press: 245–88.

Goemans, Henk E, Kristian Skrede Gleditsch, and Giacomo Chiozza. 2009. "Introducing Archigos: A Dataset of Political Leaders." *Journal of Peace Research* 46 (2). Sage Publications Sage UK: London, England: 269–83.

Goldberg, Yoav. 2017. *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Greene, Kevin T, Baekkwan Park, and Michael Colaresi. 2019. "Machine Learning Human Rights and Wrongs: How the Successes and Failures of Supervised Learning Algorithms Can Inform the Debate About Information Effects." *Political Analysis* 27 (2). Cambridge University Press: 223–30.

Halterman, Andrew. 2019. "Geolocating Political Events in Text." *NLP+CSS Workshop, NAACL*.

Hoffman, Matthew, Francis R Bach, and David M Blei. 2010. "Online Learning for Latent Dirichlet Allocation." In *Advances in Neural Information Processing Systems*, 856–64.

Honnibal, Matthew, and Ines Montani. 2017. "SpaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing." *To Appear*.

Howard, Jeremy, and Sebastian Ruder. 2018. "Universal Language Model Fine-Tuning

for Text Classification." *arXiv Preprint arXiv:1801.06146v2*.

Jones, Daniel M, Stuart A Bremer, and J David Singer. 1996. "Militarized Interstate Disputes, 1816–1992: Rationale, Coding Rules, and Empirical Patterns." *Conflict Management and Peace Science* 15 (2): 163–213.

Kornilova, Anastassia, Daniel Argyle, and Vlad Eidelman. 2018. "Party Matters: Enhancing Legislative Embeddings with Author Attributes for Vote Prediction." *arXiv Preprint arXiv:1805.08182*.

Kuhn, Harold W. 1955. "The Hungarian Method for the Assignment Problem." *Naval Research Logistics Quarterly* 2 (1-2). Wiley Online Library: 83–97.

LaFree, Gary, and Laura Dugan. 2007. "Introducing the Global Terrorism Database." *Terrorism and Political Violence* 19 (2): 181–204.

Lauretig, Adam M. 2019. "Identification, Interpretability, and Bayesian Word Embeddings." *arXiv Preprint arXiv:1904.01628*.

Le, Quoc, and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." In *International Conference on Machine Learning*, 1188–96.

Levy, Omer, and Yoav Goldberg. 2014. "Neural Word Embedding as Implicit Matrix Factorization." In *Advances in Neural Information Processing Systems*, 2177–85.

Makarov, Peter. 2018. "Automated Acquisition of Patterns for Coding Political Event Data: Two Case Studies." In *Proceedings of the Second Joint Sighum Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 103–12.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." In *Advances in Neural Information Processing Systems*, 3111–9.

Nardulli, Peter F, Scott L Althaus, and Matthew Hayes. 2015. "A Progressive Supervised-Learning Approach to Generating Rich Civil Strife Data." *Sociological Methodology* 45 (1): 148–83.

Norris, Clayton, Philip Schrodt, and John Beieler. 2017. "PETRARCH2: Another Event Coding Program." *The Journal of Open Source Software* 2 (9).

O'Connor, Brendan, Brandon Stewart, and Noah A Smith. 2013. "Learning to Extract International Relations from Political Context." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* Vol. 1.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. "The Proposition Bank: An Annotated Corpus of Semantic Roles." *Computational Linguistics* 31 (1). MIT Press: 71–106.

Palmer, Martha, Daniel Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling*. Edited by Graeme Hirst. Synthesis Lectures on Human Language Technologies. Morgan & Clay-

pool Publishers.

Pavlick, Ellie, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. 2015. "FrameNet+: Fast Paraphrastic Tripling of Framenet." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2:408–13.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Empirical Methods in Natural Language Processing (Emnlp)*, 1532–43. http://www.aclweb.org/anthology/D14-1162.

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep Contextualized Word Representations." *arXiv Preprint arXiv:1802.05365*.

Powell, Jonathan, and Clayton Thyne. 2011. "Global Instances of Coups from 1950 to 2010: A New Dataset." *Journal of Peace Research* 48 (2): 249–59.

Raleigh, Clionadh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature." *Journal of Peace Research* 47 (5): 651–60.

Raytheon BBN Technologies. 2015. "BBN Accent Event Coding Evaluation." Technical report.

Rheault, Ludovic, and Christopher Cochrane. 2019. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis*. Cambridge University Press, 1–22.

Ritter, Alan, Evan Wright, William Casey, and Tom Mitchell. 2015. "Weakly Supervised Extraction of Computer Security Events from Twitter." In *Proceedings of the 24th International Conference on World Wide Web*, 896–905. International World Wide Web Conferences Steering Committee.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, and others. 2013. "The Structural Topic Model and Applied Social Science." In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Ruder, Sebastian. 2018. "NLP's Imagenet Moment Has Arrived." *The Gradient* https://thegradient.pub/nlp-imagenet/ (July).

Rudinger, Rachel, and Benjamin Van Durme. 2014. "Is the Stanford Dependency Representation Semantic?" In *Proceedings of the Second Workshop on Events: Definition, Detection, Coreference, and Representation*, 54–58.

Salehyan, Idean, Cullen S Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams. 2012. "Social Conflict in Africa: A New Database." *International Interactions* 38 (4): 503–11.

Schrodt, Philip A. 2009. "TABARI: Textual Analysis by Augmented Replacement Instruc-

tions." *Dept. of Political Science, University of Kansas, Blake Hall, Version 0.7. 3b3*, 1–137.

Schrodt, Philip A, Shannon G Davis, and Judith L Weddle. 1994. "Political Science: KEDS—a Program for the Machine Coding of Event Data." *Social Science Computer Review* 12 (4): 561–87.

Spirling, Arthur, and P Rodriguez. 2019. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." Working paper.

Sundberg, Ralph, Kristine Eck, and Joakim Kreutz. 2012. "Introducing the UCDP Non-State Conflict Dataset." *Journal of Peace Research* 49 (2): 351–62.

Van Atteveldt, Wouter, Tamir Sheafer, Shaul R Shenhav, and Yair Fogel-Dror. 2017. "Clause Analysis: Using Syntactic Information to Automatically Extract Source, Subject, and Predicate from Texts with an Application to the 2008–2009 Gaza War." *Political Analysis* 25 (2). Cambridge University Press: 207–22.

White, Aaron Steven, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. "Universal Decompositional Semantics on Universal Dependencies." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1713–23. Austin, Texas: Association for Computational Linguistics.

Wood, Reed M, and Mark Gibney. 2010. "The Political Terror Scale (Pts): A Re-Introduction and a Comparison to Ciri." *Human Rights Quarterly* 32 (2): 367–400.

Zhukov, Yuri M, and Brandon M Stewart. 2013. "Choosing Your Neighbors: Networks of Diffusion in International Relations." *International Studies Quarterly* 57 (2). Wiley Online Library: 271–87.

6 Appendix A

```
    verbs" :  0 : "shoot kill fire attack"
              1 : "assist help provide bring"
              2 : "meet talk discuss"
              3 : "gather protest demonstrate carried march"
              4 : "capture abduct"
              5 : "deliver provide"
              6 : "seize take capture overrun recapture"
              7 : "occupy control"


 "objects" :  0 : "village town villager civilian militant"
              1 : "aid convey help village town development government"
              2 : "karzai ambassador embassy"
              3 : "demonstration chant opposition near "
              4 : "aid worker civilian innocent"
```

```
5 : "humanitarian water food aid"
6 : "base territory village control area stronghold"
7 : "base territory village area stronghold"
```