Measuring Consolidation of Power:
Extracting Event Data from Laws and Decrees

Natalie Ahn
University of California, Berkeley
natalieahn@berkeley.edu

August 24, 2016

Abstract

The allocation of government authority is an important aspect of any political system. How decision-making bodies are structured and how power is assigned within them can have widespread consequences for the policies and programs those bodies carry out. Government authority structures can change dynamically, and the allocation of power is often the subject of vigorous debate. Leaders and opponents argue over whether more streamlined personal control leads to more rapid and effective responses to crises, or to greater abuses of power. Testing these assertions requires accurately measuring changes to government authority structures. Yet collecting data on institutional change within governing bodies may be more difficult than coding governments' external actions. More often, scholars measure government institutions in terms of static states, classifying them by regime types or major constitutional provisions that don't change very often. Event data is much more common in studies of interactions between societal actors, especially violent conflict. Yet the methods used for event extraction in other areas may be applicable to institutional change as well. This project presents a novel approach to measuring institutional change through automated event extraction from government documents. The event extraction process uses emerging natural language processing tools, especially dependency parsing, to facilitate the assignment of actions and attributes to events of interest.

This paper was prepared for the annual meeting of the American Political Science Association in Philadelphia, PA, September 1-4, 2016.

Introduction

When deciding how to measure social phenomena, we may consider two general approaches: 1) measure its current state, or 2) measure changes to that state. For instance, we can count how many legislators a congress or parliament has, in a given term, and compare that to the number of seats in other terms and/or in other legislatures. Or we can record whenever legislative seats are added or removed. The first approach implies an element of change, if we look at time series data in which the same legislature has different numbers of seats in different time periods. But the primary target of measurement is the total number of seats, as a snapshot at a chosen point in time. The second approach skips straight to change itself. The researcher never needs to know how many total seats there were, only that so many more were added or removed.

Certain research questions may lend themselves more to one form of measurement or the other. In studying political systems, we may distinguish between studies of leadership selection contests, studies of the government institutions that individuals are selected to lead, and studies of the policy decisions and external actions those governing bodies carry out. Leadership selection processes include democratic elections, authoritarian succession, and/or forceful take-overs. These processes constitute active changes in the leadership of a state; they're intuitively seen as events. Policy decisions also naturally lend themselves to event data; they're actions taken by leaders to carry out some program or operation in pursuit of a policy objective.

Government institutions, on the other hand, sound more like static states, things that just exist for a period of time. When it comes to organizational structures and the rules and practices that govern them, the predominant question seems to be what the rules *are*, i.e. whether a particular law is in effect or not. Political institutions form the foundation of a political system; they stand while leaders are appointed to run them and to make policy decisions based on them.

It seems natural, then, that scholars tend to study political institutions by measuring their states at regular intervals of time, rather than directly measuring institutional change.

Yet the organizational structures and rules of public decision-making may change in very dynamic ways. Leaders are not only empowered to make decisions within the bounds of existing rules, they also have the power to change the rules. A change to executive organizational structures may have much broader effects on many policy outcomes than an individual policy decision on its own. Government authority structures might also be easier to measure in terms of change rather than absolute characteristics. One doesn't have to measure the exact distribution of executive agencies and their assigned powers, to be able to say that a new agency was created, that more powers were delegated, that additional constraints were imposed. An act to concentrate more power might even have different causes and/or consequences than a static configuration of highly concentrated power that has persisted for some time.

There are challenges to taking an event data approach to institutional change, however, which may be harder than collecting and analyzing event data on elections or policy decisions. Leadership selection usually takes place at an identifiable point in time, with enough regularity for systematic comparisons yet infrequent enough so as not to overwhelm. Comparable election contests tend to be repeated, and the selection events and their results tend to be widely publicized, because the new leaders must win the recognition and acceptance of those they govern. Policy decisions are also often well publicized once enacted, because they involve external services, regulations, or uses of force. They demonstrate leaders' evidence of what they're doing for their supporters, to fulfill their mandate and maintain their claim to power. The inner workings of the governing apparatus behind those decisions, however, are more likely to be obscure: what offices and authorities were used to arrive at the policy output on display.

Yet technological advances and increasing government transparency provide new opportunities for studying the inner workings of governments in more fine-grained and dynamic ways. Governments in both developed and developing regions have increasingly made legal and policy documents publicly available in digital form, in response to citizen and foreign pressure for transparency (Relly and Sabharwal 2009). Advances in computer-assisted natural language processing (NLP) are also making it easier to extract information from such documents, turning narratives into event data for systematic analysis. Official government documents might be especially well suited to automated event extraction, even more so than commonly used sources like news media, since they tend to use consistent formats and precise, legal language.

This project proposes a novel approach to studying institutional change that utilizes emerging methods for event extraction from text, and applies those methods to government documents to identify changes in the organizational structures, rules and practices of public decision-making. In the rest of the paper, I review the current literature on measures of government authority structures (which are dominated by static indicators), areas where event data is more prevalent (primarily in studies of violent conflict), and established methods of event extraction. I then explain this project's methodology, working with open source dependency parsers and constructing an event extraction system to operate on parsed text. Finally, I describe the application of event extraction to a set of government documents from Peru and Ecuador, used to identify a set of key changes to government authority structures. I compare the event extractor's performance when used on output from several different parsers, parsing both original Spanish and machine translated English text, along with options for aggregating the different iterations' output, given the still emerging and imperfect nature of available tools.

Literature

Existing Measures of Government Power Structures

As mentioned above, studies of political institutions tend to measure their current state at regular intervals in time, rather than changes to their state in the form of actions or events. Major datasets of political authority structures classify them first and foremost by regime type, using categories of leadership selection (i.e. democracies versus autocracies, and various subtypes) (Geddes et al 2014; Marshall and Gurr 2014). Other common variables include whether the state has a presidential or parliamentary system, whether key branches of government exist, whether those branches are separately elected or appointed, and whether they have veto powers over each other (Przeworski et al 2000; Regan et al 2009). More dynamic variables usually have to do with election rules and elected offices, such as the use of proportional representation, existence of term limits, and/or whether the governing party has a majority in the legislature (Przeworski et al 2000; Beck et al 2001), rather than about changes in the decision-making apparatus of the state.

The focus on fundamental components of the state that are established in the constitution, and that rarely change, may be due to the fact that the internal characteristics of government organizations are very complex, and it would be hard to define comparable units of analysis unless one looks at essentially the entire government as a whole, or at least at the most common branches that tend to exist across states. Scholars do recognize that the formal rules written in a constitution may not reflect reality in practice. But the usual solution is to add expert judgments about the same variables, e.g. whether in practice the legislature does have veto power over the executive, or whether the judiciary is effectively independent (Marshall and Gurr 2014). These expert judgments appear to be based on personal assessments of a variety of events or actions taken within a certain period, but those events are not individually identified or quantified.

Event Data in Conflict Studies

In contrast, event data are more popular in the study of relationships and developments that happen outside the state, between multiple states or a state and other social actors. The most common social science applications for event data have been in studies of violent conflict. The focus of these efforts has shifted from an earlier emphasis on war between states (as in the Correlates of War Project, Singer and Small 1994; and WEIS, McClelland 1999), to broader event frameworks that include non-state actors and various forms of civil unrest (as in CAMEO, Gerner et al 2002; and IDEA, Bond et al 2003). Other topics of event data include economic and business developments (Hogenboom et al 2013; Shin et al 2014) and, most prominently outside the social sciences, biomedical research (Vanegas et al 2015; Miyao et al 2008).

Conflict and business events have certain advantages for event data collection. Since they occur between societal actors, they tend to be more visible than events that occur within one organization. Studies of conflict and business events tend to use news media reports, which are produced for public consumption, tend to be more readily accessible than government documents, and may cover at least a selection of events across countries in the same language (Rose et al 2002). Also importantly, there are strong motivations to understand, predict, and respond rapidly to events involving violent conflict or unrest. The U.S. defense and intelligence community has sponsored many of the major conflict event data efforts to date (O'Brien 2010).

Yet many of the methods used to collect and analyze event data on conflicts between political organizations can also be used to study changes in the institutional structure and functions within a political unit. These efforts may complement each other, since changes to internal authority structures may serve as precursors to changes in external government programs or operations. There may even be certain advantages to extracting institutional change events from official government records, as discussed further below.

Methods of Event Extraction

In recent years, there have been major advances in automated event extraction from text. The broader fields of text analysis and NLP are rapidly evolving, and there are many choices a researcher can make, depending on what types of information the researcher is looking for and the nature of the documents the researcher has to work with. Methods for processing natural language can be grouped into two broad approaches: statistical methods that focus on word frequency or co-occurrence (i.e. "bag-of-words" analysis) and language modeling methods that make use of both semantic and syntactic information, using grammatical patterns or parsing. Bag of words methods can create powerful tools for identifying topics and themes covered in text, as well as how those topics are characterized or emphasized. These tools have been used in studies of political actors' ideology, preferences, and sentiments (Grimmer and Stewart 2013). When additional information is available about the origins and authors of the documents, it may be possible to infer that a specific action or event has occurred as well.

Syntactic parsing involves much more information than analysis of word counts alone, and therefore requires more complicated algorithms, decision rules, and detailed annotation for training and testing data. Yet grammatical parsing may be necessary to answer certain types of questions. For instance, if researchers need to know whether "the President dismissed Parliament" or "Parliament dismissed the President," a bag-of-words approach would treat these two statements the same. Word order matters, as do common connector words like prepositions that a bag of words approach might throw out as uninformative without their immediate context.

Full parsing has been a major challenge for computer scientists and computational linguists. Until recently, social scientists who needed to identify who did what to whom have tended to rely on simpler forms of pattern matching, using manually compiled dictionaries and key phrase expressions (Hogenboom et al 2011). For instance, researchers would search a given text for statements that match the pattern "[a known entity] [attacked/assaulted/ambushed] [another known entity]." Patterns of this kind can get quite complex, and can identify a lot of events. This was the approach used by the KEDS project and its successor TABARI, one of the most prominent efforts in political event data collection (Schrodt 2006). Yet this approach can be labor intensive, since it requires researchers to construct rules to handle all relevant phrase patterns, along with dictionaries of all relevant action and entity terms for a particular domain.

A breakthrough has come in the emergence of open-source NLP tools that are now able to fully parse text, not only identifying named entities and parts of speech, but dependency relationships between terms. These parsers take care of much of the work involved in event extraction, by identifying verb or noun roots, their associated direct and indirect objects, and other modifiers and connectors. The task of extracting events becomes much simpler when applied to parsed text. The researcher still needs additional rules or classifiers to determine the semantic components of specific types of events. But instead of matching patterns applied to raw text, researchers can use dependency relationships and syntactic roles as features when determining event attributes. The latest successor to the KEDS/TABARI project, PETRARCH, appears to do exactly that, using the Stanford Parser and then somewhat simpler rules to extract events from the parsed text (Schrodt 2014).

Open source parsers are more widely available than even a few years ago. Yet parsers vary greatly in terms of their underlying algorithms, the domains they cover, and the extent and quality of training data used. There are also more resources available for parsing English-language text than foreign languages, although the latter are growing. The systems tested in the CoNLL-X shared tax on multilingual dependency parsing ranged in accuracy from 37.8% to 91.7% (Buchholz and Marsi 2006), showing wide variation across parsers developed for foreign languages. Some open source frameworks for syntactic parsing can be trained on new languages, and there are also some publicly available treebanks annotated with dependency relationships, so that each individual researcher does not have to create their own (McDonald et al 2013).

A final debate is whether to construct the event extractor with hand-written rules, or with machine learning. Event extraction is a more structured problem when conducted on parsed as opposed to raw text. An event extractor can traverse a parse tree and treat the assignment of each node to an event predicate or attribute role, as a classification problem based on various features from the node and surrounding tree. Both hand-written rules and machine learning classifiers can be designed to handle that type of event attribute assignment. According to Chiticariu et al (2013), machine learning approaches to event extraction are dominant in academic work, while rule-based approaches are more common in industry. Machine learning approaches may have greater potential to develop higher-performing and more generalizable tools in the long run, yet rule-based approaches tend to be more intuitive and can produce more immediate results. Vanegas et al (2015) note that event extraction in biomedical research has been moving from rule-based approaches to machine learning, due to the limited generalizability and lower recall of purely rule-based extraction. Yet they note that rule-based systems can still beat machine learning classifiers in some contexts, and are less reliant on large, high-quality training datasets.

Methodology

Parsers

For this project, I use two open source dependency parsers, testing a Spanish and an English model for each parser, and compare the events extracted across them. The parsers used are shown in Table 1:

Table 1. Parsers used for input to event extraction process

Framework	**English**	**Spanish**
Stanford Parser: lexicalized PCFG parser Version 3.6.0 (Klein and Manning 2003)	English-language model provided with typed dependencies (de Marneffe et al 2006)	Spanish-language model provided with head attachments; grammatical relations more loosely inferred from part of speech tags
Google's SyntaxNet: neural network framework for dependency parsing (Petrov 2016)	English-language model "Parsey McParseface" provided (Petrov 2016)	Spanish-language model trained with Spanish Universal Dependencies Treebank (McDonald et al 2013)

Both the Stanford Parser and Google's SyntaxNet come with pre-trained English-language dependency parsers. More work needed to be done to utilize Spanish-language implementations. For the Google framework, I trained a Spanish-language parser on SyntaxNet using the Spanish Treebank from the Universal Dependencies project (McDonald et al 2013). The Stanford Parser package currently comes with a lexicalized PCFG model for Spanish language text, which provides output in CoNLL format that includes head attachments, but not typed dependencies. (In other words, it indicates which head a node is attached to, but not the grammatical relation, e.g. whether this node is a direct or indirect object of its predicate head.) For now, I have inferred grammatical relations more loosely from the part of speech tags provided, since my event extractor is still able to distinguish different types of objects when using connecting terms to determine their semantic roles in a particular event.

The Stanford NLP website indicates that they plan to release a Spanish-language dependency parser (complete with typed dependencies) sometime in the near future. Stanford CoreNLP also offers a neural network dependency parser, which is currently trained for English, and which may be trained for other languages using CoNLL-formatted treebanks with dependency relationships. For now, however, the Stanford lexicalized PCFG parser and Google SyntaxNet parser provide a nice comparison, in the use of two different algorithms each configured for both languages. Even the incomplete nature of the Spanish-language dependencies inferred from the Stanford Parser is useful, since that implementation still produces better event data results than the English-language parsers when used on machine-translated text (as shown in the evaluation section below).

Event Extractor

The event extractor presented here is rule-based, although I plan to add a machine learning classifiers at a later stage. I chose a rule-based approach for the initial implementation, because of its intuitiveness and ability to generate immediate results, while the project is still in an exploratory stage of development. Hand-written rules are more human-readable and straight-forward to manipulate directly than decision-making algorithms built with machine learning. Working with an initial set of fairly simple rules made it easier to try out different approaches to processing the dependency parse tree, along with intermediate steps to organize or label the tree nodes and other data structures, without having to re-design machine learning classifiers or annotate whole new sets of training data as the process evolved. The resulting event extractor may not perform at the level of state-of-the-art machine learning classifiers, but it provides a framework that can be refined with the use of other algorithms for select components within it.

The process of turning parsed text into event data involves three steps. First, I convert the token-based relationships from a dependency parser into a somewhat simplified phrase-based parse tree. This step combines less relevant determiners (e.g. articles) and modifiers (e.g. adjectives) with the more important objects (i.e. the nouns) they support. The resulting tree distinguishes more clearly which phrases are the potential moving parts in the events described, and isolates the connector words between them (e.g. prepositions) that are important to determine what attribute role the subordinate phrases play in a given event.

Second, I assign entity labels to the phrases in the simplified parse tree. This step is similar to Named Entity Recognition (NER), typically used to label proper nouns as persons, organizations, places, or things (i.e. miscellaneous). The entity labeling used in this project is broader, covering common as well as proper nouns with the use of generic terms contained within each phrase. The entity labeler still uses a dictionary of known terms, but that dictionary doesn't include names of all of the government bodies in the Peruvian and Ecuadorian governments (e.g. "Ministry of Interior", "Ministry of Finance"), only the generic terms those names may contain (e.g. "ministry", "bank", "police"). The entity labeler also labels more categories than typical NER implementations, including references to legal documents, government functions, programs, and resources. The objective is to distinguish different types of objects that might play different roles in institutional change events.

The final step in the event extraction process is to assign phrases from the entity-labeled parse tree to event predicate and attribute roles. The current attribute assignment rules use the following features to identify events and their attributes: 1) the stem of a given predicate, 2) the dependency relation assigned by the parser to a particular constituent of that predicate, 3) the entity label assigned in step 2 to the constituent, and 4) the terms connecting the constituent to the predicate in the parse tree, usually prepositions immediate preceding each constituent (although more complex connector phrases are allowed).

Application

Document Sources

For this project, I use proposed legislative bills, enacted laws, and presidential decrees from Peru and Ecuador, from the end of their last military regimes (1980 and 1979, respectively) to the present (or the subset of those years in which each type of document is publicly available in digital form). These documents were obtained from public archives on government websites. In most cases, the sites provided at least searchable lists of the documents for the full time period, including the document's title and date proposed or promulgated, often with other metadata like the sponsor (of a bill) and its current status (whether passed or later repealed). Most archives also provide the full text of published laws and decrees, although in a variety of formats, many of them image files that are not yet machine readable. For the development of the event extraction process and evaluation of different methodological options, I am relying on titles and summaries, although I plan to use more of the full text in later stages of this project when testing hypotheses about changes to government authority structures.

As mentioned in the introduction, official government documents offer advantages over news media for extracting event data. News reporting is often redundant, so in order to accurately count specific events, the researcher must identify and remove reports that represent the same event. News reporting is also influenced by many factors other than whether events occurred, such as reporters' access to the actors involved, the fluctuating interest level of the publication's target audience, and potentially investigative or branding objectives of the publication's editors (Weidmann 2015).

The language used in news articles is also more informal than in official government documents. News reporting is designed to be interesting and informative; it tends to use simple, straightforward language for rapid comprehension, but the vocabulary and phrasal structure used may vary across reports of the same type of event. News articles are a secondary source of information, they do not themselves enact the changes they report. Official government documents have a greater burden to be exact, if dry and cumbersome to read through, in order to reduce opportunities to dispute their meaning. They are more likely to use formulaic repetitions of the same terms and linguistic patterns to describe legally precise types of events.

Event Definitions

Table 2 shows a set of key changes to government power structures, which are commonly enacted in laws and decrees in Peru and Ecuador. These events constitute institutional changes to the government's internal decision-making apparatus, rather than policy decisions about external services, operations, or regulations. The events in Table 2 focus on the existence and hierarchical position of major government offices, and the powers or capabilities assigned to those offices.

Many other attributes can be added, including the location and/or duration of the event. For this project, it is not essential to identify a time frame reported in the document itself, since the documents are primary sources and the dates of introduction and publication are included in their metadata. It is also not essential to assign a location to each event, because this project focuses on the allocation of power among central government institutions (and the country of the document's origin is also in the metadata). The entity labeling step does distinguish between central government and local government offices, and other attributes could easily be added, if a location or time period were important to a particular type of event. The predicate-attribute combinations in Table 2 are the most important to understanding the direction of each type of change in power allocation, i.e. which entity gained what. These combinations are the ones used to evaluate the performance of the event extractor, below.

Table 2. Event Types for Major Changes to Government Power Structures

Predicate	**Attribute (direct object)**	**Attribute (indirect object)**
create (verb or noun form)	**body** (organizational unit created)	**assigned-to** (organizational unit to which the new body is assigned)
transfer (verb or noun form)	**body** (organizational unit transferred)	**assigned-to** (organizational unit to which the new body is transferred)
dissolve (verb or noun form)	**body** (organizational unit dissolved)	**assigned-to** (organizational unit to which the dissolved body was assigned)
delegate (verb or noun form)	**capability** (function or program delegated)	**agent** (position or organizational unit to whom power is delegated)
authorize (verb or noun form)	**capability** (function or program authorized)	**agent** (position or organizational unit authorized to act)
restrict (verb or noun form)	**capability** (function or program restricted)	**agent** (position or organizational unit subject to restrictions)

Not only might other attributes be important, but other predicates might matter to the nature of a particular event as well. Many laws and decrees begin with a different root verb, such as “propose” or “introduce,” which facilitates the main action we’re interested in. It is also common for laws to “revise” or “modify” a previous action, or to “repeal” or “nullify” one. In the latter cases, we should not interpret a mention of a delegation of power to mean that a new power is being delegated, but that the power was previously delegated and is now being altered or withdrawn. Within the list of event types in Table 2, a common combination is to see a law that “authorizes the transfer” of some function or resource to a government unit. The event extractor treats these as nested events, so that a subordinate action is seen as an object of the parent action, and the parent action indicates the status of the subordinate act.

Evaluation

As mentioned above, this is an initial implementation, based on hand-written rules for a limited set of institutional change events. The accuracy scores shown below are decent but not remarkable. They suggest that a considerable amount of good events are being identified, but with considerable noise. More importantly, the performance metrics lack context, and it’s hard to say what values should be considered good or bad. This is a rule-based system in which the same researcher wrote the rules and hand-coded the test data. Event extraction projects are best compared in a shared task environment, in which a third party provides the same annotated training and testing data for each project to use in system development and evaluation.

The best use of the tests done here is instead to compare the different options I’ve tested against each other, i.e. to treat the scores as relative rather than absolute values. I provide measures of accuracy for the event extraction process conducted on the same text after using each of the four parsers. This includes comparisons between use of the Spanish-language parsers on the original text, and use of the English-language parsers on machine-translated text (using Google Translate through the web interface at translate.google.com). Finally, I test an ensemble method that combines the even output from all four parsers, weighting their output equally, and also weighting their output based on their accuracy in the individual parser implementations.

For general reference, the KEDS/TABARI system has reported 75% to 85% accuracy on Reuters sentences with the WEIS coding scheme, and a high end of 91.7% using optimized dictionaries and the PANDA project's more fine-grained coding scheme (Schrodt 2014). Other conflict event data projects have also reported accuracy scores up to the low 90s (King and Lowe 2003). Event extraction projects in biomedical research have reported F-scores that range widely, from around 50% to 85% on major shared tasks, and even lower for more difficult tasks like negation and speculation detection (Vanegas et al 2015). Miyao et al (2008) tested multiple syntactic parsers as components in an event extraction system, evaluating each parser based on the accuracy of the events extracted using it. They reported precision scores of around 48% to 53% and recall scores of around 55% to 67% across parsers and event extraction schemes.

To evaluate the event extraction process, I hand-coded a test set of 500 government documents randomly sampled from all three types of sources (bills, laws, and decrees) from both Peru and Ecuador, coding predicates and attributes from the six types of events in Table 2. I calculated precision and recall in terms of the number of event attributes assigned the correct entity label, similar to a "labeled attachment score" (LAS) used to evaluate dependency parsers (Buccholz and Marsi 2006). "Correct" refers to the hand-coded test set, which means the event extraction process is being evaluated in comparison to human coding, as opposed to some objective truth. Precision refers to the percentage of guesses that were correct (true positives / true positives + false positives), and recall refers to the percentage of correct assignments that the extractor guessed (true positives / true positives + false negatives). F-scores are calculated as (2 x precision x recall) / (precision + recall).

Individual Parser Implementations

Table 3 shows results for event extraction using each parser separately. Note that these aren't accuracy tests for the task of parsing itself, but for the overall event extraction process.

Table 3. Accuracy Tests on Event Extractor Using Output from Each Parser

Parser used in event extraction	Precision	Recall	F-score
Stanford Parser – Spanish (original)	65.59	55.93	60.52
SyntaxNet Parser – Spanish (original)	77.94	77.18	77.56
Stanford Parser – English (translated)	69.47	50.23	58.30
SyntaxNet Parser – English (translated)	69.13	51.39	58.96

As discussed above, the Spanish model for the Stanford Parser is not a complete dependency parser, only providing head attachments but not typed dependency relations. It is worth noting that even without precise typed dependencies, the Spanish-language parser still facilitates more accurate event extraction than the use of either English-language parser on machine-translated text, indicating that considerable information is lost in the translation step that is important to accurate event attribute assignment.

Ensemble Approach

I also tested an simple ensemble method, combining the event instances extracted from the output of all four parsers. The ensemble approach is a simple form of ensemble averaging (or combining of the individual output), since the event extractor is rule-based and doesn't facilitate an ensemble learning algorithm as the means of weighting or selecting among the different parser implementations. I compare events identified for each document, determine which instances refer to the same event, and tallies up how many parsers identified that same event in the same document. Three parameters that can be used to configure the process: 1) how closely event instances need to match in order to be considered the same event, 2) how to weight the event instances identified using each of the different parsers, and 3) how many instances are enough to include an event in the collective output. In this exercise, I hand-select different values for these parameters, though an ensemble learning approach could be used instead.

For the first parameter, the strictest approach would be to require event instances to be identical, in order to consider them as referring to the same event. A looser approach would be to require event instances to have at least one attribute label assignment in common and no assignments conflicting. An attribute label assignment would conflict if the same predicate and attribute were present in both instances, but with different attribute labels. For instance, if one parser identified an event "*create* body:*directorate*" and another parser identified "*create* body:*directorate* assigned-to:*cabinet ministry*" in the same document, the stricter approach would treat these as different events, while the looser approach would assume that they refer to the same event. On the other hand, if one parser identified an event "*create* body:*directorate* assigned-to:*cabinet ministry*" and the other identified an event "*create* body:*directorate* assigned-to:*local government*", these would be considered different events in both cases.

For the second and third parameters, we can imagine the different parser implementations of the event extraction process each casting votes for possible events. Events derived from each parser's output get a certain number of votes, and then a certain number of total votes are required for an event to be included in the collective output. Each parser's output might be given the same weight (i.e. the same number of votes) or different weights (e.g. based on the accuracy of each individual parser's output before aggregation). The minimum votes required to include an event might be as low as one instance identified from one parser's output, or as high as the same event appearing in the output from all parsers.

For the ensemble approach evaluated below, I use the less strict rule for matching instances of the same event, and then compare different parameter values for parser weighting and minimum vote thresholds. The less strict rule is likely to lead to higher precision and lower recall than the strict rule, because the stricter rule would create more slightly different copies of the same events, while the less strict rule combines more events into a best guess of the common event to which they refer. The other parameters can balance this out, since a low threshold for inclusion should lead to more events being added to the collective output. In general, the less strict option for event matching (treating more event instances as potential references to the same event) would seem to get the most out of the ensemble method, which should work best when different implementations of the event extraction process have enough overlapping output to differentiate events by how many votes they each received.

The third parameter most directly affects the trade-off between precision and recall. A lower threshold of required votes means more events will be added to the collective output, most likely including more wrong guesses (lower precision) but potentially more right ones as well (higher recall). A higher threshold of required votes means that fewer events will be added to the collective output, most likely representing the best ones among all implementations' output (higher precision), but potentially losing some correct events that were more difficult for some implementations to identify (lower recall).

Table 4. Accuracy Tests on Events Aggregated from all Parsers

	Minimum instances/weight required to include event	
Weighting of output from each parser	Low threshold for inclusion (one weak parser's output suffices)	High threshold for inclusion (need most or best parser's output)
Equal weight across all parsers' output	55.89 / 85.79 / 67.68	84.76 / 41.50 / 55.72
Proportional to F-scores for each parsers' output	58.75 / 83.89 / 69.12	68.59 / 67.00 / 67.79
Heavier weighting by rank of F-scores for each parser's output	55.88 / 85.79 / 67.68	75.61 / 79.19 / 77.36

We can see that when the parsers' output is weighted equally, a lower threshold for inclusion yields better overall F-scores in the collective output. But when the parsers' output is weighted more heavily in favor of the better performing individual parser implementations, it becomes better to assemble them with a higher threshold for inclusion. In the latter case, a threshold can be chosen that allows the better performing parsers' output to meet the minimum vote count on their own, while weaker implementations still require agreement from other parsers' output in order to be included in the combined set of events.

Given that this is a simple form of ensemble averaging with manually selected weights, even with very heavy weighting in favor of the best parser, the collective output does not quite reach the performance of the best parser's individual output alone. However, the ensemble may still be useful in that it enables the researcher to refine either the precision or recall of the event extraction process, albeit at the cost of the other. For instance, a research may be especially concerned about precision, in a situation in which she/he does not need to find *all* relevant events, but wants to be highly confident that those events identified are correct ones. In that case, even with a very good parser, the use of additional parsers could help validate the events identified using the first, to make sure that only very clearly recorded events are included. On the other hand, if a researcher needs to find more total events and can handle more noise along with them (for instance if she/he plans to hand-check them after), then the addition of other parsers (and a low enough threshold for inclusion of events) can help find more true events that even the best parser implementation missed, even though it is also likely to add more incorrect ones.

Discussion

Institutional changes can have profound consequences for broader policy outcomes, including governments' decisions to use force, because of the ways in which they enable or obstruct the use of government power. Institutional changes may foreshadow major new initiatives or redirections in external government actions. When government actors seek to change the organizational structures and rules of decision-making, they are essentially changing who is capable of doing what, and may be enabling themselves to achieve something with those powers that they were not able to do under the pre-existing configuration of authority.

This project shows that emerging methods of event extraction, using dependency parsing to structure and inform the assignment of predicates and attributes to events, can perform well in identifying institutional change events recorded in government documents. Even a set of fairly simple hand-written rules for predicate-attribute assignment can be used to extract a considerable amount of good event data, with the best parser implementation tested here accurately labeling over 77% of event attributes. Accuracy scores would probably go down with the inclusion of more complex event types, but could probably be improved with the use of machine learning classifiers for the entity labeling and attribute assignment steps. The event rules (or training data used to train a machine learning classifier) could be expanded to cover many other types of events and attributes, including in nested combinations.

Even after extracting comparable events with concrete attributes, institutional change events may seem less intuitive to analyze than more tangible external actions like violent conflict. Institutional changes lack obvious measures of scale or magnitude; it may be unclear how to aggregate even similar institutional changes, such as the creation vs. the reorganization or transfer of an executive agency. Event data about institutional changes may require further processing, in order to turn event counts about obscure rule changes into more intuitive variables for analysis. It may help to view changes to government institutions as signals of underlying latent variables, such as government leaders' decisions to consolidate power, which they then pursue through a variety of institutional channels. The next step in this project will be to estimate the underlying decision to personalize or depersonalize power, by applying a method of latent variable estimation to the institutional change events identified in this stage.

Future plans for this data include using the estimates of leaders' decisions to personalize power as the dependent variable in hypothesis tests about when and why leaders choose to do so. The larger project's main hypothesis is that leaders with shorter time horizons are more likely to seek to consolidate personal control, while leaders who expect to survive and retain other societal roles after departing office are more likely to accept constraints while in power. The same measures of changing government authority can also be used to test theories about the consequences of concentrating or personalizing power, about which there are a number of ongoing debates. For instance, these measures could be used to test whether increasingly concentrated personal power enables more rapid government responses to crises, promotes long-term stability, or facilitates abuses of power and corruption, leading to public backlashes or more volatile leadership change.

References

Beck, Thorsten, George Clarke, Alberto Groff, Philip Keefer, and Patrick Walsh. 2001. "New Tools in Comparative Political Economy: The Database of Political Institutions." *World Bank Economic Review* 15(1): 165-176.

Bond, Doug, 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(5): 733-745.

Buccholz, Sabine and Erwin Marsi. 2006. "CoNLL-X shared task on Multilingual Dependency Parsing." *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*. New York City: June 2006, 149-164.

Chiticariu, Laura, Yunyao Li, and Frederick R. Reiss. 2013. "Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems!" Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 827–832, Seattle, Washington, USA, 18-21 October 2013.

de Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. 2006. "Generating Typed Dependency Parses from Phrase Structure Parses." *Language Resources and Evaluation Conference (LREC) 2006.*

Geddes, Barbara, Joseph Wright, and Erica Frantz. 2014. "Autocratic Breakdown and Regime Transitions: A New Dataset." *Perspectives on Politics* 12(2): 313-31.

Gerner, Deborah J, Philip A. Schrodt, Rajaa Abu-Jabr, and Omur Yilmaz. 2002. "Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions." Paper prepared for delivery at the Annual Meeting of the International Studies Association, New Orleans, March 2002.

Grimmer, Justin 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.

Hogenboom, Alexander, Frederik Hogenboom, Flavius Frasincar, Kim Schouten, and Otto van der Meer. 2013. "Semantics-based information extraction for detecting economic events." *Multimedia Tools and Applications* 64(1): 27-52.

Hogenboom, Frederik, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. 2011. "An Overview of Event Extraction from Text." *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011)*. Vol. 779.

King, Gary and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57: 617-642.

Klein, Dan and Christopher D. Manning. 2003. “Accurate Unlexicalized Parsing.” *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Marshall, Monty G. and Ted Robert Gurr. 2014. "Polity IV Project: Political Regime Characteristics and Transitions, 1800-2013." *Center for Systemic Peace*.

McClelland, Charles. 1999. “World Event/Interaction Survey (WEIS) Project, 1966-1978.” Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

McDonald, Ryan, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Tackstrom, Claudia Bedini, Nuria Bertomeu Castello and Jungmee Lee. 2013. “Universal Dependency Annotation for Multilingual Parsing.” Proceedings of ACL 2013.

Miyao, Yusuke, Rune Sætre, Kenji Sagae, Takuya Matsuzaki, and Jun’ichi Tsujii. 2008. “Task-oriented Evaluation of Syntactic Parsers and Their Representations.” *Proceedings of ACL-08: HLT, Association for Computer Linguistics*. Columbus, Ohio, 46–54.

Nivre, J. and J. Nilsson. 2005. “Pseudo-Projective Dependency Parsing.” In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*: 99-106.

O’Brien, Sean P. 2010. “Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research.” International Studies Review 12: 87–104.

Petrov, Slav. 2016. “Announcing SyntaxNet: The World’s Most Accurate Parser Goes Open Source.” Google Research Blog, May 12, 2016. Distributed at https://github.com/tensorflow/models/tree/master/syntaxnet

Przeworski, Adam, Michael Alvarez, José Antonio Cheibub and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950-1990*. Cambridge: Cambridge University Press.

Raleigh, Clionadh, Andrew Linke, Havard Hegre, and Joakim Karlsen. 2010. “Introducing ACLED: An Armed Conflict Location and Event Dataset.” *Journal of Peace Research* 47(5): 651-660.

Regan, Patrick M., Richard W. Frank and David H. Clark. 2009. "New Datasets on Political Institutions and Elections, 1972-2005." *Conflict Management and Peace Science* 26(3): 286-304.

Relly, J.E. and M. Sabharwal. 2009. “Perceptions of transparency of government policymaking: A cross-national study.” *Government Information Quarterly* 26(1): 148–157

Rose, Tony, Mark Stevenson and Miles Whitehead. 2002. “The Reuters Corpus Volume 1 – from Yesterday’s News to Tomorrow’s Language Resources.” *Language Resources and Evaluation Conference (LREC) 2002*.

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

Schrodt, Philip A. 2014. “TABARI: Text Analysis by Augmented Replacement Instructions.” TABARI Manual, Version 0.8.4B3: August 30, 2014. Parus Analytical Systems.

Shin, Sungho, Young-Min Kim, Choong-Nyoung Seon, Seunggyun Hong, Sa-kwang Song, and Hanmin Jung. 2014. “Business Event Extraction System Based on SSVM.” SWCIB2014 workshop, collocated with JIST2014 conference, November, 2014.

Singer, David J. and Melvin Small. 1994. “Correlates of War Project: International and Civil War Data, 1816-1992.” Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

Vanegas, Jorge A., Sérgio Matos, Fabio González, and José L. Oliveira. 2015. “An Overview of Biomolecular Event Extraction from Scientific Documents.” *Computational and Mathematical Methods in Medicine*, 2015: 571381.

Weidmann, Nils B. 2015. “On the Accuracy of Media-based Conflict Event Data.” *Journal of Conflict Resolution* 59(6): 1129-49.