A Semantically Rich Framework for Knowledge Representation of Code of Federal Regulations

KARUNA PANDE JOSHI and SRISHTY SAHA, UMBC

Federal government agencies and organizations doing business with them have to adhere to the Code of Federal Regulations (CFR). The CFRs are currently available as large text documents that are not machine processable and so require extensive manual effort to parse and comprehend, especially when sections cross-reference topics spread across various titles. We have developed a novel framework to automatically extract knowledge from CFRs and represent it using a semantically rich knowledge graph. The framework captures knowledge in the form of key terms, rules, topic summaries, relationships between various terms, semantically similar terminologies, deontic expressions, and cross-referenced facts and rules. We built our framework using deep learning technologies like TensorFlow for word embeddings and text summarization, Gensim for topic modeling, and Semantic Web technologies for building the knowledge graph. In this article, we describe our framework in detail and present the results of our analysis of the Title 48 CFR knowledge base that we have built using this framework. Our framework and knowledge graph can be adopted by federal agencies and businesses to automate their internal processes that reference the CFR rules and policies.

CCS Concepts: • **Applied computing** → **Law**; • **Computing methodologies** → **Natural language processing**; **Knowledge representation and reasoning**;

Additional Key Words and Phrases: Deep learning, legal text analytics, compliance, Semantic Web

ACM Reference format:
Karuna pande Joshi and Srishty Saha. 2020. A Semantically Rich Framework for Knowledge Representation of Code of Federal Regulations. *Digit. Gov.: Res. Pract.* 1, 3, Article 21 (November 2020), 17 pages.
https://doi.org/10.1145/3425192

1 INTRODUCTION

As core documents of the Executive Branch of the U.S. government, the Code of Federal Regulations (CFR) [25] provides the public with a comprehensive publication vehicle for all of the regulations issued by federal agencies and the president; the documents are indispensable to the government's operations and communication [2]. These regulations are published in the Federal Register and must be adhered to by every organization that wants to do business with the U.S. government. They are also an effective mechanism to provide a central repository of all federal laws that can be queried by people to understand the laws described in them.

This research was partially supported by a Department of Defense supplement to NSF award 1439663: NSF I/UCRC Center for Hybrid Multicore Productivity Research (CHMPR).
Authors' addresses: K. P. Joshi, Information Systems Department, ITE 424, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250; email: karuna.joshi@umbc.edu; S. Saha, Computer Science and Electrical Engineering Department, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250; email: srishty1@umbc.edu.

2639-0175/2020/11-ART21
https://doi.org/10.1145/3425192

The CFRs are currently available in electronic form [25] on a variety of free and paywall sites. They are, however, long and complicated documents with various rules of many regulations written across multiple sections and titles. Its semi-structured organizational structure makes it a challenge for researchers to identify and cross-reference relevant sections that answer a specific question. Although simple keyword searches of CFRs are available, they often return large numbers of possible matches requiring further human review to identify the relevant and irrelevant responses. The organizational structure of the documents also makes it difficult to find and compare relevant provisions across sections and titles since the indexing of the information (through sectional tables of contents) is at a relatively high level within the regulatory sections. Hence, currently, understanding the various rules in CFRs requires legal expertise and is a time-consuming and labor-intensive process.

Research goal. The knowledge in regulatory documents, like CFRs, is usually present in the form of definitions, state-of-the-art or domain-specific jargon, rules, and cross references. Automating this embedded knowledge by making it machine processable will help government agencies, businesses, and legal experts significantly speed up their processes that refer to these regulations. Although knowledge extraction from legal documents has been an active area of research, there has been limited work on automatically extracting rules and policies from regulatory documents. Traditional techniques of Natural Language Processing (NLP) and those of information retrieval, like the bag of words model or vectorized model, alone cannot automate the analysis process of CFRs. These techniques fail to capture the semantic relationships between various legal elements spread across the deeply hierarchical structure of regulatory or policy documents. Dealing with heterogeneous legal facts and rules in a semi-structured format like XML is challenging in terms of answering user queries and performing analysis on the various legal elements. Hence, building ontologies, or knowledge graphs, for legal documents [38] is one of the possible efficient solutions to capture various facts and rules of legal documents to perform analytics and answer queries. This research is one of the first steps in representing knowledge in CFRs in the form of an ontology or knowledge graph to make it machine processable.

Research methodology. We have developed a novel framework to parse automatically and extract knowledge from legal documents and represent it using an ontology. The framework captures knowledge in the form of key terms, rules, topic summaries, relationships between various legal terms, semantically similar terminologies, deontic expressions, and cross-referenced legal facts and rules. We have built the framework using deep learning technologies like TensorFlow [31, 32] for word embeddings and text summarization [30], Gensim [33] for topic modeling, and Semantic Web technologies for building the knowledge graph. In our previous work [4], we presented preliminary results on extracting simple rules from the Federal Acquisition Regulations System (FARS) [26], which is Title 48 of the CFR.

In this article, we describe this framework in detail and present our results of applying the same to analyzing CFR documents. Section 2 includes a discussion of how this work is relevant to e-government. Section 3 covers related work in this area, and Section 4 describes our approach, developed using information retrieval, NLP, and deep learning techniques, for creating the CFR legal knowledge graph. Section 5 details the results when applying this framework to FARS (Title 48 of the CFR), and Section 6 describes the conclusion and future work.

2 DISCUSSION

Organizations or researchers interested in learning more about a particular federal law or regulation usually begin by searching the complex and large CFR corpus that is publicly available as semi-structured XML-formatted text [25]. The corpus has been organized into 50 different titles to correspond to the different areas where government services are provided. Each CFR title focuses on a particular subject area, such as agriculture or labor. Since each topic has a large number of rules and policies associated with it, the CFR titles are further organized into chapters, with every title having an average of 50 chapters. Each chapter has an average of 100 parts, and each part has various subparts, sections, and subsections that contain the individual rules. Some titles are brief, spanning only a single slim volume, whereas others can run as many as 20 volumes long. A citation to the CFR—for example, "25 CFR 531.1"—tells us first the title of the CFR in which a rule is located (in

this example, that is Title 25), then the section number within that title where the rule appears (here, that is Section 531.1, located in Part 531) [36].

Apart from the complex hierarchical structure of CFRs, the titles often cross reference rules and regulations listed in other titles. For instance, if a researcher wants to learn about "Rules and regulations for Technology Investment in Federal Agencies," the related sections are present in various chapters of Title 48 CFR (Federal Acquisition Regulations [FAR]) and Title 32 CFR (National Defense). So manually parsing these titles to get the answers, which is the current state of technology, is a labor-intensive and time-consuming process.

The publicly available XML/HTML format for CFRs might be easy for human consumption; however, it lacks semantic details and hence makes it challenging for machines to process. So it is currently impossible to write sophisticated programs to automate organizational processes that have to comply with the CFR rules. Hence, both the federal agencies that formulate these rules and vendors that have to adhere to them have to manually monitor and audit their processes and data to ensure the government rules are met. This can be very challenging if the policy covers a large population or requires urgent government intervention, like for disaster or pandemic management. Thus, there is an urgent need to develop a more semantically rich representation of the CFR corpus to improve the digital governance capability of federal organizations and regulation adherence of contractors and vendors working with them.

We have developed innovative approaches to transform legal documents from textual databases to machine-processable graph-based datasets using Semantic Web languages and techniques from deep learning and NLP. By building a flexible CFR knowledge graph (described in Section 4), we can transform them from their current XML/text format into graph data stores that will allow complex searching capability along with reasoning on rules. Our knowledge graph can store the key terms, rules, and deontic statements present in CFR titles, which can be easily queried and reasoned over to answer complex questions that a researcher may have about any rule or policy. Our proposed framework and its artifacts, like the CFR knowledge graph or deontic rules extractor, built using novel Artificial Intelligence (AI) technologies are available as open source in the public domain to be used by any interested organization using CFR in their processes. Our framework uses AI/Semantic Web languages along with techniques from topic modeling and Word2Vec algorithms that primarily depend on the CFR document corpus and not on user feedback. We have used the feedback of a legal expert to validate the knowledge graph and its instances. The strength of our design is that it does not require the end users to be trained on using the system; they can access it using a simple user interface like a search bar.

Another significant contribution of this framework is that it takes advantage of the flexibility of Semantic Web languages to allow easy modifications to the CFR knowledge base. The Federal Register [40] that is published daily lists the changes to the various CFR rules, proposed rules, and other federal notices. Our framework makes it possible to add or remove these changes to the knowledge graph easily. We contribute to automating and thereby improving eGovernment processes that currently depend on researchers discovering, parsing, and interpreting CFR rules and regulations.

Our framework will also facilitate automating answering complex questions that the general public may have about the various federal laws and regulations that affect their lives. This framework has the potential to help reduce the digital divide that is knowledge based (education based) and resource based. Some of the artifacts and toolsets that we have developed as part of this framework can assist people who have limited or no legal education, cannot afford expensive legal services, and have limited access to digital libraries. That will contribute to narrowing the gap between governments and ordinary citizens.

3 RELATED WORK

3.1 Legal Analytics

Regulatory documents are part of the legal text corpus. Various AI techniques, like predictive analytics, have been used primarily in five areas of the legal domain: discovery, legal search, document generation, brief and memoranda generation, and prediction of case outcomes [18].

Electronic discovery (also called *e-discovery* or *eDiscovery*) refers to any process in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a civil or criminal legal case. The volume of documents routinely subject to discovery poses challenges in investigations and litigation that extend beyond eDiscovery [19]. Although predictive coding is gaining increased acceptance as a procedure for identifying responsive documents with a lesser manual review, there is less appreciation of how document analytics can add value in answering document-related research questions, or otherwise help identify and analyze documents in ways not practical with keywords alone. Having reduced reliance on manual document review to decide which documents to produce, the challenge is to determine quickly what the documents reveal about the critical issues in the case.

Legal document analytics, unlike a manual review, enables algorithms to be run across all documents across multiple datasets and dictionaries at a relatively short time and cost. Although the results of computerized document classification may not be perfect, analyzing all documents collectively reveals patterns not visible from targeted manual review. Algorithms can be used to gather individual pieces of similar information of interest across an entire database, such as pricing information contained in contracts, providing a basis for economic analysis that would otherwise be far more cumbersome to perform. In the past, keyword searches were the dominant approach used for document analytics. However, they often return responsive documents with an overwhelming set of irrelevant documents. The mere formulation of a query or keywords is complicated if the information being targeted can be described in several different ways. Moreover, simple search queries may return ambiguous uses of the searched keywords. It may retrieve hits of the words that are not relevant to an inquiry. Keyword searches generally will not retrieve any documents containing a keyword that is misspelled, either in the query or in the documents.

In contrast to traditional keyword searching based on specific words or phrases, concept searching is a more sophisticated approach for document analytics that does not require the legal parties to agree on and identify all possible keywords of interest upfront. Predictive coding is a form of concept searching that can classify documents based on concept similarity, even if all of the target words do not exist in the document. Predictive coding and context searching have also been accepted by several courts [19]. Moreover, the results offered by the existing legal analytics approaches are still not semantically rich.

3.2 Semantic Web and Legal Ontologies

We have used Semantic Web technologies to develop our CFR knowledge graph or ontology and the reasoning component of our framework. Semantic Web tools enable data to be annotated with machine-understandable meta-data, allowing the automation of their retrieval and their usage in correct contexts. Semantic Web technologies include languages such as the Resource Description Framework [12] and Web Ontology Language (OWL) [13] for defining ontologies and describing meta-data using these ontologies, as well as SPARQL [14] query language for querying triples and tools for reasoning over these descriptions. These technologies can be used to provide semantic relationships between various legal elements of the CFR.

Our most fundamental requirement is for a knowledge representation that supports interoperability at both the syntactic and semantic levels. OWL has a well-defined semantics grounded in first-order logic and model theory, allowing programs to draw inferences with the assurance that the subsequent interpretation is sound. An important advantage for OWL over many other knowledge-representation systems is that it has well-defined subset profiles guaranteeing sound and complete reasoning with various levels of reasoning complexity and is designed to work with popular implementation technologies, such as OWL QL for databases and OWL RL for rule-based systems. A second design requirement is for a language that is designed to integrate well with the Web, which is often used by researchers working with CFRs. OWL is built on basic Web standards and protocols and is evolving to remain compatible with them.

Researchers working in the intersection of AI and law have developed conceptual models for knowledge representation and reasoning, known as legal ontologies, that have been widely used by legal practitioners, scholars, and laypeople in a variety of situations, such as simulating legal actions, semantic search, indexing, and to keep up to date with the continual change of laws and regulations [38]. Valente [44] proposed five main roles for legal ontologies: to organize and structure information, reasoning and problem solving, semantic indexing and search, semantics integration and interoperation, and understanding the domain. Our proposed ontology meets the roles of organizing, understanding, and reasoning the knowledge embedded within CFRs. There have been other efforts, especially in the European Union, on developing ontologies or knowledge graphs to capture the rules existing in various legal regulations and government laws [45, 46]. The European Legislation Identifier system is one main initiative that has created ontologies to provide a descriptive framework for structuring meta-data of legislative resources and to publish them as linked data [37]. There is currently no semantically rich representation of the CFR documents, and our work is one of the first works on this. In our previous work, we developed a semantically rich ontology for Service Level Agreements (SLAs) and privacy policies for cloud-based services [3, 22, 23].

3.3 Text Document Analytics

Information extraction from text documents has been an active area of research [7, 8]. Amato et al. [1] proposed a technique for building Resource Description Framework ontologies from semi-structured legal documents. They used only text extraction techniques and did not explored topic modeling as our framework does. Rusu et al. [5] used parse trees to generate triplets as subject-predicate-object. Etzioni et al. [6] used pattern learning to extracts facts from large documents in an unsupervised manner. Another important NLP technique used for information extraction from unstructured text is noun phrase extraction [9]. The use of automated techniques for extracting permissions and obligations from legal documents, such as text mining and semantic techniques, has been explored by researchers in the past [20, 21, 34]. In our previous works [10, 11], we also extracted key SLA definitions and measures using pattern-based rules using the Stanford Part of Speech (POS) Tagger [27] and CMU Link Parser [28] and used pattern-based rules for extracting permission and obligation [22–24]. As CFR titles are much longer and complex documents than SLAs or privacy policy documents, we significantly improved and refined our previous approach for developing this framework to capture various facts and rules spread across the documents.

Researchers have also proposed approaches for extracting phrases or sentences instead of just keywords from text documents. Le and Mikolov [15] recommended the paragraph vector algorithm for extracting sentences from documents. Goldstein et al. [16] proposed an approach to multi-document summarization that builds on single-document summarization methods by using additional, available information about the document set as a whole and the relationships between the documents. McCarty [17] illustrated that a statistical parser can handle complex syntactic constructions of an appellate court judge and that a semantic interpretation of the full text of a judicial opinion can be computed automatically from the output of the parser. We have built upon these approaches of phrase or sentence extraction when building our framework and automatically populating our system.

4 FRAMEWORK

In this section, we describe our framework in detail, including the techniques we used to create the knowledge graph, which captures the overall structure of CFRs along with key instances, extracting cross-referenced rules from the knowledge graph, and identifying and classifying rules into basic deontic expressions. Figure 1 illustrates the high-level overview of the framework. It consists of the following four main phases detailed in the following sections: data collection and preprocessing, building the CFR knowledge graph, CFR entities and relations population, and deontic rules extraction and population.

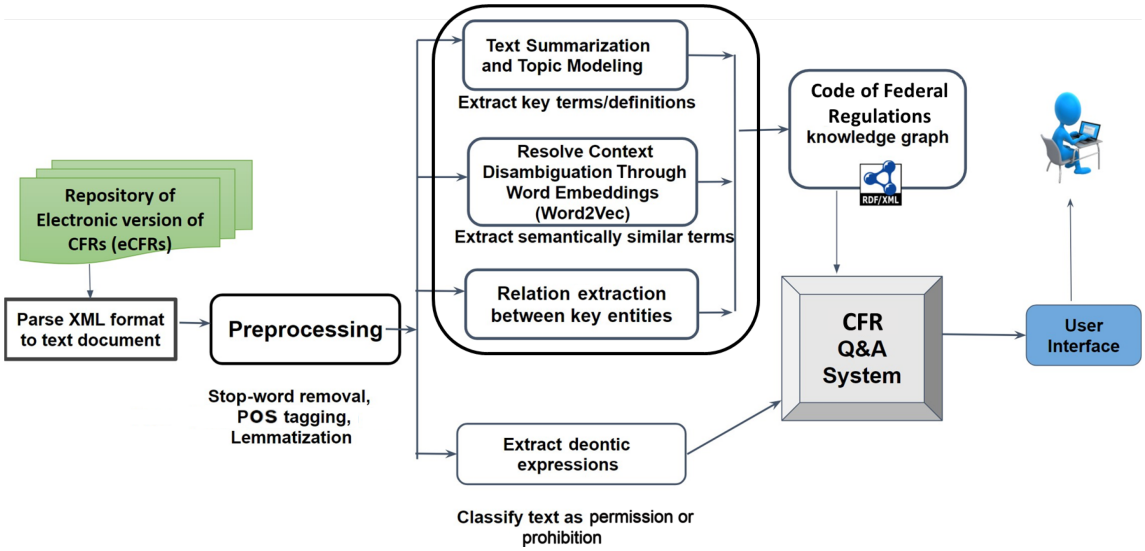

Fig. 1. Overall framework to build and populate the CFR knowledge graph.

4.1 Data Collection and Preprocessing

The online electronic version of the CFR is available in XML format in a hierarchical structure having tables and figures. We created a repository of all of the titles of CFRs. Using the ElementTree Python Library [29], we extracted the text portion from these documents. We observed that <Div> element contains the structure of XML, whereas <p> tags contain the text portion of the document. We preprocessed the extracted text using standard NLP techniques like conversion to lowercase, removal of stop-words, lemmatization [39], and POS tagging. For our analysis, we did not remove certain stop-words like "should" or "must" from the corpus, as these might semantically refer to words like "prohibition," "permission," or "authorization" rule, which could be useful in resolving the issue of context disambiguation. In addition, we did not remove alpha-numeric characters and numbers from the text, as they might represent the knowledge of the document organization. Thus, while extracting text from XML format, as well as preprocessing the text, we maintained the numbered hierarchical structure of the document.

4.2 Building the CFR Knowledge Graph

We created the knowledge graph (or ontology) by reviewing the overall document structure and the concepts, rules, and provenance embedded in the CFR titles. As stated before, CFR documents include 50 titles, and each title has on average around 50 chapters, which follow a hierarchical structure—chapters have subchapters, each subchapter has parts, each part has a subpart, each subpart has a section, and each section has subsection. In our study, we observed that end users often query for a specific section number in the CFR that they are interested in. Therefore, one of the key objectives while building the knowledge graph, in addition to capturing key terms and rules, was to capture the provenance of the rule, which was observed to be tied to the structure of the document. Hence, we considered this existing hierarchy in our design when identifying the main entities and relations of the CFR knowledge graph. This hierarchy also helps us efficiently query cross references in the regulations.

The main attributes captured include description, location on the document (chapter number, subchapter number, part number, section number, subsection number), and semantically similar words of entities. Figure 2 illustrates the overall structure of our CFR knowledge graph. The highest-level class is the *CFR* class, which has subclasses *Title, Chapter, Section*, and *Subsection*. Each *Subsection* includes the *SubSection Entity List* class whose

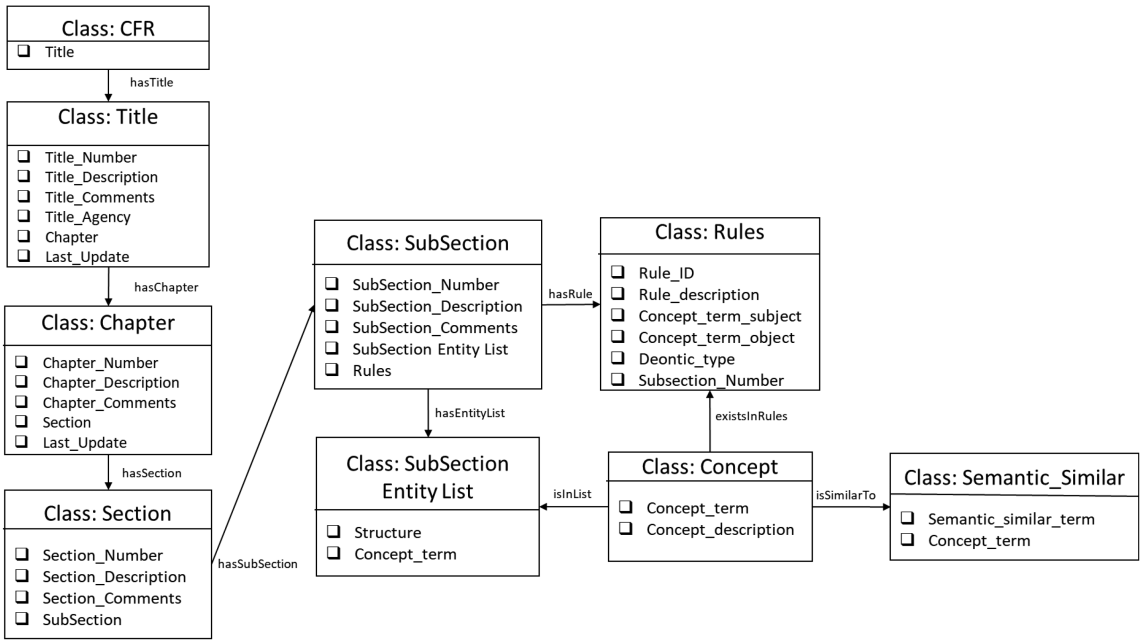

Fig. 2. Knowledge graph for the CFR.

data property is *Structure* and object property is *Concept_term*. The *Structure* stores the provenance of the *Concept_term*. Since a Concept term may occur in multiple locations in the document at different levels of hierarchy, the *structure, concept* combination allows us to query faster for a key term occurring in different sections. The *SubSection* class also contains subclass *Rules* that captures the rules and policy statements existing in the CFR subsection. Our system also classifies the rules into the four deontic types detailed in Section 4.4 and stores the deontic_type as a property of the Rules class. The class *Concept* stores concepts and descriptions as its instances. *Concept* is a superclass that is capturing an idea in the subsection (the arrow indicates the hierarchy). The same concept may be included in the various subsections or rules, and so the class *Concept* is referenced in classes *Rules* and *SubSection Entity List*. The class *Concept* has subclass *Semantic Similar*, which captures terms that are semantically similar to a *Concept* class's instance. Since the design follows a hierarchical approach, the *Title* is the superclass of *Chapter*, which is the superclass of *Section* and so forth. This design allows us to easily reason across the classes using SPARQL [14] commands. That is the reason for connecting individual entities and rules at the lowest hierarchy *subsection* of the CFR document. We created this knowledge graph in OWL [13] by using the open source protégé tool [41].

Validation of the knowledge graph. Since the CFRs are lengthy and very complicated, identification and validation of the main entities and relations for building the knowledge graph was a challenge, as there are no existing ontologies, like DBpedia or Freebase, for legal documents that we can use as ground truth. We validated this knowledge graph with the help of our legal expert collaborator, Ms. Renee Frank.

4.3 CFR Knowledge Graph Instances and Relations Population

We used the following three-step approach to populate the knowledge graph automatically. These steps are described in detail in the following sections:

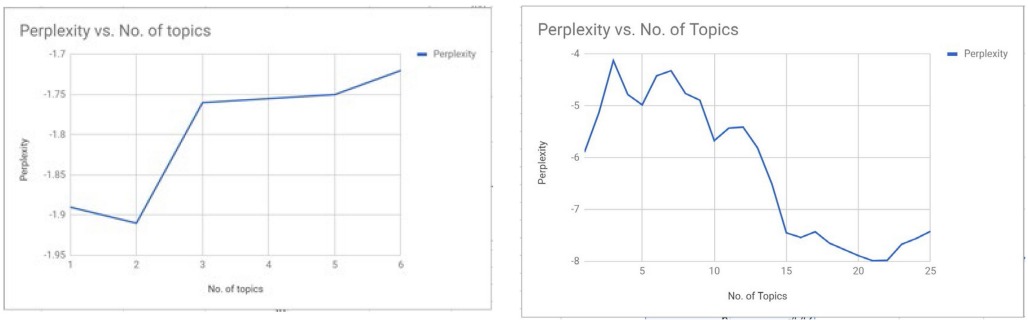

Fig. 3. Perplexity curves for subsections having 1 (left) and 51 (right) sentences.

—Extraction and validation of the main entity instances
—Extraction of semantically similar terminologies and population
—Extraction of relations between the main entities.

4.3.1 Extraction and validation of knowledge graph instances. The titles of CFRs have various chapters, parts, subparts, sections, and subsections of a varied number of sentences representing facts and rules. To extract the instances for our ontology, we used the three steps discussed next.

Step 1: Extractive text summarization of each paragraph of the subsections. Each CFR title document contains a varied number of sentences in a paragraph. For instance, Title 48 of the CFR contains paragraphs ranging from 4 to 200 sentences. We used text summarization to capture only vital information from the lengthy paragraphs of sections and subsections of a title of the CFR. We ran the TensorFlow extractive text summarization model [30] on each paragraph of the whole document. This model considered both words and word phrases to create the summary, as that does not lead to information loss.

Step 2: Extraction of topics from summarized subsections. After summarizing the text, we implemented the Latent Dirichlet Allocation [33] Topic Model on the summarized text to extract top-k topics for each section. For example, if a summarized section A has 10 sentences, then 5 topics were extracted, whereas another summarized section B having 50 sentences had 15 extracted topics. We used the perplexity measure to determine the number of topics extracted from the subsection. Perplexity is a statistical measure determining how well a probability model predicts a sample. The lower the perplexity, the better is the topic model. For example, in Chapter 1, the lowest number and highest number of sentences in a subsection were 1 and 51, respectively. We chose to extract 2 and 21 topics from 1 and 51 sentences, respectively. We used the lowest perplexity to determine the number of topics. Figure 3 illustrates the perplexity curves for the number of sentences 1 and 51, respectively.

We used the help of a legal expert and a legal dictionary [35] to validate the topics, which would be the main entities in our knowledge graph. After identifying the list of the entities, our toolset extracts definitions or descriptions of those entities from the document. In our previous work [22, 23], we developed a topic descriptor system for extracting definitions using pattern-based rules like Stanford POS Tagger [27] and CMU Link Parser [28]. We used the same system to extract the description of entities from each title of the CFR. For each entity, we also extracted the associated section number, part number, chapter number, and the title number (i.e., the location where the entity occurs in the document). This helped us capture the provenance of the rule, description, and location of each of the entities that will form a set of attributes for each entity of our CFR knowledge graph.

Algorithm 1 details the approach to extract the best number of topics from each subsection of a title of the CFR.

ALGORITHM 1: Entity Extraction

```
Let Topics be the hashmap where keys of the hashmap store the location of the paragraph and the values of
hashmap store array of topics extracted from Latent Dirichlet Allocation Topic Modeling.
Let Min represents minimum perplexity
Let N be a Hashmap where key-value pairs are perplexity values and corresponding topic models.
1.    input Preprocessed raw text
2.      For Chapter number in each Chapter within Title:
3.      For Part number in each Part within Chapter:
4.        For Subpart number in each Subpart within Part:
5.          For Section number in each Section within Subpart:
6.            For Subsection number in Subsection within section:
7.              For each_topic in set of topic models:
8.                Min = minimum value of perplexity of each topic
9.            Topics [Subsection number] = Min
Output: Hashmap Topics will store topics with lower perplexity as value for each subsection. This results in
fetching relevant topics from each subsection.
```

Step 3: Extracting instances and their description from selected topics. After extracting the topics from each subsection, we next identified the main instances of our knowledge graph from these topics. As each topic consists of a set of key terms arranged in the decreasing order of probability of likelihood, we needed to extract relevant key terms from each topic. After choosing relevant key terms for each topic, we assigned a topic label from those key terms. We next extracted definitions of the key entities from the document. The assignment of the topic label was validated through our legal expert. We validated the topics identified as essential entities for our knowledge graph with the help of a legal dictionary [35] and our legal expert.

4.3.2 Extraction of Semantically Similar Terminologies and Ontology Populations. In the CFR, various chapters are related to each other, and for a novice user, it becomes challenging to co-relate semantically similar terminologies found across various chapters. For example, semantically similar meanings of word "publication" are found across various chapters of the CFR as "findings" or "document." To resolve context disambiguation, we used TensorFlow Word2Vec deep learning architecture [31, 32] to generate a word embedding model for capturing semantically similar words. This model is essentially a neural network architecture utilizing a continuous bag-of-words model or skip-gram model to predict similar words. We configured parameters like batch size, number of skips, and skip-window to build and train the skip-gram model. The skip-window represents the number of words to be considered at the left and right of the target word. Num-of-skips represents the number of output words that will be picked in the span of a single word in an (input, output) tuples. We used the set of target words that we are interested in to evaluate the similarity on every specific step; the model is evaluated by looking at the most related words of those target words. In this process, in every epoch of the neural network while training, we find the probability $P(w)$ of target word w being "compatible" or "semantically similar" to other words in the raw text. We define V to describe the set of words in the skip-window used to predict semantically similar words of target word w, and K is the size of V. Unlike traditional deep neural network architecture where activation functions being used are usually *tanh* or *sigmoid* functions, we have used the *softmax* function [42] as an activation function in the hidden and fully connected layer of our deep learning neural network architecture. The probability $P(w)$ is calculated in a fully connected layer of the deep neural network architecture after every epoch.

$$P(w) = \frac{exp^{w}}{\sum_{i=1}^{k} \exp(\mathrm{w_i})} \tag{1}$$

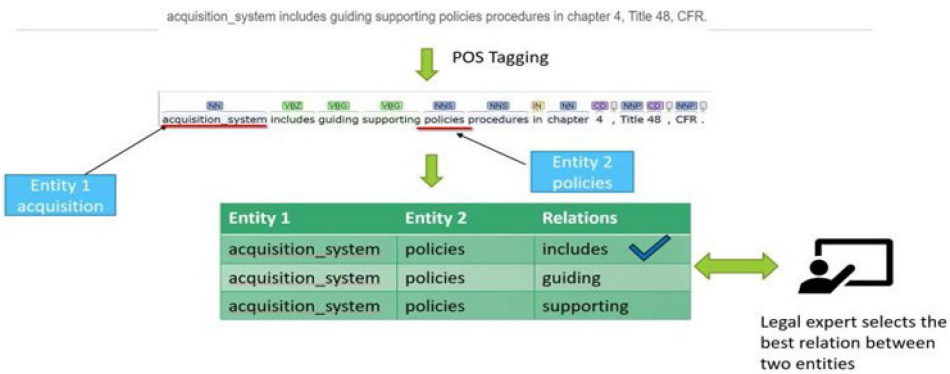

Fig. 4. Extracting the best relation between entities: acquisition system and policies.

Equation (1) describes the probability calculation using the *softmax* function. To maximize the likelihood of probability $P(w)$, we apply a logarithmic function to $P(w)$. Equation (2) describes the maximization of probability $P(w)$.

$$Maximized\ P(w) \ = \log \frac{exp^{w}}{\sum_{i=1}^{k} \exp(\mathrm{w_i})} \tag{2}$$

The program stops after 100,000 steps, and the loss and similar words result will be optimized. So for each of the entities, we obtained semantically similar terminologies and used it for the ontology population. Section 4 describes the results in detail. After creating sets of vital key entities and definitions, with the input from our legal expert and extracted vital key terms and semantically similar terms from the corpus, we can populate our knowledge graph with the facts and rules contained in the CFR.

4.3.3 Extraction of Relations Between Key Entities. To extract relations between key entities, using text mining techniques, we first extracted the description of each entity from the raw text (as explained in Section 4.3.1). We applied Stanford POS Tagger on the raw text to generate the subject-object-predicate rule. The list of key entities extracted was used as subjects or objects, and associated actions or predicates were considered as relations. We extracted all associated relations of entities. To establish the relationship between two entities, we calculated the frequency of occurrence of entity-relation in the text. The most frequent entity-relation occurrences were considered for the knowledge graph. We validated our results through a legal dictionary [35] and with the help of our legal expert. Figures 4 and 5 illustrate the extraction and validation process by the legal expert.

4.4 Rules Identification and Classification Using Deontic Logic

In this phase, we classify the extracted components, such as definitions of key terms and rules in the sections, into basic deontic expressions using modal logic [43]. These key terms and rules listed in the CFR define the rights, obligations, and prohibitions for the key stakeholders, such as the federal agency, organizations, and researchers. Deontic rules can be reasoned over by our framework to answer questions like "What are the responsibilities of a Contracting Officer?" The answer to such questions should clearly specify four deontic expressions.

4.4.1 Theory of Modal/Deontic Logic. Modal logic is a broad term used to cover various other forms of logic, such as temporal logic and deontic logic [43]. Deontic logic describes statements about permissions and obligations, and temporal logic describes the temporal expressions. Deontic logic further consists of four types of modalities:

— *Permissions/Rights*: Permissions are expressions or rules that describe the rights or authorizations for an entity.

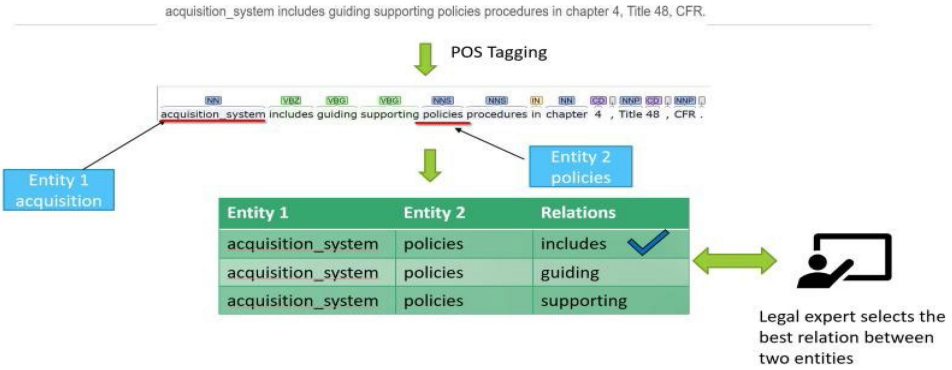

Fig. 5. Extracting best relation between entities: policies and CFR.

— *Obligations*: Obligations expressions are the mandatory actions that an entity must perform.
— *Dispensations*: Dispensations that describe optional expressions and describe non-mandatory conditions.
— *Prohibitions*: Prohibitions are the expressions that specify the actions that are prohibited.

In our previous work, we used text mining techniques to extract deontic rules from cloud SLA documents [13, 14]. We used similar techniques to classify sentences into permissions and prohibitions. We implemented Stanford POS Tagger [27] for each of the sentences in the document comprising vital components. Next, we formulated grammatical rules based on the POS tags to obtain rules in the form of permissions and prohibition. The following grammar rules were used to classify text into the deontic expression:

Permissions: < *Noun/Pronoun* > < *deontic* > < *verb* >
Obligations: < *Noun/Pronoun* > < *deontic* > < *adverb* > < *verb* >
Prohibitions: < *Noun/Pronoun* > < *deontic* > < *negation* > < *verb* >
Dispensation: < *Noun/Pronoun* > < *deontic* > < *negation* > < *adverb* > < *verb* >

We used the following modal verbs for extraction of deontic expressions:

— *Prohibition*: should not, must not, shall not
— *Permission*: can, may, could, might
— *Obligations*: should, must, shall
— *Permission*: can not, may not, could not, might not

After extracting all statements containing the deontic expression, we stored each statement as an instance of the Rules class in our ontology. We also classified each rule into one of the four deontic types using the modal verbs present in the rule. Section 5.2 lists the results we obtained for this phase.

5 VALIDATION RESULTS

In this section, we describe the results of the validation of our framework by applying it to Title 48 of the CFR, which describes the FAR. We collected and preprocessed the FARS document and ran it against the last two phases of our framework—that is, population of the knowledge graph with instances of entities and relations and capturing the deontic rules in each subsection. We validated our results with the help of our legal expert and from the legal dictionary.

Use case. FAR contain the government-wide acquisition regulations jointly issued by the General Services Administration, the Department of Defense, and the National Aeronautics and Space Administration. All procurement-related rules that must be adhered to by organizations doing business with the U.S. federal agencies are included in the FAR.

Table 1. Main Key Entities and Descriptions in Title 48 of the CFR That Were Automatically Extracted by Our System. Multiple descriptions were concatenated.

Legal Terms	Definitions
Acquisition	Acquiring by contract with appropriated funds of supplies, services for the use of the federal government through purchase or lease, whether the supplies or services are already in existence must be created, developed, demonstrated, and evaluated
Affiliate	Associated business concerns, individuals control one or other
Claim	Written demand or written assertion by one of the contracting parties
Component	Any item supplied to the government as part of an end item or of another component
Contract	Mutually binding legal relationship obligating the seller to furnish the supplies or services and the buyer to pay for them
Contracting officer	Person with the authority to enter into, administer, or terminate contracts, or make related determinations, findings
Conviction	Judgment or conviction of a criminal offense by any court of competent jurisdiction
Depreciation	Charge to current operations that distributes the cost of a tangible capital asset, less estimated residual value, over the estimated useful life of the asset in a systematic and logical manner
Debarment	Action taken by a debarring official under 9.406 to exclude a contractor from government contracting and government-approved subcontracting for a reasonable, specified period
Federal agency	Executive agency or any independent establishment in the legislative or judicial branch of the government
Information security	Protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction
Servicing agency	Agency that will conduct an assisted acquisition on behalf of the requesting agency
Bid sample	A product sample required to be submitted by an offeror to show characteristics of the offered products that cannot adequately be described by specifications, purchase descriptions, or the solicitation
Chief Acquisition Officer	Executive-level acquisition official responsible for agency performance of acquisition activities and acquisition programs created pursuant
Conviction	A judgment or conviction of a criminal offense by any court of competent jurisdiction, whether entered upon a verdict or a plea and includes a conviction entered upon a plea of nolo contendere

Consider a researcher in a government contracting company who is interested in knowing the rules for responding to the General Services Administration's Request for Proposal. Title 48 of the CFR has a total of 99 chapters and 9,999 parts, and so manually parsing the document will be very time consuming and labor intensive. The researcher will be able to use a simple interface to ask a question in regular English, and the system will automatically parse the question, identify the terms and keywords the question contains, and then query the knowledge graph to return a definition or list of all rules in the knowledge graph that contain those key terms and also terms that are semantically similar to the terms in the question.

5.1 Populating the CFR Knowledge Graph with FAR Terms and Rules

To populate the knowledge graph, we used the third phase of our framework to extract the vital key entities and definitions, semantically similar terms, and relations between key entities from the FAR document. Table 1 shows

Table 2. Semantically Similar Words Extracted from the Word Embedding Model

Query Words	Analogous Words
Acquisition	Acquisitions, procurement, subpart, department wide, purchases
Certification	Certifications, proprietorships, rationale, approval, balances
Debarment	Suspension, action, ineligibility, actions, protest, debarring, suspended
Request	Waiver, obtain, invite, requested, approval, submit, provide
Signature	Derive, requisition, turpitude, authentication
Patent	Invention, application, experiments
Publication	Document, findings, survey, certification
Agency	Authority, office, DHS, official
Rule	Guidelines, terms, provision, regulations
Enforcement	Memoranda, obligation, legislate
Violation	Immorality, iniquitousness, iniquity
Invoice	Financing, entry, recommendation, demilitarization, certified
Database	Attorney, bulletin, prospect
Patent	Intellectual, invention, tolerance, court-jurisdiction

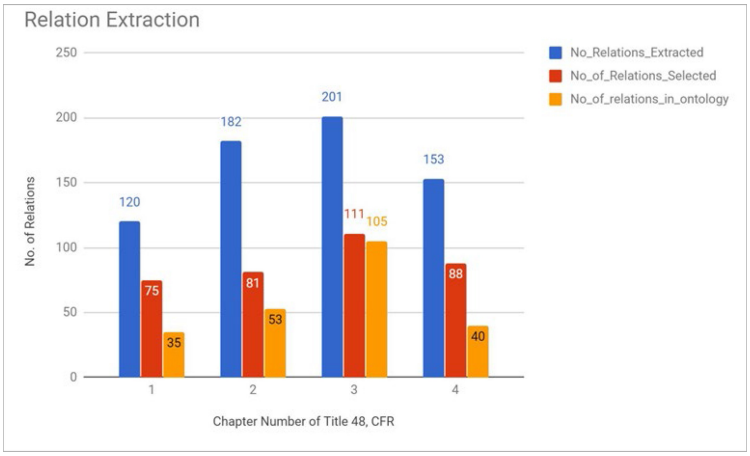

Fig. 6. Analysis of number of relations selected for ontology for Chapter 1-4, CFR Title 48.

some of the automatically extracted key terms and definitions embedded in the document. While extracting semantically similar terms, we got some interesting results. For example, for a query keyword like "acquisition," some of the words extracted by our framework were "procurement" and "purchase," all of which are semantically similar. Table 2 lists some of the semantically similar terms that were extracted by the framework. We used the extracted vital keys and semantic relationship between the key terms for the ontology population. Figure 6 illustrates the statistics of the number of relations added to the ontology. Figure 7 illustrates a partial view of instances that were populated in the knowledge graph. Our legal expert validated the results of this section.

5.2 Classifying Rules into Deontic Expressions

Using our framework, we extracted the deontic expressions embedded in all chapters of FARS. We classified each sentence into one of the deontic types. In total, 9,084 deontic expressions were extracted. With help from our

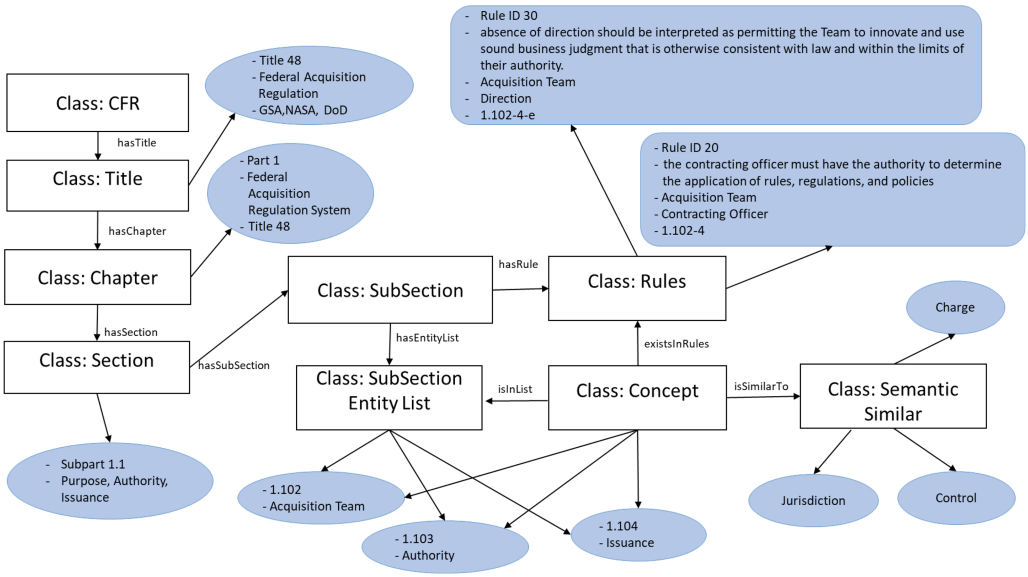

Fig. 7. Partial view of the populated knowledge graph for Part 1 of the CFR, Title 48. The rectangles represent the classes and ovals are the instances populated by our framework.

legal expert, we classified each sentence into four types: permission, prohibition, obligations, and dispensation. Table 3 and Figure 8 illustrate the results of the classification of sentences into deontic expressions.

Some deontic rules in Title 48 of the CFR that were successfully extracted by our system include the following.

Permission. "The contracting officer may request from CCC and any other sources whatever additional information is necessary to make the responsibility determination." [Subpart 209.1, Part 209, Subchapter B, Chapter 2, Title 48].

Dispensation. "Matters related to legal sufficiency reviews that cannot be resolved between the respective CO and SOL Attorney-Advisor must be submitted..." [Subpart 1401.7001-2, Part 1401, Subchapter A, Chapter 14, Title 48].

Obligation. "The military departments and defense agencies shall provide a rolling annual forecast of acquisitions at the end of each quarter (i.e., March 31; June 30; September 30; December 31), to the Deputy Director, Defense Procurement and Acquisition Policy (Contract Policy and International Contracting)." [Subpart 201.170, Part 201, Subchapter B, Chapter 2, Title 48].

Prohibition. "The Secretary of Defense determines in writing that it should not be practicable to carry out the acquisition without continuing to use a contractor to perform lead system integrator functions and that doing so is in the best interest of DoD [Department of Defense]. The authority to make this determination may not be delegated below the level of the Under Secretary of Defense for Acquisition, Technology, and Logistics..." [Subpart 209.5, Part 209, Subchapter B, Chapter 2, Title 48].

6 CONCLUSION AND FUTURE WORK

Currently, legal documents like the CFRs are managed and analyzed as long and complex text documents. The manual analysis and retrieval of relevant information across various titles and chapters is a complicated and time-consuming process. In this article, we presented a novel framework built using AI technologies like Semantic

Table 3. Total Number of Deontic Expressions

Deontic Type	Total Extracted Sentences
Permission	710
Obligation	698
Prohibition	479
Dispensation	149

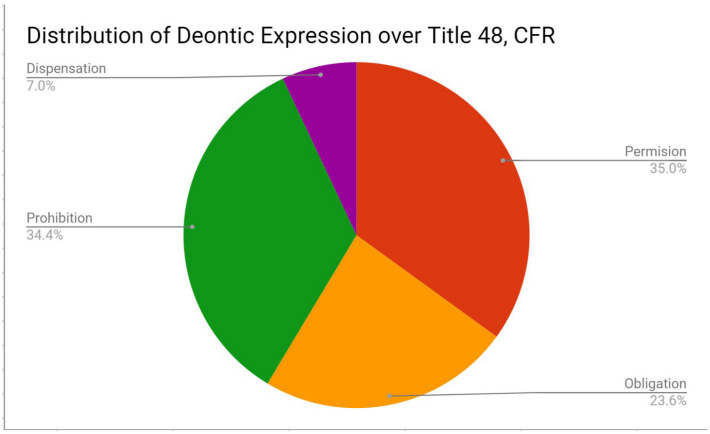

Fig. 8. Distribution of deontic expressions embedded in Title 48 of the CFR.

Web, text analytics, and information retrieval, which can facilitate automation of organizational processes that depend on the knowledge embedded in the CFR titles. We developed techniques to automate the extraction of essential key terms/definitions, semantically similar terms for the ontological representation of the legal knowledge base. In addition to this, we classified text into deontic expressions as permission prohibition, obligation, and dispensation using pattern-based rules. For this work, we focused on the validation results of analyzing FARS (Title 48 of the CFR) using this framework. The long-term goal of this research is to build an efficient and automated legal question and answer system for all CFR titles. This framework contributes to automating and thereby improving eGovernment processes that currently heavily depend on researchers discovering, parsing, and interpreting CFR rules and regulations.

6.1 Ongoing and Future Work

We are working on developing a user interface for our framework where researchers can ask a question in regular English, and the system will automatically parse the question, identify the terms and keywords the question contains, and then query the knowledge graph to return a definition or list of all rules in the knowledge graph that contain those key terms and also terms that are semantically similar to the terms in the question. This interface will make it convenient for non-tech-savvy people to use the framework.

As part of our future work, we will populate our CFR knowledge graph with data in all other titles of the CFR. This knowledge graph will eventually be a vital part of the integrated legal knowledge base that we are building.

We will be developing techniques to handle the hierarchy within the rules and regulations. These will address situations where a rule will supersede another rule or will be negated by another rule. Currently, all deontic rules are ranked with the same priority by the system. We will extend our framework to allow prioritization of the deontic rules (the prohibitions will override the dispensations, etc.). These enhancements will also help identify discrepancies in the rules, such as when the same rule is prohibited by one CFR title yet permitted by another CFR title. As part of our ongoing work, we are getting other domain experts and users involved in validating and improving the knowledge graph.

Although users are currently able to query the CFR knowledge graph for rules and definitions, the framework cannot provide legal advice without human validation. Actionable legal advice is the future focus of this research. We are also working to develop a question and answer system that, for any given action or question on the CFR, can highlight all terms, policies, rules, and stakeholder entities that might be applicable to it and offer preliminary guidance to a researcher or government official.

ACKNOWLEDGMENTS

We would like to thank our legal expert Renee Frank for constant guidance and support while validating results. We would also like to thank Jiayong Lin (UMBC) and Michael Aebig (UMBC) for technical help with this work.

REFERENCES

[1] F. Amato, A. Mazzeo, A. Penta, and A. Picariello. 2008. Building RDF ontologies from semi-structured legal documents. In *Proceedings of the 2008 International Conference on Complex, Intelligent and Software Intensive Systems*. 2008, 997–1002.

[2] Richard J. McKinney. 2002. A research guide to the Federal Register and the Code of Federal Regulations. In *LLSDC's Legislative Sourcebook*. Law Librarians' Society, Washington, DC, 10–15.

[3] K. P. Joshi, Y. Yesha, and T. Finin. 2014. Automating cloud services life cycle through semantic technologies, services computing. *IEEE Transactions on Service Computing* 7, 1 (2014), 109–122.

[4] Srishty Saha, Karuna Joshi, Renee Frank, Michael Aebig, and Jiayong Lin. 2017. Automated knowledge extraction from the Federal Acquisition Regulations System (FARS). In *Proceedings of IEEE 2nd International Workshop on Enterprise Big Data Semantic and Analytics Modeling*.

[5] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, and D. Mladenic. 2007. Triplet extraction from sentences. In *Proceedings of the 10th International Multiconference Information Society (IS'07)*. 812. http://plato.stanford.edu/entries/logicdeontic/.

[6] O. Etzioni, M. Cafarella, D. Downey, A. M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. 2005. Unsupervised named entity extraction from the Web: An experimental study. *Artificial Intelligence* 165, 1 (2005), 91–134.

[7] P. Cimiano, S. Staab, and J. Tane. 2003. Automatic acquisition of taxonomies from text: FCA meets NLP. In *Proceedings of the International Workshop and Tutorial on Adaptive Text Extraction and Mining Held in Conjunction with the 14th European Conference on Machine Learning and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*.

[8] J. Cowie and W. Lehnert. 1996. Information extraction. *Communications of the ACM* 39, 1 (1996), 80–91.

[9] K. Barker and N. Cornacchia. 2000. Using noun phrase heads to extract document key phrases. In *Advances in Artificial Intelligence*. Lecture Notes in Computer Science, Vol. 1822. Springer, 40–52.

[10] K. Joshi and T. Finin. Ontology for Cloud Services SLA (Service Level Agreement). Retrieved October 22, 2020 from http://ebiquity.umbc.edu/resource/html/id/344.

[11] K. Joshi. Ontology for Services on the Cloud. Retrieved October 22, 2020 from http://ebiquity.umbc.edu/resource/html/id/318/ OntologyforServicesontheCloud.

[12] O. Lassila and R. Swick (Eds.). 1999. *Resource Description Framework (RDF) Model and Syntax Specification*. World Wide Web Consortium.

[13] D. McGuinness and F. Van Harmelen (Eds.). 2004. *OWL Web Ontology Language Overview: W3C Recommendation*. World Wide Web Consortium.

[14] W3C. 2013. SPARQL 1.1 Overview. Retrieved October 22, 2020 from http://www.w3.org/TR/sparql11-overview/.

[15] Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*.

[16] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization (NAACL-ANLP-AutoSum'00)*, Vol. 4. 40–48.

[17] L. Thorne McCarty. 2007. Deep semantic interpretations of legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL'07)*.

[18] John O. McGinnis and Russell G. Pearce. 2014. The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services. *Fordham Law Review* 82, 6 (2014), 1–26.

[19] Rand Ghayad, Paul Hinton, Mark Sarro, and Michael Cragg, the Brattle Group Inc., and David Cohen and Reed Smith. 2015. Making the Most of Document Analytics. Retrieved February 24, 2020 from https://www.law360.com/articles/730189.

[20] Travis D. Breaux, Matthew W. Vail, and Annie I. Anton. 2006. Towards compliance: Extracting rights and obligations to align requirements with regulations. In *Proceedings of the IEEE 14th International Requirements Engineering Conference (RE'06)*. 49–58.

[21] Travis D. Breaux and Annie I. Anton. 2005. Analyzing goal semantics for rights, permissions, and obligations. In *Proceedings of the IEEE 13th International Requirements Engineering Conference (RE'05)*. 177–186.

[22] Sudip Mittal, Karuna Pande Joshi, Claudia Pearce, and Anupam Joshi. 2016. Automatic extraction of metrics from SLAs for cloud service management. In *Proceedings of the IEEE International Conference on Cloud Engineering (IC2E'16)*.

[23] Sudip Mittal, Karuna Joshi, Claudia Pearce, and Anupam Joshi. 2015. Parallelizing natural language techniques for knowledge extraction from cloud service level agreements. *Poster presented at the IEEE International Conference on Big Data*.

[24] Karuna P. Joshi, Aditi Gupta, Sudip Mittal, Claudia Pearce, Anupam Joshi, and Tim Finin. 2016. Semantic approach to automating management of big data privacy policies. In *Proceedings of the IEEE Big Data Conference*.

[25] e-CFR. 2020. Electronic Code of Federal Regulations. Retrieved October 22, 2020 from https://www.ecfr.gov/.

[26] e-CFR. 2020. Title 48 Federal Acquisition Regulations System. Retrieved October 22, 2020 from at https://www.ecfr.gov/cgi-bin/text-idx?SID=0bc8c1900bc75327bde6606d1791872d&mc=true&tpl=/ecfrbrowse/Title48/48tab_02.tpl.

[27] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*. 252–259.
[28] Carnegie Mellon University. n.d. Link Grammar Parser. Retrieved October 22, 2020 from http://www.link.cs.cmu.edu/link/.
[29] Python. n.d. ElementTree Python Library. Retrieved October 22, 2020 from https://docs.python.org/2/library/xml.etree.elementtree.html.
[30] Peter Liu and Xin Pan. 2016. Google AI Blog: Text Summarization with TensorFlow. Retrieved November 16, 2020 from https://ai.googleblog.com/2016/08/text-summarization-with-tensorflow.html.
[31] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781
[32] Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, et al. 2016. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467
[33] Gensim. n.d. Gensim for Topic Modelling. Retrieved October 22, 2020 from https://radimrehurek.com/gensim/.
[34] N. Kiyavitskaya, N. Zeni, T. D. Breaux, A. I. Anton, J. Cordy, L. Mich, and J. Mylopoulos. 2008. Automating the extraction of rights and obligations for regulatory compliance. In *Proceedings of the 27th International Conference on Conceptual Modeling (ER'08)*.
[35] Law.com. n.d. Legal Dictionary. Retrieved October 22, 2020 from https://dictionary.law.com/.
[36] Barbara Bavis. 2014. How to Trace Federal Regulations—A Research Guide. Retrieved February 16, 2020 from https://blogs.loc.gov/law/2014/11/how-to-trace-federal-regulations-a-research-guide/.
[37] Publications Office of the European Union. n.d. European Legislation Identifier (ELI) Ontology. Retrieved February 16, 2020 from https://op.europa.eu/en/web/eu-vocabularies/eli.
[38] Cleyton Mário de Oliveira Rodrigues, Frederico Luiz Gonçalves de Freitas, Emanoel Francisco Spósito Barreiros, Ryan Ribeiro de Azevedo, and Adauto Trigueiro de Almeida Filho. 2019. Legal ontologies over time: A systematic mapping study. *Expert Systems with Applications* 130 (2019), 12–30.
[39] Stanford. n.d. Stemming and Lemmatization. Retrieved October 22, 2020 from https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.
[40] Gov.info. n.d. Federal Register. Retrieved February 19, 2020 from https://www.govinfo.gov/app/collection/fr/2020.
[41] Mark A. Musen. 2015. The protégé project: A look back and a look forward. *AI Matters* 1, 4 (June 2015), 4–12. DOI: 10.1145/2557001.25757003
[42] Data Science Bootcamp. 2018. Understanding the Softmax Function in Minutes. Retrieved February 21, 2020 from https://medium.com/data-science-bootcamp/understand-the-softmax-function-in-minutes-f3a59641e86d.
[43] Stanford Encyclopedia of Philosophy. n.d. Modal Logic. Retrieved October 22, 2020 from http://plato.stanford.edu/entries/logic-modal/.
[44] A. Valente. 2005. Types and roles of legal ontologies. In *Law and the Semantic Web*. Lecture Notes in Computer Science, Vol. 3369. Springer, 65–76.
[45] A. Gangemi, M. T. Sagri, and D. Tiscornia. 2005. A constructive framework for legal ontologies. In *Law and the Semantic Web*. Lecture Notes in Computer Science, Vol. 3369. Springer, 97–124.
[46] A. Gómez-Pérez, F. Ortiz-Rodríguez, and B. Villazón-Terrazas. 2005. Legal ontologies for the Spanish e-government. In *Current Topics in Artificial Intelligence*. Lecture Notes in Computer Science, Vol. 4177. Springer, 301–310.

Received August 2019; revised July 2020; accepted September 2020