Comparing NLP Methods for Identifying Policy Decisions in Government Documents[*]

Natalie Ahn
University of California, Berkeley

Abstract

Advances in natural language processing (NLP) are making it easier to extract events from text, not only to analyze topics or tones, but to dissect who did what to whom. In political studies, these methods have mainly been applied to international conflict events as reported in news media. Much less work has been done to study internal government decisions in similar ways. Yet government policies and institutions also change dynamically, through decisions and actions that are recorded in laws and decrees. A major challenge is that existing event extraction systems do not extend well to new types of text or substantive domains. News-based event data projects rely heavily on pattern matching and supervised machine learning, using manually constructed event frames, entity taxonomies, and annotated corpora. Yet government documents may lend themselves well to less supervised forms of event identification and extraction. In this paper, I explore and compare rule-based, supervised, and unsupervised computational approaches to labeling the contents of executive decrees, in order to produce data on different types of policy actions and the entities they affect. I apply these methods to executive decrees from Peru, as part of a broader study on presidential power in developing states. I present two different forms of evaluation to determine the best methods for the research goals, including accuracy tests in relation to human coding, and usefulness in capturing relationships to other real-world phenomena.

[*] Paper Prepared for the 2017 Poli Informatics Workshop

1 Introduction

Questions about government policy decisions and institutional change are difficult to study. Concrete actions like the creation of new offices or programs, delegation of authorities, and imposition of regulations, are recorded in publicly available laws and executive orders or decrees. Yet those laws and decrees are written in human language, not codified in data structures or numeric values. Since laws and decrees are used to enact many different types of policy decisions, the total volume of such documents is often a poor measure of specific actions or events they might contain. To dissect different policy actions and the widely varying factors that motivate or constrain them, we may need to look at what the laws or decrees actually say they do, and yet do so for a large volume of such documents over time.

Advances in natural language processing (NLP) are making it easier to extract events automatically from text, to identify not only themes but specific actions and participants. In political studies, most event extraction systems rely on pattern matching and supervised machine learning, using manually constructed event templates, domain-specific entity taxonomies, and large-scale annotated training corpora. The costs of constructing these resources make it difficult to apply the same methods to new languages or domains, such as to public records in developing countries. Yet government documents may be well suited to automated classification or information extraction, potentially enabling less resource-intensive approaches.

In this paper, I explore computational approaches to measure certain types of government decisions recorded in executive decrees, as part of a broader study on presidential power in developing states. I compare different options for input features, coding scheme, and classification algorithms, including rule-based information extraction, supervised machine learning, and unsupervised clustering and topic modeling. I also present two forms of evaluation: standard accuracy tests in relation to human coding, and an explorative effort to assess the resulting data's usefulness for studying relationships to other real-world phenomena.

The paper proceeds as follows. Section 2 discusses related work and the types of documents and methods typically used to study similar research questions. Section 3 describes this project's choices of source documents and computational algorithms. Section 4 presents accuracy tests in relation to human coding, comparing the different methods chosen, along with an alternative form of assessment that may aid in the discovery of useful concepts or categories. Section 5 concludes.

2 Related Work

2.1 Sources of Information on Government Action

Researchers studying government action and political institutions often collect data from secondary news media, surveys and polls, or abstract indicators compiled and coded by experts. The use of raw text reports as sources of data has become more common in recent years, in part due to the explosion of information available in digital news and other online media (Tanev et al., 2008). With regard to government actors, scholars tend to focus on activities directed outward, such as violent conflicts between states, which are frequently observable in secondary news reporting (Bond et al., 2003; Raleigh et al., 2010). News reports tend to be informative and straightforward, structured with the most important details of a story at the top, which are useful for monitoring crises and detecting major emerging events (Tanev et al., 2008).

However, news media pose challenges to measuring events consistently over time. News reporting is often redundant, so that researchers must deconflict reports about the same event to get accurate counts. News reporting is also influenced by many factors other than whether events occurred, such as resource constraints, reporter access, target audience interest, and the publication's business objectives (Kepplinger, 2002; Althaus et al., 2011; Weidmann, 2015). In terms of scope, news reports are not designed to account for every minor policy decision or external action, and are more useful for studying large-scale attention-grabbing events like new wars or regime changes, rather than everyday decisions and interactions (Ortiz et al., 2005).

There are many research questions, however, that involve the day-to-day business of government and the evolution of public offices and authorities. The best or only sources of those everyday activities are often official government records. Governments are increasingly making legislative and executive records publicly available in digital archives, offering new opportunities to study government authority and action in systematic detail. NLP tools have been used to identify topics and sentiment in legislation and congressional debates (Thomas et al., 2006; Purpura and Hillard, 2006), and to extract key details from court rulings to assist legal researchers in finding relevant prior cases (Brüninghaus and Ashley, 2001; Jackson et al., 2003).

2.2 Methods for Extracting Actions and Entities

In political science research, the most common approaches to automatically categorize text utilize the frequency of words appearing in each document, regardless of grammar or word order, i.e. "bag-of-words" (BOG) techniques (Hopkins and King, 2010; Grimmer and Stewart, 2013; Biagioli et al., 2005). These word frequencies are often used to assign a pre-defined label or category to each document as a whole. Researchers hand label a set of training examples, then train supervised machine learning models to predict which combinations of words should be assigned to which categories (Biagioli et al., 2005; Thomas et al., 2006; Purpura and Hillard, 2006).

Alternatively, if researchers want to learn unknown categories, unsupervised or inductive approaches can be used to look for new patterns. The most popular is Latent Dirichlet Allocation (LDA) for topic modeling (Grimmer, 2010; Gerrish and Blei, 2012; Roberts et al., 2014). Topic modeling is also typically performed using bag-of-words features, so the learned topics are represented by distributions of words. In this way, bag-of-words features are often used to reveal documents' themes or tones, but they do not indicate which elements in the text play which functional roles.

A less common but growing approach to text-based data in the social sciences is automated event extraction, using more complex processes to parse who did what to whom. Earlier efforts relied on keyword searches and pattern matching, using shallow parsing and manually constructed dictionaries and phrase patterns, as in prominent English-language projects on violent conflict (King and Lowe, 57; Schrodt, 2006) and systems to automate the indexing and retrieval of case law (Brüninghaus and Ashley, 2001; Jackson et al., 2003). These systems tend to be designed and built wholesale for each project, requiring large teams and considerable resources to implement.

In recent years, open source NLP tools have made the task of event extraction more accessible. These tools include **part-of-speech taggers**, which tag words as verbs, nouns, prepositions, etc, and **dependency parsers**, which label words' grammatical relations like the subject, direct and indirect objects of a given verb (Klein and Manning, 2003; de Marneffe et al., 2006). Instead of searching for fixed-order phrases or sentence patterns, researchers may construct logical rules, using automatically labeled parts of speech and grammatical relations, to identify which actors played which roles in events (Schrodt, 2014). For instance, a researcher might look for any form of the verb "attack", and assign the roles of *attacker* to its subject and *target* to its direct object. Researchers have also used supervised machine learning to assign semantic roles, by first annotating training corpora at the word level, although such annotation is especially labor-intensive (Kim et al., 2008; Caselli et al., 2011).

Computational linguists have also begun to experiment with unsupervised or inductive approaches to learning event templates. Some researchers have used pipeline approaches in which they separately cluster verbs into event types and those verbs' noun phrase arguments into event roles (Chambers and Jurafsky, 2011; Ahn, 2017). Others have used generative models to jointly learn event types and entity roles within each document (Chambers, 2013; Cheung et al., 2013; Nguyen et al., 2015; Sha et al., 2016). These inductive methods are not yet very effective at inducing the same event types and roles that researchers might specify; the best accuracy scores are usually in the 40s when compared to hand-labeled test documents. There are also not yet available off-the-shelf tools for social scientists to use these methods for new empirical applications. In this paper, I make use of some of the ideas from this emerging body of work, while evaluating remaining challenges and potential areas for improvement.

3 Methodology and Empirical Application

3.1 Case Selection

This paper is part of a larger project studying the use of presidential power in five South American countries: Peru, Ecuador, Bolivia, Colombia, and Venezuela. There is growing concern in many countries about strong presidents making decisions on their own through executive orders or decrees. Yet there is also often considerable support for strengthened executive power, especially amid gridlock or corruption in other parts of the political system. Debates about concentrated power are especially relevant in countries with developing or transitioning political institutions, such as in South America after the last wave of military regimes departed in the late 20th century. In the Andean states, multiple presidents have been removed before the end of their terms, and many former presidents have been prosecuted for corruption or other abuses of power (Perez-Linan, 2007; Valenzuela, 2004; Conaghan, 2012).

I chose to use Peru for an initial exploratory study, to refine the overall project's substantive hypotheses and develop the data collection and measurement strategy, which is presented in this paper. Peru has experienced considerable variation in the power of the president in recent decades, including the rise and eventual removal of a strongman who dismissed the legislature and governing by decree, in the face of economic crisis and congressional opposition (Conaghan, 1996; Montero, 2001). While subsequent presidents have been more constrained, decrees continue to be an important decision-making tool used in a variety of ways (Wright, 2014). Peru has also made decades of legislation and decrees accessible through digital archives, as part of efforts to improve government transparency (Miguel-Stearns, 2011). I discuss the collection of those documents in the next section.

3.2 Document Sources

I have collected information on over 18,000 executive decrees issued in Peru from 1995 to 2016, the period since the most recent Constitution was enacted. I include the three main types of executive decrees in Peru (*decretos supremos*, *decretos legislativos*, and *decretos de urgencia*) to maximize coverage of the various policy decisions made through decrees, and to make the data comparable to the other countries in the larger study, which distinguish types of decrees in different ways.

I've chosen to extract data about presidential decisions and actions directly from official decrees for multiple reasons. As discussed in the literature section, news reports and other digital media have many limitations in terms of coverage and biases. In contrast, laws and decrees are primary source documents that themselves enact the policy decisions or institutional changes they report. This means that each law or decree constitutes a separate action, which do not need to be deconflicted for redundancy, and researchers can assemble a set of all relevant decisions enacted through a particular channel, if a reasonably complete archive is maintained.

The language used in government documents is also more formal and consistent than news reports, since laws are designed to be authoritative, rather than attention-grabbing or quick reads. Laws and decrees use exact legal terminology, with official organization names written out, and the same verb phrase regularly represents the same type of policy action. For instance, the verbs *nombrar* ("to name") and *designar* ("to designate") are consistently used for official appointments, while news reports might use a wider variety of colloquialisms like "to pick" or "to tap" a person for a job. Also, fewer event roles may need to be extracted, because the agent or doer can usually be interpreted to be the authority enacting the law or decree. Researchers may only need to find the main verb or action and the main object of that action, turning questions of "**who** did **what** to **whom**" into "**what** was done to **whom**".

However, researchers still face many challenges when seeking to collect and process a large volume of full text government documents, especially in developing countries. Despite progress in the availability of public records, there are still considerable differences in quality, completeness, and usability across different archives. There are often multiple archives for the same type of document in each country, which may not have identical contents. Historic decrees may be published in an official gazette or daily registry, a subset may be available on the president's website, or that of an agency responsible for public records, and additional collections may be available as part of the legislature's archives or maintained by an office of the judiciary. In Peru, decrees appear in the official gazette (Diario Oficial) El Peruano[1], the Digital Archive of Legislation of Peru[2] maintained by the Congress, and the Peruvian System of Legal Information[3] maintained by the Ministry of Justice and Human Rights.

In terms of access and use, laws and decrees are considered public domain in many countries, and transparency laws often require government records to be made publicly available.[4] Yet to cover the costs of archival services and manage server load, some databases charge fees and/or impose download limits, even while the same documents are available for free or without limit on another website. As with news media, norms are still developing regarding database access for those seeking to crawl, scrape, or otherwise mine archives for research purposes, especially in countries outside the United States (Truyens and Eecke, 2014).

Finally, it is more common to find lists of metadata for government documents than full text files, especially going back more than a few years. When historic full text documents are available, they are often scanned images that are not machine readable. Optical character recognition software for non-English text still produces many errors, especially when used on grainy images. Even when machine readable, full text laws and decrees are also difficult to parse. Legal norms may contain substantial non-operational preamble language, and may be annotated with legislative histories, alternative terminology, or other explanatory notes throughout.

For this project, I've chosen to use the titles of decrees, which are most widely available across countries, archives, and years. The titles also tend to have similar linguistic structure even across countries, centered around a recognizable policy action phrase and a target entity name which the law or decree governs. The titles can be

[1] http://www.elperuano.com.pe/

[2] http://www.leyes.congreso.gob.pe/

[3] http://spij.minjus.gob.pe/

[4] For instance, Peru's LAW No. 27806 of 2003, "Law on Transparency and Access to Public Information," http://www.peru.gob.pe/normas/docs/LEY_27806.pdf

thought of as combining some of the brevity and directness of news reporting with the authority and formality of official records. Law titles are designed to clearly identify the law by the gist of its main provision(s), in the correct legal terms and with enough detail to distinguish it from similar documents, yet short enough to fit on a few lines. Titles may be inadequate for more detailed information extraction, such as when seeking to map the history of specific sub-law provisions. But titles generally suffice to identify the main actions and target entities that I seek to measure for this project.

3.3 Measurement Strategy

The task of automatically encoding text into comparable measures for analysis involves several decisions: 1) what information (features) to use from the original text, 2) what the resulting categories or labels should look like, and 3) what encoding process to use to transform the first into the second. These decisions involve a number of trade-offs, including what type and how much training data is needed, how much effort is required to implement the encoding process, and how well the resulting data captures the concepts of interest for the research goals.

Since I'm studying types of documents and policy actions for which there are not yet established best practices, I've chosen to compare multiple options for answering each of the above questions. I've selected several algorithms used for document classification and information extraction, including popular off-the-shelf classifiers and more original processes that required project-specific implementation. I've also chosen to compare algorithms that use different types of input features – bag-of-words term frequencies versus more complex syntactic features – and to compare supervised algorithms with unsupervised learning approaches.

The chosen options are summarized in Table 1 and explained in sections below.

Table 1: Summary of Methodological Options Selected for Comparison

1. Input Features	**pros**	**cons**
Bag-of-words (doc term freqs)	Easy to extract, existing tools, little preprocessing required	Discards word order, grammar which might be important
Structured features (main verbs and noun objects)	Enables extraction of info on events/relations, may produce more accurate doc-level classification too	More effort to extract, requires parsing (existing tools), turning select components into numeric vectors (no general-use tools)

2. Output Labels	**pros**	**cons**
Doc-level labels	Easier to hand-label training data, good for topics/tones	Does not indicate which components fill which roles
Word/clause-level labels	Enables extraction of info on event roles and entity relations	More work to annotate training docs or write extraction rules
Intermediate-level labels	One main action label per document, One target entity label per document	
Few high-level categories	Simpler classification task, need fewer training docs	May oversimplify, labels too abstract to learn or be useful
More low-level categories	Preserves more distinctions, may matter for research goals	Requires more training docs to cover all categories

3. Encoding Models	**pros**	**cons**
Rule-based info extraction *(project-specific)*	Requires less labor to get easy cases right, may get decent scores with fairly simple rules, very fast	Hard to specify rare/subtle cases, tail takes more effort with diminishing returns, may be an upper limit on potential accuracy
Logistic regression *(scikit-learn)*	Should be capable of highest accuracy with enough training data	Need lots of labeled training data, can only reproduce human coding, biases in input retained in output
Agglomerative clustering *(w/ scikit-learn)*	Needs no labeled training data, can use any distance metric, learn new concepts	Hard to evaluate, possibilities require more data/parameters to converge on useful output
LDA topic modeling *(gensim)*	Needs no labeled training data, can use popular off-the-shelf tools	Hard to evaluate, must specify num topics, off-the-shelf tools only use doc-level BOG features

3.3.1 Coding Scheme, Output Labels

For this project, I am interested in identifying what type of decision or change the president enacted through each decree, and who the primary target of that action was. For instance, decrees that create new executive agencies or transfer resources to them are more likely to expand the president's power and enable a wider range of future executive actions. Decrees that impose regulations on private actors, restructure taxes, or approve one-off public works projects, may matter a great deal in terms of who gains or loses. But the latter external actions do not themselves alter which internal government actors have the power to make future policy decisions.

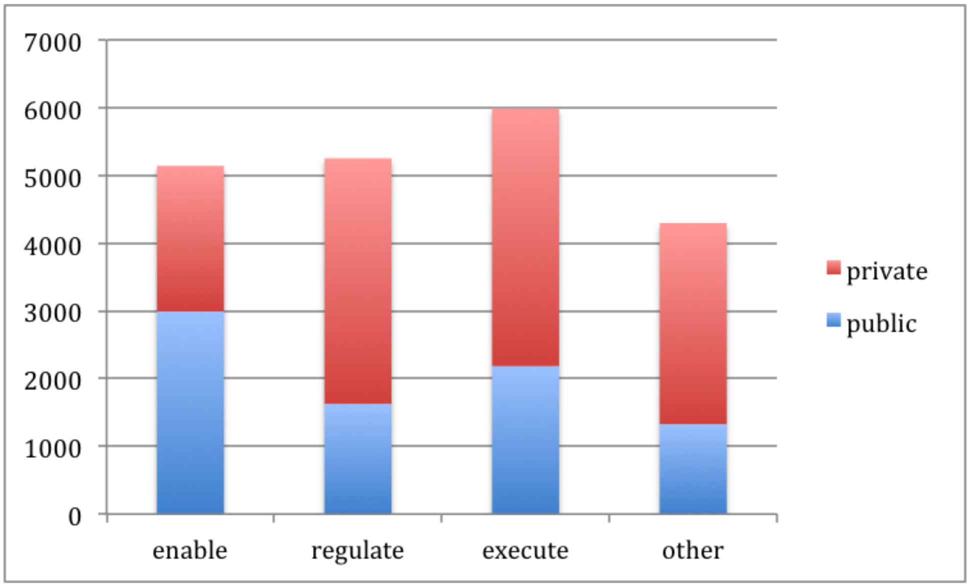
The desired information involves a bit more than an overall document theme, but does not require an encoding of every single aspect of each decree event, such as duration or target location. I've chosen a coding scheme that balances the research goals with feasibility constraints; the resulting scheme falls between a single document-level category (e.g. a topic) and a full event schema (e.g. with agent, target, time, place, and means). I seek to identify two pieces of information for each document: a main action and a target entity, both grouped into a limited set of categories.

Table 2 shows the chosen categories. In them, I've preserved two levels of granularity, to explore what works best: a higher level with only a few broad action and entity groups, and a lower level with more subtypes. In the target entity list, there are more subtypes for public actors than private. This is because distinctions between executive agencies and other government actors are likely to matter more to a policy-maker's motivations, than the wider variety of private actors that decrees may affect. Figure 1 summarizes decrees in the project dataset (Peru from 1995 to 2016), broken down by high-level category using the rule-based classifier defined later on.

Table 2: Coding Scheme with main action and target entity categories

Category	Subcategories	Example Terms
Action		
enable	*empower_enable,* *finance_enable,* *modify_enable*	create, appoint, authorize ... transfer, fund ...
regulate	*initiate_regulate,* *modify_regulate*	regulate, require, limit ...
execute	*initiate_execute,* *modify_execute*	plan, transact, implement ...
other	*enact_other,* *modify_other*	adopt, ratify ...
Target		
public	*public_executive,* *public_legislature,* *public_judiciary,* *public_local,* *public_other*	ministry, cabinet ... legislature, congress ... court, tribunal ... municipality, province ...
private	*private_business,* *private_other*	corporation, industry ... youth, workers, voters ...

Figure 1: Decrees by action (x-axis) and target (stacked) in the dataset

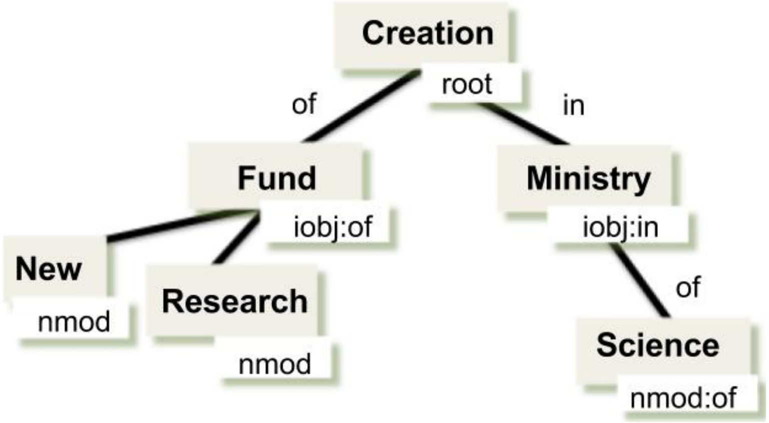


3.3.2 Preprocessing and Feature Extraction

Bag-of-words features: Bag-of-words features require very little preprocessing of the raw text. The main step is to construct a document-term matrix, i.e. a vector of term counts for each document. I use the top 1000 most frequent terms in the corpus as the columns in this matrix. To reduce the feature space, I lemmatize terms to their root form (i.e. infinitive for verbs, singular male for nouns), and weight the count vectors by term frequency and inverse document frequency (TF-IDF) to emphasize terms more likely to predict specific labels.

Structured features: To extract features that capture more linguistic structure from the text, I perform several preprocessing steps involving off-the-shelf NLP tools, existing general-domain knowledge resources, and a limited degree of project-specific knowledge engineering. The first step is to apply part-of-speech tagging and dependency parsing, for which I use the Spanish language models in the Stanford CoreNLP toolkit (Manning et al., 2014). These tools produce a tree structure for each sentence rooted at its main verb, as shown in Figure 2.

Figure 2: Parse tree for a typical decree title:
Creation of New Research Fund in the Ministry of Science and Technology



Main action verbs: The next step is to extract each document's main action verb and target entity noun phrase. To do so, I search each parsed decree title for verbs and nouns in certain grammatical positions. For verbs, the highest-level active verb in the parse tree usually appears in the "root" dependency position, or immediately after an enabling verb like "propose" or "declare", as in "Propose to *create* a new office ...". Actions may be in nominal form, as long as the noun falls under the hypernym for "event" in WordNet (discussed below), as in "Propose the *creation* of a new office ...". Throughout the paper, I refer to these action terms as "verbs" for readability, althought they may be eventful noun predicates too.

Noun objects: For target entities, I am interested in the target objects of the main action verbs. The targets might be direct objects, as in "Create *the ministry of* ...", or indirect objects that appear after a preposition like "to" or "in", as in "Transfer funds *to the ministry* ...". In the latter case, the direct object ("funds") is a general resource that might be transferred to any organization. The direct object clarifies what action is being taken (i.e. a financing operation), but not which entity is being funded. The indirect object following "to" ("the ministry") is the organizational target of the action. For simplicity, in this preprocessing step, I extract all direct and indirect objects of the main action verbs to use as potential target entities.

Hypernyms: Finally, we need a way to begin to group the extracted terms, either directly through rules, or probabilistically based on their similarities. It is much more difficult to specify every possible attack-invoking pattern like "Russia attacked China" or "North Korea attacked Japan", than it is to specify a template like "[Country A] attacked [Country B]" and then label noun phrases in the corpus as country names so that they can be matched to the looser template pattern. However, defining a dictionary of all possible terms in the corpus with their associated entity types for every project would be incredibly costly.

Instead, I use WordNet, a lexical database of over 100,000 word senses with definitions and hierarchical relationships. A Spanish language version of the Multilingual Central Repository (MCR) (Gonzalez-Agirre et al., 2012) is available through the Open Multilingual WordNet interface in the Python Natural Language Toolkit (NLTK) (Bond and Paik, 2012). I look up noun objects in the WordNet hierarchy to identify their higher-level parent terms (i.e. hypernyms), and include those in the decree title's structured features. Since WordNet's word senses and hierarchy do not always capture the most relevant meaning or relationships for each specific project, I've chosen to augment this general resource with a limited project-specific taxonomy. I define a set of terms that take on a particular meaning in the context of Spanish legal documents, and the hypernyms that should be assigned to them when they appear in the dataset. For all other terms, I look them up in WordNet and assign comparable hypernym features, keeping the added effort to a minimum.

3.3.3 Classification Algorithms

For classification algorithms, I've chosen two document-level classifiers that are popular among social scientists: a supervised machine learning classifier and unsupervised topic modeling. For the supervised classifier, I use Logistic Regression (LR), which performed best among *scikit-learn*'s supervised classifiers in testing, and the *gensim* package's implementation of Latent Dirichlet Allocation (LDA) for topic modeling.

Since the supervised machine learning and LDA models operate at the document level, I run all classifiers twice on the full document set, once to assign each document a main action label and once to assign a target entity label. I repeat this process using the bag-of-words features and the more structured features as inputs. For bag-of-words models, the same vectors are used to classify both actions and target entities. For the structured features, slightly different vectors are used that roughly parallel the information I use in the rule-based approach below. I use both main verbs and their noun objects to classify action types, since “establish a fund” is an enabling action, while “establish a regulation” is a restricting action. I only include the noun objects (and their hypernyms) as input features when assigning target entity categories.

3.3.4 Information Extraction Algorithms

I’ve also chosen two approaches that operate at the sub-document level and are more deterministic (as opposed to probabilistic), allowing more direct manipulation of how certain verb and noun features may be encoded into action and target categories. I again use one method with pre-defined labels (as in supervised learning, but this time using fixed rules) and one unsupervised method (using agglomerative clustering).

For the **rule-based system**, I define conditions for which documents should be labeled with each of the categories in the project coding scheme, using the features extracted in preprocessing. Each rule includes a condition for the document’s main action verb, plus potential conditions for that verb’s objects. The object conditions specify a dependency relation (e.g. the verb’s direct object or indirect object) and permitted hypernyms for the noun that should appear in that object position. If a decree title has a main verb matching the permitted verbs for a given rule, and that verb also has a noun with a matching hypernym in the right object position, the document is assigned the action category corresponding to that rule. Rules for target categories follow the same format, but any main verb is allowed, only the noun object conditions matter. Table 3 below shows example rules.

For **unsupervised learning** of action and target entity categories, I take the same main verbs and noun objects extracted in preprocessing, and cluster them into action and entity groups. I use agglomerative clustering, which builds a hierarchy of clusters by assigning each term to its own cluster and then merging clusters one

Table 3: Example rules to match verb-object clauses to action or target labels

category	**subcategory**	**= verb + dependency relation : [hypernym]**
Action		
enable	*finance_enable*	= “financiar” + [any object]
enable	*finance_enable*	= “transferir” + dobj : [assets]
Target		
public	*public_executive*	= [any verb] + iobj : [government department]

by one in order of their distance apart. This process is based on a more complex implementation developed for news reports in Ahn (2017).

For the distances between pairs of verbs or nouns, I construct feature vectors and then calculate the cosine distance between them. For main verbs, the features used are counts of: 1) terms that appear as the verb’s dependents throughout the corpus, 2) hypernyms of the verb’s dependents throughout the corpus, and 3) the verb’s synonyms in WordNet (i.e. other terms that share one of the verb’s WordNet synsets). For noun objects, the features are counts of: 1) the noun’s hypernyms (in the project-specific taxonomy or in WordNet otherwise), 2) the noun’s WordNet synonyms, and 3) verbs of which the noun is a dependent throughout the corpus.

I cluster verbs and nouns separately, using average linkage distances between the elements of two clusters, and stop clustering at a maximum distance between clusters optimized on a training set (defined in the evaluation section). I also apply constraints against merging terms that seem highly unlikely to be of the same type. For verbs, I prevent two verbs from ending up in the same cluster if they share no common synonyms or dependent terms. For nouns, I constrain all clusters to have only one of the following high-level entity types: 1) Person or Organization, 2) Location, 3) Physical Object, 4) Document, or 5) Other; similar to the entity types used in event schema induction papers (Chambers and Jurafsky, 2011; Cheung et al., 2013).

4 Evaluation

4.1 Accuracy in Relation to Human Coder

The most common way to evaluate automated document classification is to test a model's accuracy against a baseline of hand-coded test examples. This type of test is appropriate if our goal is to train a machine to be able to replicate the same coding decisions that we've made by hand, meaning that it makes the most sense for supervised classification. Accuracy tests are straight-forward to implement and interpret; a machine that performs well may be very valuable for automating previously labor-intensive processes. I test all models on a random sample of 500 decrees from the project dataset, which I've hand labeled using the project coding scheme.

To ensure that no model has overfit (or simply memorized) the training examples, we need to conduct accuracy tests using a "held-out" test set. For the supervised machine learning models, I use 10-fold cross validation, iteratively reserving a different 10th of the labeled documents, training on the remaining 9/10ths, testing on the held-out 10th, then averaging the scores from all folds. For the unsupervised models, I use a comparable method, in which I iteratively optimize the necessary parameters (maximum distance for clustering; number of topics and *alpha* parameter for LDA) on 9/10ths of the documents, then test the model on the remaining 10th.

Finally, for the rule-based extraction system, I wrote the rules while hand-coding the first 100 documents, then refined them through tests on the next 50 documents. At that point, I froze the coding scheme and rules, and hand coded the remaining 350 out of 500 total documents without any further changes or error correction tests. I evaluate the accuracy of the rule-based system on the latter 350 labeled documents.

4.1.1 Results

I sought to test all combinations of the methodological options that made sense, varying the classification model, input features, and granularity of output labels. Precision, recall, and F1-scores are reported in Table 4 (in percentages for readability).

Table 4: Evaluation against human coding, comparing models and label granularity

			Many labels			Few labels		
	Model	Features	P	R	F1	P	R	F1
Known labels	Rules	Structured	71	77	**74**	78	83	**80**
	LR	BOG	61	61	**61**	76	76	**76**
	LR	Structured	69	70	**69**	80	80	**80**
Induced labels	Cluster	Structured	78	43	**55**	84	31	**45**
	LDA	BOG	71	41	**51**	72	34	**45**
	LDA	Structured	67	42	**50**	75	33	**45**

The rule-based system performs best overall, with an F1-score of 0.737 for extracting the more fine-grained categories, and 0.804 for extracting the higher-level categories. Supervised machine learning (using logistic regression, LR in the table) nearly matches the rule-based system with an F1-score of 0.798, when using structured verb and noun features and learning only the simpler high-level labels. More than 500 labeled documents would be needed to achieve comparable accuracy with supervised machine learning for more complex coding schemes. Unsupervised (inductive) models perform worst, which is expected, since the computer has not been told what labels to use. The "correct" labels in the project coding scheme were only mapped, afterwards, to the most similar of the clusters the machine induced.

In terms of input features, models that utilize more structured main verb and noun object features generally perform better than models that rely on bag-of-words features, although the differences vary across models. Some models are only suited to one type of features, such as the rule-based and clustering approaches, which I designed to operate at the level of specific verbs and nouns in a parse tree, rather than at the document level where simple word frequencies are relevant.

Notably, the output labels that are easiest to learn with supervised models are not the same as those best learned with unsupervised models, as highlighted in gray in Table 4. Supervised classifiers do better at learning a smaller number of higher-level concepts, but induced labels more accurately map to a larger number of lower-level labels. The results indicate that it may make sense to induce fairly fine-grained categories, when using unsupervised methods, then use rules or more supervised selection to aggregate those concepts into larger groups relevant to the substantive application.

Consider, for instance, the target categories of *public* and *private*, and their subcategories *public_executive*, *public_legislature*, and *private_business*. A supervised classifier can easily learn that a user wants terms like both “cabinet” and “congress” to be assigned to the public category. But these terms might be used very differently in the corpus, so that similarities in usage and context won’t be enough to group them inductively. While an unsupervised classifier might easily group terms commonly used for executive agencies, it is much less likely to learn the abstract concept of public sector affiliation that links those executive agencies to legislative bodies, and might find more similarities to the set of terms that represent private corporations.

The unsupervised classifiers also have much lower recall scores than precision, meaning that the closest matching induced categories tend to capture a good subset of a labeled category (without a lot of wrong guesses or false positives), but the induced categories are too narrow and still miss many other decrees in the same labeled category (i.e. many false negatives). Again this indicates that unsupervised learning may be best for identifying low-level similarities between small groups of concrete terms or concepts, but less good at accurately guessing the full degree of conceptual abstraction that researchers choose to incorporate into higher-level categories.

While rule-based or supervised classifiers perform best at replicating human coding, a downside is that they can only learn categories that are known and sufficiently well defined by the researcher beforehand. The resulting data will preserve whatever biases were encoded into the training data. The same would be true of even unsupervised classifiers that were able to achieve a high level of accuracy against a hand-coded test set. While accuracy tests can evaluate the effectiveness of replacing human coding with an automated process that produces the same output, accuracy tests can't tell us whether the resulting data capture the most useful underlying concepts for the actual research objectives.

4.2 Usefulness in Distinguishing Empirical Trends

I've constructed a second form of evaluation to assess how well the resulting data capture anticipated relationships to other real-world variables. I take a set of factors that are theoretically expected to affect decrees as a whole, and look for differences in how they relate to the encoded decree categories, to justify breaking down the data in this way. In other words, if it is useful to study decrees based on their main actions and target entities, then we should see one type of decree increase under certain conditions, and another type of decree increase under other conditions, beyond the rise and fall of the total number of decrees overall.

To compare to the encoded decree categories, I have selected factors from the literature on presidential power, which can be thought of as the most likely established explanations for when we should see an increase in decrees. Leaders might issue more decrees if they face obstacles to enacting their policy agenda through the legislature, such as when the president's party controls less congressional seats. Leaders might also issue more decrees if they face urgent demands for policy action, such as during economic downturns. Finally, leaders might issue more decrees as their remaining time in office grows shorter, whether they're trying to win support for a preferred successor in the next election or to bolster their own legacy.

Table 5 defines the variables used to represent these factors in the subsequent tests. I also tested other related economic and political indicators, such as the rate of passage of executive-sponsored bills in the legislature, and the global prices of Peru's top export commodities. Those indicators showed very similar results, so for brevity, I present one indicator for each type of motivation here.

Table 5: Influencing Factors Expected to Correlate with Decrees

Variable	**Description**	**Motivation**	**Source**
Exec seats in congress	Share of seats in Congress controlled by president's party	Circumvent obstacles in legislature (decrees up when exec loses control of legis)	DPI[5]
Industrial production	Year-on-year % change in industrial production	Respond to urgent demands (decrees up when econ worse)	IADB[6]
Months left in office	Log months left until the president departed office	Finish agenda in time (decrees up as time runs out)	New[7]

4.2.1 Set-up of Bivariate Regressions

To prepare the data for comparison to other historic variables, I aggregate decrees by month for each of the action and target categories, then divide by the total decrees for that month, so that the dependent variable is the percentage of decrees issued in a given month that contain a particular type of action and/or target entity. The goal is to identify relationships between other factors expected to contribute to the use of decrees, and changing proportions of different types of policy decisions, to show that

[5] Database of Political Institutions (DPI), Inter-American Development Bank/World Bank

[6] Data Indicators, Inter-American Development Bank (IADB)

[7] Hand coded based on entry/exit dates in Wikipedia presidents list

the categories are useful break-downs of the data.

I then regress each decree category on each external factor, using ordinary least squares and clustering standard errors by presidential term, since there may be serial correlation between months within the same president's tenure. I do not combine multiple external factors into the same model, since they are correlated, and testing select combinations of factors can produce unstable results. The goal is to compare pairwise relationships between different influencing factors and different subsets of decrees, rather than test a comprehensive theoretical model and interpret the results in absolute terms. Some tests may capture spurious correlations, when important variables are omitted, but there should be even stronger relationships in the tests in which the more directly related variables are included.

4.2.2 Results for Decrees Categorized by Supervised Methods

Table 6 shows the coefficients and standard errors from the bivariate regressions using data from the rule-based system. There are many significant relationships between the percentage of decrees in certain categories and certain external factors, which appear across the versions of the dataset that were classified using supervised models. The rule-based data is shown here, since it had the highest accuracy for labeling the predefined action and target categories in section 4.1. These results offer support for using the project's coding scheme and automated classification methods, to produce measures of different types of decree decisions for use in substantive analysis.

The results suggest that in the face of legislative opposition (a low percentage of executive-controlled seats in congress is shown here), presidents tend to issue more decrees targeting *public* offices, while the percentage of decrees targeting *private* actors declines. The largest increases (i.e. largest negative coefficient) occur in decrees that *enable* public actors. There is also a significant negative relationship for decrees that

Table 6: Regression of Rule-Based Decree Categories on External Factors

		public targets		
	all acts	*enable*	*regulate*	*other*
exec seats in congress	−0.1425 **	−0.2162 ***	−0.0497 **	0.0861 **
	(0.061)	(0.053)	(0.025)	(0.041)
industrial production	−0.2274 **	−0.3160 ***	0.502	0.0374
	(0.089)	(0.077)	(0.037)	(0.060)
months left in office	0.1450 **	0.0358	−0.0405	0.1517 ***
	(0.061)	(0.055)	(0.025)	(0.041)

		private targets		
	all acts	*enable*	*regulate*	*other*
exec seats in congress	0.2633 ***	0.0633 *	−0.0301	0.2304 ***
	(0.056)	(0.038)	(0.046)	(0.058)
industrial production	0.1858 **	−0.1853 ***	0.1726 ***	0.1873 **
	(0.085)	(0.054)	(0.066)	(0.086)
months left in office	−0.1222 **	−0.0868 **	−0.1123 **	0.0489
	(0.059)	(0.038)	(0.046)	(0.060)

Note: Standard errors clustered by presidential term. * p <0.1; ** p <0.05; *** p <0.01.

regulate public actors, which are somewhat different from those regulating private actors, since internal regulations delineate administrative procedures and authorities. In other words, when presidents face legislative opposition or gridlock, they appear to issue more interally-focused power-changing decrees, possibly to empower their own executive offices to act on their own, or to weaken rival branches or agencies.

The relationships between different types of decrees and economic performance vary more with the type of action than with the decree's target entity. (For economic performance, percent change in industrial production is shown here; other economic indicators showed similar results.) Decrees that *enable* both public and private actors tend to go up when the economy *worsens*, which may indicate that those decrees are designed to stimulate the economy or provide countercyclical measures to alleviate hardship. In contrast, decrees that *regulate* private actors go up when the economy performs *better*. This also seems intuitive, that government leaders would be more likely to impose regulations (especially on private economic activity) when the economy is doing well and they can afford to impose limits to protect broader values or address other social goals.

Finally, with regard to the president's remaining time in office, more decrees target public actors early in the president's tenure (i.e. with more months left), while more decrees target private actors toward the end (i.e. with fewer months left). The story regarding specific types of decree actions is harder to interpret; it appears that presidents shift from everyday actions ("execute" and "other" categories) targeting the public sector early in their tenure, to focus more on enabling and regulating private actors in their final months in office. This might indicate that they're trying to secure a positive legacy, or benefit certain private organizations within which they hope to win a future position after they leave government.

4.2.3 Results for Unsupervised Classification

I ran the same bivariate regressions on decrees categorized using the unsupervised methods as well. Since these methods are much less accurate at replicating human coding, the goal here is not to compare the same pre-specified action or target categories to the external factors. Instead, unsupervised methods should offer value if they are able to learn *other* categories which were not previously known or specified, yet which may be useful to the research objectives.

Table 7 shows the results of the tests performed on clusters of main verbs' noun objects, along with a sample of the terms in each of the clusters shown. By looking for clusters with significant relationships to the external factors expected to influence decrees, then inspecting the terms that comprise those clusters, we may find more

Table 7: Regression of Clustered Decree Categories on External Factors

	clust 48		*clust 23*		*clust 152*		*clust 122*	
exec seats in congress	0.0001		−0.0029	**	−0.0022		0.0033	*
	(0.0002)		(0.0012)		(0.0014)		(0.0018)	
industrial production	−0.0002		0.0058	***	0.0036	*	−0.0012	
	(0.0003)		(0.0017)		(0.0020)		(0.0027)	
months left in office	−0.0007		0.0019		0.0007		0.0008	
	(0.0002)	***	(0.0012)		(0.0014)		(0.0018)	

Note: Standard errors clustered by presidential term. * p <0.1; ** p <0.05; *** p <0.01.

	Sample of Cluster Terms	Potential Label
clust 48	*beneficiario, victima, ciudadano, contribuyente, usuario, nino*	social welfare programs/services
clust 23	*tasa, renta, pension, tributo, impuesto, isc, aduanas*	Taxes, duties, and fees
clust 152	*sunarp, indecopi, conasev, licencia, auditoria, accidentes, tejido*	Consumer/market licensing, registration and regulation
clust 122	*callao, vilcabamba, tayacaja, chincheros, huanuco, pariahuanca*	Local cities and provinces

precisely which groups of related terms capture the trends of interest.

For instance, cluster 48 had the only significant negative relationship to the number of months left in a president's tenure. From the cluster's terms, it appears that presidents at the end of their mandate tend to target slightly more decrees at private beneficiaries or social welfare groups, potentially as "quick wins" to shore up their legacy or win support for a future role in civil society.

In contrast, cluster 23 appears to maintain a larger share of total decrees earlier in a president's tenure, though not quite significant, and also when the president controls fewer seats in the legislature. Both clusters 23 and 152 are positively correlated with economic performance. Cluster 23 refers to taxes and customs duties, which may be priority agenda items to reform early on, or when the country can afford adjustments. Cluster 152 contains the acronyms of Peruvian agencies responsible for regulating capital markets and intellectual property, along with terms for licensing, audits, and controlled goods. This offers more detailed evidence that executives seek to regulate private markets more when the economy is doing well.

Cluster 122 contains place names for provincial and local political units or communities. It is not always clear whether to treat these as public entities (as in a transfer of funds to a municipal government) or private communities (as in disaster relief to benefit the people of a particular city). Decrees targeting local entities represent a larger portion of total decrees when the executive has greater control of the legislature. While the relationship to economic performance is not significant, it is negative, which might suggest that presidents take greater unilateral action to address local needs or local initiatives during economic downturns.

The analysis in this section is exploratory, as are most uses of unsupervised learning in the social sciences. Open-ended clustering or topic modeling can help us identify unforeseen groups of policy actions or target entities that might be important for our substantive research questions, if we assess induced categories based on expected relationships to other real-world phenomena. That said, using *new* explanatory factors in this analysis and then using the same variables in hypothesis tests would be tautological, since we would be effectively selecting data measures that are engineered to have significant relationships in hypothesis tests. Instead, well-established control variables like those presented here are probably the best heuristics for identifying useful new measures, which might then be used to analyze *other* hypothesized relationships.

5 Discussion

In this paper, I investigate practical computational methods for measuring certain policy decisions recorded in executive decrees, to be used in a broader study about presidential power. I discuss the growing advantages of studying government action directly through public records, which are increasingly available in digital archives, and which report formally enacted decisions in consistent legal terminology. I also compare several established and emerging methods for classifying those documents, based on their main actions and target entities, in order to arrive at useful data for substantive analyses. Finally, I present two forms of evaluation of the resulting data: accuracy tests in relation to human coding, and a more open-ended analysis of how well the resulting data capture expected relationships to other real-world phenomena.

The evaluation lends support for the use of a rule-based classification system, when good preprocessing tools like dependency parsers and general lexical resources like WordNet are available, to facilitate the extraction of structured information with limited additional effort. A rule-based system may then be able to reach a reasonable level of accuracy with less time and effort than more popular approaches to supervised machine learning. The basic architecture for the rule-based system did need to be built for this project, while supervised machine learning can be done with off-the-shelf tools. But much of the same architecture was used to extract structured verb, noun, and hyperym features that were also used as inputs to the most accurate supervised classifier. The rules also had to be written by hand, but this was done in conjunction with defining the coding scheme and labeling the initial set of training documents, both of which needed to be done for supervised classification as well.

Rules are easier to write for the clearest and most common linguistic patterns than for more rare or subtle configurations of relevant terms. This means that it may be less costly to develop a rule-based system that can achieve a reasonable level of accuracy – in this case 80% – but becomes more costly if researchers seek accuracy above 90% or 95%. Supervised classifiers also require more labeled documents to reach a higher level of accuracy, but it may seem faster to assign a label to each document (once the coding scheme is set) than to write out more complex rules for trickier cases. However, this trade-off probably diminishes as researchers seek to label more sub-document elements even for supervised learning, and if they have already extracted the necessary sentence components for sufficiently complex rules.

There is still a great deal of researcher judgment and bias that goes into selecting a coding scheme for any rule-based or supervised classification process. To reduce those biases and be able to identify useful patterns in documented events that might not have been foreseen, unsupervised methods offer a great deal of potential. Yet the possible patterns those methods might extract are so numerous, with many of them seemingly meaningless, that more work must go into guiding and constraining unsupervised algorithms in order to yield data comparable to the types of concepts that humans choose to encode. The second type of evaluation in this paper offers a potential avenue to move in that direction. By identifying induced clusters that capture significant and substantively meaningful relationships to other phenomena, we might be able to inform the induction process to select categories that capture useful measures of real-world trends, without specifying exactly which ones we want.

There is still much work to be done to develop computational methods for extracting events from documents, especially in less supervised or less resource-intensive ways. Some projects may require more complex event schemas, involving multiple entities such as the source and destination of a transfer, the instrument or mechanism of a change, or the location or duration of the action. Knowledge-based engineering of structured processes, features and rules, in combination with appropriate forms of machine learning, as well as unsupervised frameworks in combination with some supervised heuristics of good induced concepts, may be the best ways to arrive at tools that perform well on complex tasks and yet can be more easily extended to new languages and domains.

References

Natalie Ahn, 2017. "Inducing Event Types and Roles in Reverse: Using Function to Discover Theme". "Proceedings of the ACL Workshop of Events and Stories in the News", .

Scott L. Althaus, Nathaniel Swigger, Svitlana Chernykh, David J. Henry, Sergio C. Wals, and Christopher Tiwald, 2011. "Assumed Transmission in Political Science: A Call for Bringing Description Back In". *Journal of Politics*, 73(4):1065–1080.

Carlo Biagioli, Enrico Francesconi, Andrea Passerini, and Claudia Soria, 2005. "Automatic semantics extraction in law documents". "Proceedings of the Tenth International Conference on Artificial Intelligence and Law", .

Doug Bond, Joe Bond, Churl Oh, J. Craig Jenkins, and Charles Lewis Taylor, 2003. "Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development". *Journal of Peace Research*, 40(6):733–745.

Francis Bond and Kyonghee Paik, 2012. "A survey of wordnets and their licenses". "Proceedings of the 6th Global WordNet Conference (GWC)", .

Stefanie Brüninghaus and Kevin D. Ashley, 2001. "Improving the representation of legal case texts with information extraction methods". "Proceedings of the 8th International Conference on Artificial Intelligence and Law", .

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof, 2011. "Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank". "Proceedings of the Fifth Law Workshop (Law V) of the Association for Computational Linguistics", .

Nathanael Chambers, 2013. "Event schema induction with a probabilistic entity-driven model". "Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing", .

Nathanael Chambers and Dan Jurafsky, 2011. "Template-based information extraction without the templates". "Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies", .

Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende, 2013. "Probabilistic frame induction". "Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", .

Catherine M. Conaghan, 1996. “A Deficit of Democratic Authenticity: Political Linkage and the Public in Andean Polities”. *Studies in Comparative International Development*, 31(3):32–55.

Catherine M. Conaghan, 2012. “Prosecuting Presidents: The Politics within Ecuador’s Corruption Cases”. *Journal of Latin American Studies*, 44(4):649–78.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning, 2006. “Generating Typed Dependency Parses from Phrase Structure Parses”. “Proceedings of the International Conference on Language Resources and Evaluation (LREC)”, .

Sean M. Gerrish and David M. Blei, 2012. “How They Vote: Issue-Adjusted Models of Legislative Behavior”. “Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)”, .

A. Gonzalez-Agirre, E. Laparra, and G. Rigau, 2012. “Multilingual central repository version 3.0”. “Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)”, .

Justin Grimmer, 2010. “A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases”. *Political Analysis*, 18:1–35.

Justin Grimmer and Brandon M. Stewart, 2013. “Text as data: The promise and pitfalls of automated content analysis methods for political texts”. *Political Analysis*, 21(3):267–97.

Daniel Hopkins and Gary King, 2010. “A Method of Automated Nonparametric Content Analysis for Social Science”. *American Journal of Political Science*, 54(1):229–47.

Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell, and Arun Vachher, 2003. “Information extraction from case law and retrieval of prior cases”. *Artificial Intelligence*, 150(1-2):239–290.

Hans Mathias Kepplinger, 2002. “Mediatization of Politics: Theory and Data”. *Journal of Communication*, 52(4):972–86.

Jin-Dong Kim, Tomoko Ohta, and Jun’ichi Tsujii, 2008. “Corpus annotation for mining biomedical events from literature”. *BMC Bioinformatics*, 9(10).

Gary King and Will Lowe, 57. “An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design”. *International Organization*, 3(617-42).

Dan Klein and Christopher D. Manning, 2003. “Accurate unlexicalized parsing”. “Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL)”, .

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky, 2014. “The Stanford CoreNLP natural language processing toolkit”. “Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations”, .

Teresa M. Miguel-Stearns, 2011. “The Digital Legal Landscape in South America: Government Transparency and Access to Information”. Technical report, Yale Law School – Lilian Goldman Law Library, Yale University.

Mercedes Garcia Montero, 2001. “La Década de Fujimori: Ascenso, Mantenimiento y Caída de un Líder Antipolítico”. *América Latina Hoy: Revista de Ciencias Sociales*, 29:49–86.

Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besancon, 2015. “Generative event schema induction with entity disambiguation”. “Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing”, .

David G. Ortiz, Daniel J. Myers, N. Eugene Walls, and Maria-Elena D. Diaz, 2005. “Where Do We Stand with Newspaper Data”. *Mobilization: An International Journal*, 10(3):397–419.

Anibal Perez-Linan, 2007. *Presidential Impeachment and the New Political Instability in Latin America*. Cambridge University Press.

Stephen Purpura and Dustin Hillard, 2006. “Automated classificiation of congressional legislation”. “Proceedings of the International Conference on Digital Government Research”, .

Clionadh Raleigh, Andrew Linke, Havard Hegre, and Joakim Karlsen, 2010. “Introducing ACLED: An Armed Conflict Location and Event Dataset”. 47, editor, “Journal of Peace Research”, volume 5, pages 651–60.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand, 2014. “Structural Topic Models for Open-Ended Survey Responses”. *American Journal of Political Science*, 58(4):1064–82.

Philip A. Schrodt, 2006. “Twenty years of the Kansas Event Data System project”. *The Political Methodologist, Newsletter of the Political Methodology Section, APSA*, 14(1):2–6.

Philip A. Schrodt, 2014. *TABARI: Text Analysis by Augmented Replacement Instructions.* Parus Analytical Systems, version 0.8.4b3 edition.

Lei Sha, Sujian Li, Baobao Chang, and Zhifang Sui, 2016. "Joint learning templates and slots for event schema induction". "In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", .

Hristo Tanev, Jakub Piskorski, and Martin Atkinson, 2008. "Real-Time News Event Extraction for Global Crisis Monitoring". "Proceedings of the 13th International Conference on Applications of Natural Language to Information Systems, NLDB", .

Matt Thomas, Bo Pang, and Lilliean Lee, 2006. "Get out the vote: determining support or opposition from congressional floor-debate transcripts". "Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)", .

Maarten Truyens and Patrick Van Eecke, 2014. "Legal aspects of text mining". "Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)", .

Arturo Valenzuela, 2004. "Latin American Presidencies Interrupted". *Journal of Democracy*, 15(4):5–19.

Nils Weidmann, 2015. "On the accuracy of media-based conflict event data". *Journal of Conflict Resolution*, 59(6):1129–49.

Claire Wright, 2014. "Executives and emergencies: presidential decrees of exception in Bolivia, Ecuador, and Peru". *Instituto de Estudios Latinoamericanos (IELAT), Universidad de Alcala*, 60.