Unsupervised Extraction of Workplace Rights and Duties from Collective Bargaining Agreements

Elliott Ash
ETH Zurich
ashe@ethz.ch

Jeff Jacobs
Columbia University
jpj2122@columbia.edu

Bentley MacLeod
Columbia University
bentley.macleod@columbia.edu

Suresh Naidu
Columbia University
sn2430@columbia.edu

Dominik Stammbach
ETH Zurich
dominik.stammbach@gess.ethz.ch

***Abstract*—This paper describes an unsupervised legal document parser which performs a decomposition of labor union contracts into discrete assignments of rights and duties among agents of interest. We use insights from deontic logic applied to modal categories and other linguistic patterns to generate topic-specific measures of relative legal authority. We illustrate the consistency and efficiency of the pipeline by applying it to a large corpus of 35K contracts and validating the resulting outputs.**

***Index Terms*—Unsupervised Extraction, Information Mining, Legal Corpus Analysis**

I. Introduction

The language of legal contracts, commonly referred to as “legalese”, stands somewhere between informal natural language and the precise statements of a formal system of deontic logic. At the same time, the ambiguity that legalese ”inherits” from natural language can be a central factor in judicial decision-making – see, for example, *United States v. Premises Known as 171-02 Liberty Ave* (1989), where the verdict hinged upon whether an ”or” was read as exclusive or inclusive [1]. With the scale and scope of contracts expanding every day – including digital ”smart contracts” – it has become increasingly important to study the contents of contracts in a systematic and quantitative manner.

These issues are especially relevant in unionized workplaces, where collective bargaining agreements lay out the bundle of promises made between the company and its employees. These promises matter: for example, how workplace grievances are processed, or how legal disagreements are arbitrated, can make a big difference for the lives of workers and the profits of shareholders. Negotiation of these provisions takes time and expensive expert attention to contract text.

To provide computational support of such an analysis, we introduce a contract-parsing pipeline that automatically generates detailed statistics on the key dimensions of labor union contracts. Our system leverages natural language processing tools to decompose a contract into its constituent parts – rights and duties bearing on the respective parties to the contract. The statistics include counts of entitlements, permissions, obligations, and prohibitions, by party (worker, employer, union, manager) and by topic (produced with topic models and document embeddings).

Our hope is that the resulting outputs can help researchers and practitioners understand the balance of responsibilities between various parties within a workplace. For example, one could use the resulting statistics to quickly summarize the gains and losses from different proposals during collective bargaining. Or they could be used for automating grievance processing in contract arbitration. In labor economics, the pipeline could for example be used to measure non-wage aspects of contracts [2].

This report lays out the pipeline and provides some validation and illustration of its outputs. Section 2 briefly describes the law and economics underlying the terms used for parsing the corpus. Section 3 outlines the methods, while Section 4 reports statistical results from our pipeline as applied to a large historical corpus of Canadian contracts, presenting a set of stylized facts about what these contracts do. Section 5 zooms in on a particular case study (the 2005 automaker crisis) to show how the parser can inform understanding of particular collective bargaining events. Section 6 situates our approach in the literature, while Section 7 concludes.

II. Economic and Legal Background

Our approach is motivated by Herbert Simon’s [3] distinction between a sales contract and an employment contract. A sales contract between a worker and a firm would entail a specified wage in exchange for a specific output, where the characteristics of the output are clearly defined in advance. The contract consists of an *obligation* to provide the specified output. This setup describes a one-off fee-for-service contract, such as when a house cleaner who arrives to clean a house, supplies his/her own equipment, and hence would be self-employed for tax purposes.

Simon observes that in most employment situations one cannot anticipate the tasks that are required in advance. This problem is solved with an “employment contract” under which the worker agrees to supply units of *time*, rather than units of *output*. The firm, after observing the relevant events, assigns the workers to different tasks. The worker has an *obligation* to be at work for the specified hours, while the firm has *permission* to manage the worker and specify tasks.

Simon points out that this contract is not optimal. In particular, it cannot be optimal that the firm may ask the worker to carry out any conceivable task. If the permission to manage is unlimited, the employer might ask for too much, such as requiring the worker to use dangerous equipment, or to come to work even if they are ill.

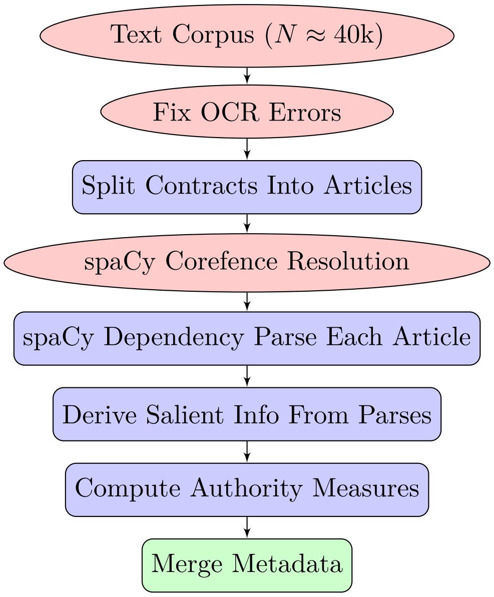

Text Corpus ($N \approx 40$k)
Fix OCR Errors
Split Contracts Into Articles
spaCy Corefence Resolution
spaCy Dependency Parse Each Article
Derive Salient Info From Parses
Compute Authority Measures
Merge Metadata

Fig. 1. Pipeline Overview

To solve this problem, the firm and worker can agree to limits on the scope of work in the contract. Such limits include *prohibitions* upon the firm from asking the worker to carry out certain tasks, such as using dangerous equipment. Similarly, the contract can specify *entitlements* to the workers, such as having paid sick leave.

These contracts are implemented using written legal documents. Because the particular sets of obligations, permissions, prohibitions, and entitlements are predicted to vary according to the work environment [4] [5], moving forward with a scientific analysis of contracts requires first that we measure these respective quantities. Thus, the goal of this paper is to make progress in characterizing how contract language is mapped to different types of economically and legally relevant statements.

III. Methods

Figure 1 presents a schematic of the pipeline's constituent steps. In this section we describe each step in detail.

A. *Contracts Corpus*

Throughout the development and evaluation of the pipeline, we relied on a novel corpus of over 35K Canadian collective bargaining agreements obtained from the Canadian Department of Employment and Social Development. In what follows we use the English-language versions of all contracts. We also collected a comprehensive set of numeric metadata variables linking the textual contracts with information on strikes, inflation, wage increases, political election outcomes, and so on. In future work, we plan to explore how our predicted measurements interact with the collected metadata.

B. *Optical Character Recognition*

Our contract corpus consists of documents stored as PDFs, a format which is not machine-readable. So, the first step consists of transforming the PDFs into TXT-files via OCR. We use three separate OCR engines: ABBBY FineReader, Adobe, and Tesseract. Agreement of the three engines in general is high. In case of disagreement, we choose the majority prediction of a word from the three engines. If all engines decide differently, we fall back to the prediction made by ABBBY FineReader, which has the lowest overall error rate.

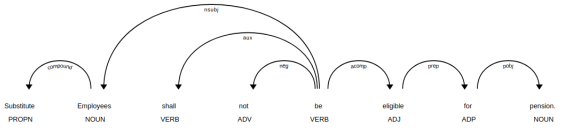
C. *Document Splitting*

Next, we split the resulting file into its corresponding sections. We use a custom-built splitter based on Regular Expressions (listed in Appendix A), We then apply the spaCy [6] sentence tokenizer to segment each section into a list of sentences. The Canadian Corpus consists of 980,909 contract sections (32.9 per contract) and 10.8 million sentences (11.06 per section).

D. *Dependency Parsing*

We use spaCy to perform dependency parsing. Besides being free and open-source, spaCy is widely used, fast, and accurate. We use the provided pre-trained English spaCy model, which is built on a bloom embedding strategy combined with subword features, followed by convolutional layers with residual connections, layer normalization, and a maxout nonlinearity. Among other things, spaCy can be used for tagging, parsing, and named entity recognition. We use its functionality to tokenize article sections into sentences, sentences into words, lemmatization, and dependency parsing of each sentence.

Fig. 2. Dependency Parse for Modal Verb Structures



The output is a parse tree, which represents the relations between words in a recursive hierarchical structure. Figure 2 shows such an example parse tree. The edge labels indicate the relation between two words in a sentence, e.g. *employees* is the subject of the sentence, the main verb is *be eligible* which is modified by an auxiliary modal verb *shall* and a negation *not*. The prepositional object which they are eligible for is *pension*. Our pipeline extracts clauses of the form *(subject, verb, object)*, so we would extract *(employees, shall not be, eligible for pension)* from this sentence.

E. *Coreference Resolution*

Our pipeline uses a dependency parser and therefore would not be able to map the legal relevance of statements where the subject is referred to by a pronoun. For example, consider the snippet: *The worker must ..., he then is entitled to ....* Without coreference resolution, we would only capture the first item as a worker obligation, but we could not attribute the second item to any identifiable role. Thus, we resolve coreferences and replace each subject being part of a coreference chain with the root of the first entry in the respective chain. We use huggingface's NeuralCoref 4.0 for this task, which integrates nicely in the spaCy NLP pipeline. In total, we extract 17'709'754 statements and resolve 1'353'704 coreferences (8% of all statements).

F. Derive Salient Info from Parses

The pipeline applies domain knowledge from the literature on deontic modality and forensic linguistics to create a set of deontic-logical rules mapping dependency trees into legal categories. We extract clauses of the form *(subject, verb, object)* from the dependency parses.

We start with deontic modal verb structures. Formally speaking, modality prescribes a favored action within a possible world [7]. In contracts, these statements create legal obligations and entitlements, featuring the modal verbs *shall, will, may, must, and can*. We distinguish between strict (*shall,will,must*) and permissive (*may,can*) modal verbs. Unlike general language, in contracts these verbs are always used to indicate modality, rather than for example to indicate contingency or future tense.

There are a few other important features of verb structures which we obtain from the dependency parse. Statements can be negative (*"shall not"* rather than *"shall"*), and active (*"shall hire"*) or passive (*"shall be hired"*). These annotations will be useful for determining the type of legal provision in any given statement.

TABLE I
"Special Verbs" indicative of rights and duties

Category	Verbs
Obligation	be required, be expected, be compelled, be obliged, be obligated, have to, ought to, agree, promise
Prohibition	be prohibited, be forbidden, be banned, be barred, be restricted, be proscribed
Permission	be allowed, be permitted, be authorized
Entitlement	have, receive, retain

Modal verbs are not the only way in which contracts impose obligations and give entitlements. Table I lists a small set of "special verbs", which are common in the contracts and delineate legal requirements without using a modal structure. We built this list from extensive inspection of the frequent verbs used in the corpus.

TABLE II
Pseudocode for Provision Type Assignment

Clause Type	**Patterns**
Obligation	strict modal + active verb + not negation not permissive modal + obligation verb + not negation
Prohibition	modal + not obligation verb + negation strict modal + prohibition verb + not negation permission verb + negation
Permission	permission verb + not negation permissive modal + not special verb + not negation prohibition verb + negation
Entitlement	entitlement verb + not negation strict modal + passive + not negation obligation verb + negation

Table II provides a full specification of the tags and rules underlying our clause-type assignment procedure. First, an *obligation* is characterized by one of two structures: 1) a non-negated strict modal followed by an active verb ('employer shall pay"), or 2) a non-negated non-permissive modal (either a non-modal or a strict modal) followed by an obligation verb ("employer is expected to"). *Prohibitions* are characterized by 1) a negated modal ("employer shall not"), a negated permission verb ("worker is not allowed"), or a non-negated obligation verb ("union shall be prohibited from"). *Permissions* are characterized by a 1) non-negated permission verb ("workers are allowed to"), 2) a non-negated permissive modal followed by a non-special verb ("workers may take"), or a 3) negated prohibition verb ("employer is not prohibited from"). Finally, *entitlements* are characterized by 1) a non-negated entitlement verb ("company retains the right"), 2) a non-negated strict modal followed by a passive verb ("employee shall be considered"), or 3) a negated obligation verb ("worker is not obligated to").

From detailed inspection of the semantic subjects in contract statements, we discovered that there are four central subject categories in labor contracts: *worker, union, employer, and manager*. Our pipeline maps agents of clauses heuristically to one of these categories via a dictionary of synonyms. In terms of the relative power in the workplace, we think of workers and unions on one side and employers and managers on the other. There are a handful of other actors mentioned in the contracts, such as arbitrators, who are not clearly in one of these camps. They are not especially relevant for understanding workplace authority and our pipeline excludes them from statistics.

G. Compute Authority Measures

We use the rule-based annotations of clauses to build measures of relative authority between the different types of agents. We understand obligations and prohibitions as reducing an agent's authority. We understand permissions and entitlements as expanding an agent's authority. Therefore, we can produce simple measures of authority for each agent group by counting the clause types – duties (obligations and prohibitions) on the decreasing side and rights (entitlements and permissions) on the increasing side.

The particular functional form one uses for authority would depend on the downstream application. For example, we could collapse down to two agent groups: (1) workers and unions, and (2) employers and managers. Relative worker authority could then be computed as *worker rights + employer duties - (worker duties + employer rights)*. One could normalize this measure based on total contract length, or use ratios instead. We analyze several statistics below and leave exploration of this problem to future work.

H. Section Topics

In addition, we would like to compute measures of worker/employer authority separately across the topics or types of clauses. One can imagine, for example, that some workers get more power over their office environment, while others get more control over their vacation schedule. We take two approaches to extract topic content.

a) Embedding of object branch content: To produce topic-specific authority measures, we start by extracting each statement's object clause: The tokens remaining after subject and verb have been removed. This material can be understood as including most information in the sentence besides what we have already extracted and used to compute authority measures across agent types. We then construct documents at the contract section level via concatenation of statement-level object clauses.

We embed these object-branch documents into a low-dimensional vector space using a topic-based dimensionality-reduction algorithm. We have experimented with Latent Dirichlet Allocation (LDA) and paragraph vectors (Doc2Vec). LDA results were more interpretable than Doc2Vec's, so the latter is not reported. LDA produced intuitive results with $K = 20$ topics, while interpretability declined steeply below $K = 20$. Redundancy across topics increased noticeably for values of K above 20.

b) Embedding of section headers: To complement this object-branch-based approach, we also derive topics for each section based on their title, when available. Taking the full text of each title and filtering out titular stopwords like "Section" and "Article", along with digits and Roman numerals, we extracted 550K section titles with 163K unique values. More detail on the process is reported in Appendix A.

We then compute two embedding vectors for each section title: the average of the GloVe embeddings produced by spaCy for each token in the title, and the contextual Sentence-BERT [8] (S-BERT) embedding of the title as a whole. S-BERT has led to more interpretable downstream topic labels, so results based on these are reported below.

To dimension-reduce the vectors, we applied K-means clustering to these embeddings. K-means clustering finds an optimal partition of the title embeddings into K "clusters", such that no other partition achieves a lower total "inertia" (within-cluster variance). Titles whose embeddings are in the same cluster are thus considered to be about the same topic. To automatically estimate the optimal number of clusters, we use the elbow method (see Appendix Figure 5).

An advantage of section-header clustering, relative to LDA, is that the resulting labels are intuitive by construction (as they use the contract drafter's provided summary language). A limitation is that some contracts do not have semantically meaningful or informative section headers. Contracts often label sections with a number or letter, for example, rather than providing a textual header. To unpack the topic content in sections without informative headers, we rely on the LDA approach.

IV. Results

A. Inspection of Representative Clauses

To provide some initial intuitive validation of the pipeline results, we inspect the most frequent extracted items. The key ingredients are subject, negation, modality, and main verb. These items are used to assign agent type and provision type: 47.2% of the statements in the corpus are assigned to a provision category, and 42.3% of these can be mapped to one of the four agent categories. In total, 20% of all statements in the corpus are assigned to a provision and agent category.

Table III lists the most frequently occurring subject-verb prefixes of all extracted clauses (absolute frequencies for each clause are reported in parentheses), grouped by agent type and provision category. The example clauses show at a granular level how our parser decomposes contract statements. They fit quite intuitively: ”employee is required” is the most common worker obligation, for example, while ”company agrees” is the most common firm obligation. Similarly, the most common worker entitlement is ”employee shall be paid”, while the most common union entitlement is ”union shall have”. Overall, the listed clauses all fit intuitively into the intended bins. Meanwhile, the frequent items in the ”other” categories (e.g., ”there will be”, ”employee is”) are not informative about rights and duties in the contracts.

B. Annotation of Most Frequent Tuples

To produce accuracy statistics, a trained lawyer manually inspected a sample of subject-verb tuples and annotated their accuracy. The annotator observed the tuples and associated machine labels directly and annotated each item as correct, probably correct, indeterminate, probably incorrect, or incorrect. An annotation of ”probably correct” was based on an assessment that this tuple structure could sometimes imply a different provision type. For example, ”company has” is coded as an entitlement by our system; the annotator judged that as ”probably correct” because one can imagine counterexamples, such as ”company has an obligation”.

The annotator coded the top 10-15 items for each class (agent-provision), for a total of 317 tuple annotations. Because the most frequent tuples were annotated, this procedure worked out effectively to the annotation of 3.8 million statements. This is about 19 percent of all statements extracted in the corpus.

Based on the annotations, we computed measures of system precision. A measure of ”baseline precision” is given by the share of correct and probably-correct statements. We also report ”lower-bound precision” where we only treat correct statements as true positives (that is, probably-correct is treated as a false positive). For both of these measures, we weight statements by their frequency in the corpus.

Our rule-based classification of subject-verb prefixes turns out to be quite precise. The baseline measure produces an overall baseline precision = .99, with lower-bound precision = 0.91. The respective values without frequency weighting are .96 and .84. For the 16 main agent-provision categories, the (weighted) precision is .99 (lower bound = .85). For the 9 categories containing ”other”, the precision is .99 (lower bound = .95).

We found it illuminating to inspect some of the errors generated by the system. For example, consider the statement ”employee shall not lose”. Our system codes this tuple as a

TABLE III
MOST FREQUENTLY OCCURRING SUBJECT-VERB PREFIXES

subject	obligations	prohibitions	permissions	entitlement	others
worker	employee is required (41789) employee shall be (21968) employees shall be (14350) employee will be (12868)	employee shall not lose (3578) employee shall not be (3517) employee will not be (2997) employees shall not be (1944)	employee may request (11120) employee may elect (9148) employee shall be allowed (7524) employee may be (5583)	employee shall be paid (61643) employee shall receive (57367) employee has (54772) employee shall be entitled (42847)	employee is (181457) employee works (42449) employees are (24868) employee is laid off (21863)
firm	company agrees (83488) employer agrees (76739) employer shall provide (19909) company will provide	board shall not be authorized (2397) company shall not be (1403) company will not be (1133) hospital shall not contract out (1062)	employer may require (4992) employer may grant (4307) company may require (2705) company may grant (1998)	company has (9725) board shall have (7767) employer has (7506) company shall have (6831)	employer recognizes (13744) company recognizes (13531) company is (9089) employer is (7880)
union	union agrees (46060) union shall notify (6113) member is required (3034) union shall advise (2893)	union will not cause (967) union will not engage (590) representatives shall not suffer (316) union will not discriminate (308)	representative may be (3452) union may refer (1983) union may submit (1785) union may request (1761)	union shall have (9463) union has (5231) member shall receive (4184) union shall be notify (3789)	union recognizes (15091) member is (12139) union is (10315) union acknowledges (10280)
manager	supervisor shall give (1278) management agrees (1272) manager shall give (1057) manager shall render (1012)	supervisors shall not perform (343) supervisors will not perform (284) management will not take (139) supervisors shall not work (91)	administrator may desire (566) director may grant (384) administrator may grant (377) manager may grant (345)	principal shall receive (973) administrator may have (808) principal shall be paid (516) principals shall be paid (507)	supervisor is (2126) management is fix (1375) management is vest (1334) principal is (1126)
other	there shall be (73307) parties agree (70143) there will be (33167) who is required (25927)	provisions shall apply (4494) leave shall not exceed (4242) release shall not be (3667) provision shall not apply (2760)	case may be (14213) which may arise (6131) which may be (6042) party may request (5266)	who has (36467) leave shall be granted (15557) leave will be granted (10311) who receives (9439)	who is (141114) there is (116098) it is understood (102328) it is (80531)

prohibition on workers. However, it should more intuitively be coded as a kind of protection for workers (if anything). In principle, we could build out additional vocabulary words based on these annotations. We leave this for future work.

This annotation scheme does not get at more subtle errors made by our parser, for example due to the presence of exterior framing clauses or contingencies. We could imagine sentences of the form: ”It is not the case that workers shall extend hours without overtime.” Our parser identifies ”workers shall extend” as an obligation, but the leading clause would negate this obligation. Annotating these types of errors is more subtle and often requires legal expertise as well as a more sophisticated parser. We therefore leave it for future work.

C. Summary Statistics on Rights and Duties

TABLE IV
FREQUENCY DISTRIBUTION: ROLES AND PROVISIONS

role	obligation	prohibition	permission	entitlement	total
worker	579K (16.2%)	83K (2.3%)	266K (7.4%)	1115K (31.2%)	2045 (57.1%)
firm	787K (22.0%)	46K (1.3%)	109K (3.1%)	90K (2.5%)	1033 (28.9%)
union	207K (5.8%)	17K (0.5%)	62K (1.8%)	130K (3.6%)	418 (11.7%)
manager	54K (1.5%)	3K (0.1%)	11K (0.3%)	16K (0.4%)	85 (2.4%)
total	1628 (45.5%)	151 (4.2%)	451 (12.6%)	1352 (37.7%)	3582K

In Table IV, we show the summed frequencies of the different authority measures across roles for the Canadian Union Contracts corpus. In parentheses, we show the percentage of a specific authority measure for a role given all computed counts.

This simple matrix provides a detailed view of the topology of union contracts. Starting with the row and column totals, we see that clauses are mostly about workers (57.1%), followed by the firm/employer (28.9%), and relatively few clauses about the union (11.7%) and managers (2.4%). In our set of parsed clauses, they mostly consist of obligations (45.5%) and entitlements (37.7%). There are relatively few permissions (12.6%) and even fewer prohibitions (4.2%).

Looking more closely, we see that the single most important set of clauses, comprising 31.2% of all clauses, is worker entitlements. The second most important category is firm obligations, at 22%. To the extent that firm obligations can be understood as a kind of worker entitlement (that is, the firm promises to do something for the workers), our parser shows that over half (53.2%) of the legally important content in contracts serves to empower workers. Worker permissions (7.4%) could also be added to this worker-rights ledger.

From a legal perspective, this emphasis on worker rights is intuitive because residual authority over workplace decisions is reserved to the employer. That is, in situations where the contract does not speak, the firm's preference (rather than the worker's) wins out legally. Therefore, if workers want a guarantee over some entitlement or authorization, that should be articulated in the contract.

On the other side, worker obligations are the next most substantial component of the union contracts (16.2%). This shows workers also make some long-term promises that the firm can rely on. The remaining items in the matrix are less important. Union obligations comprise 5.8%, the other items contribute all less than 5% of the share, adding up to 23.2% of the extracted clauses.

D. Analyzing Authority Across Topics

a) LDA applied to object branches: Having applied LDA to object branches, we identify 20 topics in the data. These topics, with most-associated words and some hand-generated labels, are reported in Appendix Table VI. The topics are mostly quite intuitive in comprising the relevant factors in the workplace, such as scheduling, amenities, insurance, benefits, maternity, pensions, strikes, safety, and harrassment.

To demonstrate the information that can be generated by the topic model, we show the authority measures across topics for workers and managers. Figure 3 reports these statistics, which provide some helpful intuition about the content of collective bargaining agreements. We can see that workers have more entitlements than obligations, while managers consistently have more obligations than entitlements, across all categories of clauses. The highest worker entitlements can be found in the topics concerning *holidays, parenthood, and vacation*, which are intuitively related to non-wage benefits. Conversely, the topics containing the least entitlements for workers are *strikes, rules, and harassment*, which are not related to benefits. Under

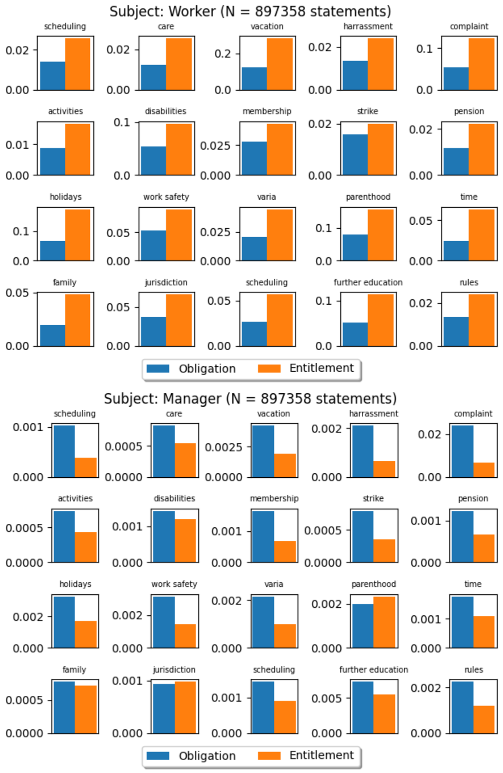


Fig. 3. Authority for workers/managers by topic

strikes, it is interesting that workers have many obligations, as employers would like to put limits on this potentially destructive activity.

b) Clustered embeddings of section headings: K-means clustering applied to embedded section headers produced intuitive and interesting clusters for dividing up clauses. The elbow method selected a cluster count of 40, which we used in experiments (Appendix Figure 5).

The resulting clusters, with lists of associated section header strings, frequencies, and inertia, are reported in Appendix Table VII. We found that relatively frequent clusters, and especially lower-inertia clusters, were more familiar and coherent in terms of their legal content. The clusters successfully group similar headings, e.g., about vacations (vacations; paid holidays; vacation; annual vacations; annual vacation), grievances (discrimination; grievances; grievance; disciplinary action; sexual harassment), and health/safety (health and safety; safety and health; health and welfare; safety education; occupational health and safety).

Table V reports some information on the top 24 clusters as ranked by a joint index based on frequency and inertia. They are labeled by the first (most frequent) associated section header string. Besides frequency (N), we also report the average number of net pro-worker entitlements (worker entitlements minus employer entitlements) and net pro-worker obligations (firm obligations minus worker obligations). The rows are sorted by the third column (pro-worker entitlements).

TABLE V
WORKER AUTHORITY METRICS BY CLUSTERED SECTION TITLES

label	N	proworker entitlement	proworker obligations
seniority	4208	0.76	-0.05
statutory holidays	2750	0.76	-0.04
recognition	6385	0.7	-0.05
benefits	3491	0.61	-0.07
employee benefits	3823	0.61	0.03
probationary period	3619	0.57	-0.02
leave of absence	4247	0.55	0.02
grievance procedure	5179	0.53	-0.02
training	7819	0.5	0.0
wages	5747	0.48	-0.02
overtime	1865	0.47	-0.05
severance pay	7525	0.44	0.05
leaves of absence	3498	0.38	0.04
general	2946	0.35	0.02
definitions	9342	0.32	0.02
arbitration	1983	0.3	-0.0
sick leave	5413	0.29	-0.02
union security	6829	0.28	0.0
term of agreement	5326	0.23	0.02
bulletin boards	8122	0.22	0.02
hours of work	3526	0.17	-0.0
working conditions	6675	0.11	0.0
vacations	4922	0.1	-0.01
management rights	2927	0.09	0.01

We see some intuitive patterns for the section-header topics and relative worker rights. In particular, the two topics with most entitlements are ”seniority” and ”holidays”, both contract sections where you would expect entitlements. Sections about ”benefits” rank 4th in terms of proworker entitlements. In turn, management rights refers to sections where firms reserve all unspecified rights; thus it is intuitive this clause scores low for worker authority.

V. Case Study

The usefulness of our metrics can be seen by looking at the history of contracts between the Canadian Auto Workers (CAW, later Unifor) and Chrysler, all of which are contained in our data. These contracts are important beyond this particular bargaining pair because the auto industry in Canada has ”pattern bargaining”, where contracts across Ford, General Motors, and Chrysler share most characteristics.

For the case study of our contract parser, we focus on the 2005 auto industry crisis. Post-WTO imports and high oil prices led to strong competitive pressures on North American automakers. In response, unions had to give up not just wages but also important benefits, resulting in a fall in the net entitlements favoring workers, with no offsetting increase in the other dimensions. While accounting datasets can easily document when unions are forced to negotiate somewhat smaller wage increases, documenting changes in benefits is more challenging because those are articulated in the contract text. Our methodology can show that the unions clearly wound up giving away a lot of benefits and workplace rights as well.

Figure 4 shows a set of worker-authority measures by contract start year from 1996 to 2012, together with the negotiated wage increases. Here, ”net” means the worker/union count, minus the employer/manager count. We can see that, prior to 2005, permissions, wage increases, and especially entitlements are quite high. However, in the new contract in 2005, there is a large decrease, especially in pro-worker entitlements. Meanwhile, obligations and constraints (prohibitions) on workers go up slightly. Interestingly, as the auto industry has recovered after the 2008 financial crisis, entitlements and

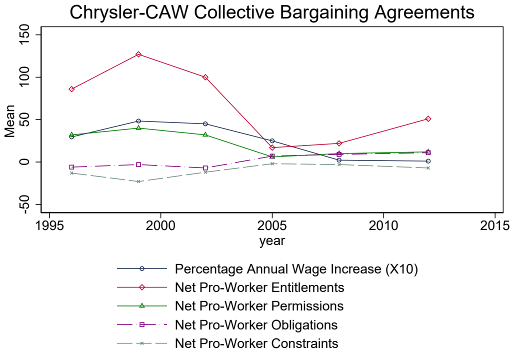


Fig. 4. Case Study: Chrysler-CAW

benefit increases have started to recover, even as base wage increases have not.

That 2005 was a critical watershed in union bargaining in auto can be seen from both press reports and the CAW's "bargaining reports", which summarize the contracts sent to the members. In 2002, CAW president Buzz Hargrove prefaced the bargaining report with: "This tentative agreement provides for significant but responsible improvements for our members in wages, benefits, job security, income security, time off the job, and a signing bonus. We've improved our benefits package in several important ways." In 2005, by comparison, the New York Times reported that the contracts were "the most modest ever signed by the C.A.W., " and Hargrove prefaced the 2005 bargaining report with: "totally unprecedented....there was 'no business as usual' in this round of bargaining". He concluded: "The companies started bargaining by demanding big concessions: like replacing wage increases with lump sums, abandoning COLA (even for pensioners), 10% co-pays on prescriptions, and giving up a week of paid time off per year."

The close tracking of our measures of authority with the history of collective bargaining outcomes suggests that a valuable use case for our methodology could be in summarizing the wins and losses in a collective bargaining outcome.

VI. Related Work

This work is not the first to extract rights and obligations from legal text. Related work can be divided into extracting such statements using rule-based approaches or using machine learning.

A. Rule-based Extraction

A number of systems have been proposed to automatically extract requirements, rights, and obligations from regulatory text. A potential advantage of contracts over corpora of government regulations is that there are many more independent documents, with no cross-document references. Rule-based approaches work to a certain extent for legal language because they are well-suited to the strict modal logic in legal language.

Early work shows that requirements are often phrased using imperative words, e.g., *shall, must, is required to, should, will, etc.* [9]. Implementing a combination of syntactic parsing and keyword-based rules can extract knowledge from regulatory text by analyzing a text's logical structure [10]. Rights and obligations can be extracted using contextual patterns and keywords, combined with object and phrase grammars [11]. [12] use a combination of WordNet information, syntax, and logic-based extractions to extract rules from legal documents.

Closer to our approach, [13] express rules found in regulation using conditional and deontic rules. They focus on agent and theme, deontic modals, main verbs, exception clauses, and conditional sentences. Using such rules, they report perfect precision and recall for obligations and permissions. Similarly, [14] implement a web-based tool where clauses containing an agent, action and modality are extracted from text. A user can then post-edit the extracted clauses. From these annotations, rules are automatically generated forming a model which can check unseen text for the user.

B. Machine Learning Extraction

Given the imprecision and complexity in language, deontic logic fails to accurately classify all statements. For example, *may* could refer to permission but could also imply possibility. Obligations could be easily framed to fool our parser – e.g., "workers are *asked* to do X". Machine learning approaches can address these challenges.

In this vein, [15] annotate sentences into *obligations, prohibitions, and permissions* and train neural networks to predict a held-out test set. They report an accuracy of 80% on that task. Similarly, [16] use hierarchical RNNs to automatically detect contractual obligations and prohibitions. Not surprisingly, they find that self-attention and hierarchical RNNs improve performance. [17] automatically extract features from a grammar tree, e.g., the modal verb, whether a negation is present, and the clause signal. Then an SVM is trained to classify a sentence into different norm types (authorization, prohibition) followed by heuristics to extract the different elements of a norm.

VII. Conclusions and Future Work

We have presented a legal contract analyzer, outlined the novel approaches developed for several steps of its pipeline, and showed its efficacy and consistency on validation tasks. In future work, we plan to integrate recent breakthroughs in legal text processing which have been obtained via state-of-the-art neural network architectures. In addition, to facilitate social-scientific study of how terms of contract between a union and a firm *evolve* over time, we are currently developing a contract evolution visualizer whereby users can trace the appearance, disappearance, and modification of various clauses of interest across a specified time series of contracts within the broader corpus.

References

[1] L. Solan, *The Language of Judges*, 1993. [Online]. Available: https://www.press.uchicago.edu/ucp/books/book/chicago/L/bo3684176.html

[2] L. Lagos, "Labor market institutions and the composition of firm compensation: Evidence from brazilian collective bargaining," 2019.

[3] H. A. Simon, "A formal theory of the employment relationship," *Econometrica*, vol. 19, pp. 293–305, July 1951.

[4] O. Hart, "Incomplete contracts and control," *American Economic Review*, vol. 107, no. 7, pp. 1731–1752, JUL 2017.

[5] B. Holmstrom, "Pay for Performance and Beyond," *American Economic Reveiw*, vol. 107, no. 7, pp. 1753–1777, JUL 2017.

[6] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[7] A. Kratzer, "Modality," *Semantik/Semantics. An International Handbook of Contemporary Research*, pp. 639–650, 1991.

[8] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019.

[9] W. M. Wilson, L. H. Rosenberg, and L. E. Hyatt, "Automated analysis of requirement specifications," in *Proceedings of the (19th) International Conference on Software Engineering*, 1997, pp. 161–171.

[10] B. Moulin and D. Rousseau, "Automated knowledge acquisition from regulatory texts," *IEEE Expert*, vol. 7, no. 5, pp. 27–35, 1992.

[11] N. Kiyavitskaya, N. Zeni, T. Breaux, A. Antón, J. Cordy, L. Mich, and J. Mylopoulos, "Automating the extraction of rights and obligations for regulatory compliance," vol. 5231, 01 2008, pp. 154–168.

[12] M. Dragoni, S. Villata, W. Rizzi, and G. Governatori, "Combining nlp approaches for rule extraction from legal documents," 01 2016.

[13] A. Wyner and W. Peters, "On rule extraction from regulations," *Frontiers in Artificial Intelligence and Applications*, vol. 235, 01 2011.

[14] J. J. Camilleri, M. R. Haghshenas, and G. Schneider, "A web-based tool for analysing normative documents in english," 2017.

[15] J. O'Neill, P. Buitelaar, C. Robin, and L. O'Brien, "Classifying sentential modality in legal language: A use case in financial regulations, acts and directives," 06 2017, pp. 159–169.

[16] I. Chalkidis, I. Androutsopoulos, and A. Michos, "Obligation and prohibition extraction using hierarchical RNNs," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 254–259. [Online]. Available: https://www.aclweb.org/anthology/P18-2041

[17] X. Gao and M. Singh, "Extracting normative relationships from business contracts," *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, vol. 1, pp. 101–108, 01 2014.

Appendix

Section Segmenter

The following regular expression was used for detecting article headers:

```
^.{0,3}(?:S ?E ?C ?T ?I? ?O? ?N? ? |
A ?R ?T ?I ?C ?L? ?E? |
A ?r ?t ?i ?c ?l? ?e? |
S ?e ?c ?t ?i? ?o? ?an?) (?: |-)
([0-9IVXl -]{1,4}|[0-9]{1,2}[ABCD])(.+)?$
```

(with Python `re` flags `re.M` and `re.UNICODE`)

The regular expressions for detecting the end of the body of a contract were simply disjunctions of phrases preceded by the ^ character (which asserts that they must appear at the beginning of a line).

These phrases were (manually) categorized based on how precisely they tended to demarcate the boundary between the contract body and the remainder of the document, for easy inclusion and exclusion during development. The "high precision" terms which tend to come immediately after the last article were { *in witness whereof, in witness, wereof, whereof, wereof* }. The "medium precision" terms which tend to indicate signature lines were { *signed, dated at, for the company, for the union, effective date, signatories of* }. The "low precision" terms indicated appendices and schedules which could serve as rough end markers if the previous two sets failed e.g. due to particularly bad OCR errors

```
appendix , schedule
```

Fig. 5. Number of clusters, elbow method

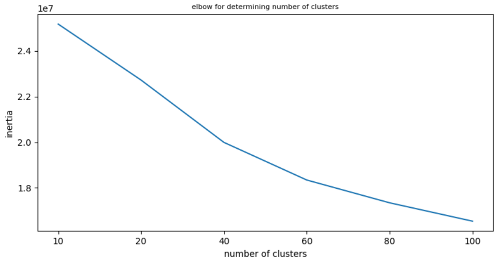


TABLE VI
LDA topics

topic words	topic label
tour teacher statutori preced weekli rest extra start meal fridai	scheduling
coverag care nurs function disciplin insur plant exclus direct locat	care
vacanc recal qualif abil plant abl skill displac address bump	vacation
harass discrimin sexual statu origin human membership complaint activ orient	harrassment
complaint file alleg violat disciplinari final settl disciplin hear interpret	complaint
activ differenti line membership plant interfer discrimin safeti premis pension	activities
disabl insur injuri percent worker safeti accid coverag certif physician	disabilities
membership contract rest recal sundai contractor monthli meal project fee	membership
strike lockout mean law picket spous monthli stoppag fee claim	strike
plant special cent district test pension start contract vacanc elect	pension
statutori sundai saturdai mondai lieu fall fridai rest straight start	holidays
safeti equip travel transport area vehicl driver site locat protect	work safety
vacanc abil area recal plant divis letter qualif minut offer	varia
parent retir pension child matern pregnanc adopt januari ag insur	parenthood
june contract juli march april septemb august renew preced joint	time
law occup death vacanc safeti father mother spous sister brother	family
court sever wit law juri elect death insur travel subpoena	jurisdiction
teach cours univers program experi academ educ trade level skill	further education
regul evalu rule disciplin contribut exclus direct fund function program	rules

Section Header Clustering

We pre-processed all section header titles by lowercasing them. We removed punctuation and all words containing numbers or roman numerals. Eventually, we removed all numbers, roman numerals and the following stopwords: *continued, cont, article, section, sec, appendix, note, rule, clause, append, table of content*. We excluded section titles where each word only contains one character (be it because of failed OCR, be it because a section is, for example, just named "A 5"). These rules combined filter out 263K section titles out of the total 555K section titles (47.4%).

To determine the optimal amount of clusters, we use the elbow method. We compute the overall inertia for all data points for a range of possible cluster counts. We can then plot the inertia as a function of the number of clusters used. The optimal value for that hyper-parameter is then located at the elbow of the graph. In our case, the number of clusters should be set to 40 (Figure 5).

TABLE VII
SECTION CLUSTERS

cluster labels	N	inertia
overtime; overtime d; overtime distribution; overtime and calls; overtime conversion	1865	4297
arbitration; arbitration procedure; expedited arbitration; arbitration d; procedure for arbitration	1983	6465
discrimination; grievances; grievance; disciplinary action; sexual harassment	1728	13556
hours of work; hours of work and overtime; working hours; hours of work d; hours of work and overtime d	3526	13904
health and safety; safety and health; health and welfare; safety education; occupational health and safety	2226	14756
no discrimination; no strikes or lockouts; no interruption of work; non discrimination; no strike no lockout	2060	15027
and following apply to full time nurses only; graduate nurse; and nurses who are transferred on a permanent basis may be; a v and and b applies to nurses only; there shall be no loss of seniority or service for a nurse during such	1651	15880
management rights; human rights; management; management functions; management s rights	2927	15886
seniority; seniority d; retention of seniority; seniority status; seniority and promotions	4208	15974
dental plan; medical examinations; has been complied with in giving such notice the hospital; medical examination; for the hospital	1724	17505
statutory holidays; holidays; general holidays; floating holidays; holiday work	2750	17817
general; general provisions; schedule d; correspondence; schedule a	2946	22190
a t u local no collective agreement; no work; where the nurse refuses the opportunity to; b will not apply to the ad hoc use of agency or registry; shall not apply	2934	23141
grievance procedure; compensation issues; grievance procedure d; grievance and arbitration procedure; adjustment of grievances	5179	23763
ontario nurses association; applies to employees covered by an ontario college under the regulated health; calgary police association; chrysler canada ltd; the edmonton police association collective agreement	2718	27466
maternity leave; custodial collective agreement september to august; seniority and service will accrue for an adoptive parent or a natural; party of the second part; maintenance collective agreement september august	2602	27550
technological change; promotions; promotions and staff changes; additional pay; transfer and promotion	3034	27623
leave of absence; bereavement leave; leave of absence d; parental leave; lay off and recall	4247	28769
leaves of absence; ill; leave; other leave; exclusions	3498	29079
loss of seniority; long term disability; bereavement; marine disaster; interruption of work	2858	29131
benefits; safety; fringe benefits; compassionate leave; effective	3491	29246
probationary period; layoff and recall; rest periods; probationary employees; probation	3619	30694
meals; part time employees; meal allowance; applies to full time and regular part time employees only; applies to full time employees only	2935	31676
employee benefits; for the company; employee status; for the employer; employee benefits d	3823	33859
vacations; paid holidays; vacation; annual vacations; annual vacation	4922	34317
term of agreement; collective agreement; memorandum of agreement; duration of agreement; of the collective agreement	5326	35596
days off and schedule of shifts; basic work week; annual; adjustment in hours; leaves of absence exceeds thirty continuous calendar days the	3684	38075
wages; rates of pay; salaries; salary schedule; wage scale	5747	41214
recognition; purpose; letter of understanding; discipline; agreement	6385	43831
year; party of the first part; time limit; xxv; chapter as amended	4140	45048
sick leave; strikes and lockouts; termination; discharge of non members; starting and stopping work	5413	48407
jury duty; application for membership; joint committee; board of trustees; general bargaining unit	4976	51577
union security; union representation; union recognition; union shop; union dues	6829	53678
shift differential; tools; transportation; protective clothing; c u p e local public works transportation	4706	54011
working conditions; job posting; job security; job postings; work standards	6675	60329
severance pay; pension plan; severance allowance; compensation; remuneration	7525	67371
training; qualifying conditions; preamble; check off; shall apply	7819	71590
definitions; step; representation; definition; of the second part	9342	75120
bulletin boards; contracting out; c u p e local collective agreement; iuoe contract; group insurance	8122	78938
cupe local; cta; start; wsw; scope	14021	107396

LDA Topic Model

In Table VI, we show the most important words for all topics after having applied LDA on the corpus. The topic label was determined manually after inspection of the topics.

Section Title Clusters

We clustered all section titles based on SBERT embeddings. In Table VII we show all resulting clusters. We find that the clusters are adequately described by the most frequently occurring section titles appearing in the cluster. Thus we simply show these section titles. The table reports the number of items in each cluster and overall cluster inertia – that is, the summed distance of all items from the cluster centroid. Lower inertia implies a more cohesive cluster, because items are, on average, closer to the centroid.