# Cognitively Rich Framework to Automate Extraction and Representation of Legal Knowledge

**Srishty Saha** and **Karuna P. Joshi**

**Abstract** With the explosive growth in cloud based services, businesses are increasingly maintaining large datasets containing information about their consumers to provide a seamless user experience. To ensure privacy and security of these datasets, regulatory bodies have specified rules and compliance policies that must be adhered to by organizations. These regulatory policies are currently available as text documents that are not machine processable and so require extensive manual effort to monitor them continuously to ensure data compliance. We have developed a cognitive framework to automatically parse and extract knowledge from legal documents and represent it using an Ontology. The framework captures knowledge in 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 built the framework using Deep Learning technologies like Tensorflow, for word embeddings and text summarization, Gensim for topic modeling and Se- mantic Web technologies for building the knowledge graph. We have applied this framework to the United States government's Code of Federal Regulations (CFR) which includes facts and rules for individuals and organizations seeking to do business with the US Federal government. In this paper we describe our framework in detail and present results of the CFR legal knowledge base that we have built using this framework. Our framework can be adopted by businesses to build their automated compliance monitoring system.

**Keywords** Deep Learning · Legal Text Analytics · Compliance · Semantic Web

Srishty Saha
Computer Science and Electrical Engineering Department, UMBC, USA
E-mail: srishty1@umbc.edu

Karuna P. Joshi
Information Systems Department, UMBC,USA
karuna.joshi@umbc.edu

## 1 Introduction

There has been an exponential growth in digitized legal documents in this century, especially after the rapid adoption of Internet or Cloud based services by businesses to simultaneously cater to millions of their consumers. Regulatory bodies require organizations to comply with strict privacy and security rules pertaining to their consumers Privately Identifiable Information (PII). Hence, organizations today are maintaining large data sets of legal contracts (like terms of services, privacy policies, purchase agreements, etc.) that they have signed with their customers, employees and contractors. While these contracts are finalized by the legal counsels of the companies, often these contracts are monitored by different departments that deal directly with the other party, like Human Resources monitors employee contracts, Sales and marketing manage customer agreements and IT department deals with IT contractors or software vendor agreements, etc. This process of managing and monitoring an ever increasing dataset of legal contracts, regulations and compliance is still a very manual and labour intensive effort and can prove to be a bottleneck in the smooth functioning of an enterprise. Automating this is hard because while the information is digitally available as text documents, it is not represented in a machine understandable way.

Representation of legal documents has been an active area of research. However, there has been limited work on automatically extracting rules and policies from regulatory documents that need to be complied upon to ensure data security and privacy. In our previous work [5][13][14][32], we have identified the various compliance regulations that apply to data managed on the cloud. As part of this project, we have also been analyzing regulatory policies issued by the United States (US) federal government. These regulations can be intimidating for novices and veterans because many areas of regulation are complex and voluminous and may be scattered across multiple sections making it hard to cross-reference critical procedures and rules. Documents like the Code of Federal Regulations (CFR) System [41] is a long and complex document. Currently the analysis of CFRs require legal expertise and is a time consuming and labour intensive process. Developing cognitive assistant for such long and complex documents will help businesses as well as legal experts to analyze the legal elements easily and efficiently. The CFR is available in electronic form [41] on a variety of free and paywall sites, but its semi-structured organizational structure makes it a challenge to find all of the relevant sections that a user may need to review to answer a particular question. Keyword searches may also return vast numbers of possible matches requiring large amounts of human review to analyze and sort the relevant and irrelevant responses. The organizational structure of the document also makes it difficult to find and compare relevant provisions across sections and titles because indexing of the information (through sectional tables of contents) is carried out at relatively high levels within the regulatory sections.

Traditional techniques of Natural Language Processing and Information Retrieval techniques like Bag of words model or vectorized model alone cannot automate the analysis process of legal documents. This is because, it fails to captures the semantic relationships between various legal elements spread across the deep hierarchical structure of legal documents. Dealing with heterogeneous legal facts and rules in semi-structured format like XML is difficult in terms of answering user queries and performing analysis on various legal element. Hence, building ontologies for legal documents is one of the possible efficient solutions to capture various facts and rules of legal documents in order to perform analytics and answer queries.

As part of our Automated Legal Document Analytics (ALDA) [2] project, we have been developing 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 Natural Language Processing (NLP). Our long term goal is to develop a system that for any given action or question, can highlight all the statutes, policies, laws and case law that might be applicable on it and offer preliminary guidance to a counsel. As a shorter term vision, we're looking to see if we can automatically extract elements from compliance and regulatory legal documents that govern Information Technology (IT) outsourcing/cloud computing and automatically answer the question, "Is the running system in compliance with the policies agreed to by the consumer and provider?"

We have developed a cognitive framework to automatically parse and extract knowledge from legal documents and represent it using an Ontology. The framework captures knowledge in 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 [47][48], for word embeddings and text summarization[46], Gensim[49] for topic modeling and Semantic Web technologies for building the knowledge graph. In this paper, we describe this framework in detail and also present our results of applying the same to analyzing CFR documents. Section 2 covers related work in this area. Section 3 describes the methodology we developed using Information Retrieval, Natural Language Processing and Deep Learning techniques for creating legal knowledge graph. Section 4 details our results and Section 5 describes conclusion and future work.

## 2 Related Work

Machine intelligence, specially developments in predictive analytics, is dramatically changing five areas in the legal domain: (1) discovery (2) legal search (3) document generation (4) brief and memoranda generation and (5) prediction of case outcomes [25].

Electronic discovery (also called ediscovery or e Discovery) 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 [26]. While predictive coding is gaining increased acceptance as a procedure for identifying responsive documents with less manual review, there is less appreciation of how document analytics can add value in answering document related research questions, or otherwise helping to 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 manual review, enables algorithms to be run across all documents across multiple datasets and dictionaries at relatively short time and cost. While the results of computerized document classification may not be perfect, analyzing all documents collectively reveals patterns not visible from targeted manual review. For example, important patterns of communication concerning particular topics may only become apparent once all messages are analyzed and mapped. Furthermore, algorithms can be used to gather individual pieces of similar information of interest across an entire database, for example 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 in many cases keyword searches can be overinclusive. That is, they return responsive documents with an overwhelming set of irrelevant documents. They can also be under inclusive. For example, the lack of standardized terms used in conversations and documents makes it hard to retrieve all documents relevant to a given set of search terms. Searching for the words automobile and car will miss references to BMW and Mercedes. The mere formulation of a query or keywords is difficult if the information being targeted can be described in several different ways. Moreover, simple search queries may return ambiguous uses of the keywords being searched. It may retrieve hits of the words that are not really 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 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 the target words are not contained in the document. Predictive coding and context searching have also been accepted by a number of courts [26]. However, existing TAR (Technology Assisted Review) technologies and predictive analytics offerings from eDiscovery vendors are not being adopted readily by the legal community due to Usability issues and data/system maintainability and per case config phrases and metadata in the training set, essentially converting text to data. Just as human reviewers reach different decisions on the relevance of the same document, a predictive coding model may make predictions that do not match an attorneys decisions in every instance. Moreover, the results offered of the existing legal analytics approaches are still not semantically rich. Some researchers have attempted to address these concerns.

An innovative approach we want to use for building our system is to extend the text analytics to clauses, or whole sentences in the document instead of limiting it to keyword search. Researchers have proposed approaches for extracting sentences instead of just keywords from text documents. Le and Mikolov [21] have recommended the paragraph vector algorithm for extracting sentences from documents. Goldstein et. al. [22] have 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 in [23] illustrated that a statistical parser can handle complex syntactic constructions of an appellate court judge, and that a deep semantic interpretation of the full text of a judicial opinion can be computed automatically from the output of the parser. We will explore and build upon all these approaches when building and automatically populating our system.

In our previous work, we have developed a semantically rich ontology for Service Level Agreements(SLAs) and Privacy Policies for cloud based services [5][13][14][32]. The Semantic Web deals primarily with data instead of documents. It enables data to be annotated with machine understandable metadata, allowing the automation of their retrieval and their usage in correct contexts. Semantic Web technologies include languages such as Resource Description Framework (RDF) [50] and Web Ontology Language (OWL) [51] for defining ontologies and describing metadata using these ontologies as well as tools for reasoning over these descriptions. These technologies can be used to provide semantic relationships between various legal elements of Code for Federal Regulations. Information extraction from text documents have been active area of research. Rusu et. al. [6] used parse trees to generate triplets as subject-predicate-object. Etzioni et. al. [7] used pattern learning to generate 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. Use of automated techniques for extracting permissions and obligations from legal documents, such as text mining and semantic techniques have been explored by researchers in the past [52][53][54].

In our previous, we also extracted key SLA definitions and measures from these documents using pattern-based rules using the Stanford PoS Tagger [43] and CMU Link Parser [44] and also used pattern based rules for extracting permission and obligation [13][14].

But, CFRs titles are much longer and complex documents than Service Level Agreements or Privacy Policies of cloud services, we need to improve and redefine our existing approach for developing an ontology for CFRs in order to capture various facts and rules spread over the documents. In Section 3, we have described our methodology to capture vital key-entities, semantic relationships between key-entities, contexually similar terminologies and identifying basic deontic expressions. Section 4 describes the results for automated extraction and representation of legal knowledge.

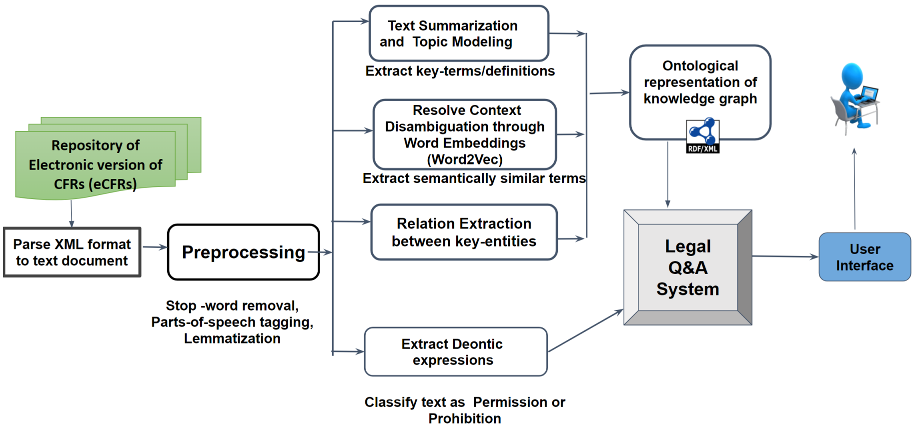
Repository of Electronic version of CFRs (eCFRs)
Parse XML format to text document
Preprocessing
Stop -word removal, Parts-of-speech tagging, Lemmatization
Text Summarization and Topic Modeling
Extract key-terms/definitions
Resolve Context Disambiguation through Word Embeddings (Word2Vec)
Extract semantically similar terms
Relation Extraction between key-entities
Extract Deontic expressions
Classify text as Permission or Prohibition
Ontological representation of knowledge graph
RDF/XML
Legal Q&A System
User Interface

Fig. 1: Overall architecture of cognitive framework for automated extraction and representation of legal knowledge

## 3 Methodology

The Code of Federal Regulations (CFRs) [41] is a semi-structured XML formatted legal text. The document is complex and lengthy. It has 50 Titles and each title has on-an average 50 chapters. Each chapter has 100 parts ( avg) and each part has various subparts, sections and subsections. Apart from the complex hierarchical structure of CFRs, the rules and regulations have been cross-referenced across document of CFRs. For example, if a user queries about "Rules and regulations for Technology Investment in Federal Agencies", the related answers are present in various chapters of Title 48, CFR (Federal Acquisition Regulations) and Title 32, CFR (National Defense). Manual querying of answers is labor intensive and time consuming process. To resolve this issue, we intend to build an automated knowledge extraction system for answering legal questions using Natural Language Processing, Information Retrieval, Deep Learning and Semantic web techniques.

In this section, we describe the methodology of building ontological representation of legal knowledge graph capturing key-information as well as structure of CFRs, extracting cross-referenced rules from the legal knowledge graph and identifying and classifying rules into basic deontic expressions. Figure 1 describes the overall architecture of automated knowledge extraction system.

### 3.1 Data Collection and Preprocessing

The online electronic version of Code of Federal Regulations are in XML format in a hierarchical structure having tables and figures. We created a repository of all the titles of CFRs.The text portion from these documents is extracted using ElementTree python library [45]. Then, we preprocessed the extracted text using Natural Language Processing techniques such as conversion to lowercase,removal of stop words, lemmatization and parts of speech 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. Also, we did not remove alpha-numeric characters and numbers from the text as they might represent the structure of the document. So, while extracting text from XML format as well as preprocessing the text, we intentionally maintained the numbered hierarchical structure of the document.

## 3.2 Ontological representation of legal knowledge graph

Considering the Code of Federal Regulations being a lengthy and very complex document.The identification and validation of legal key-entities and relations for building ontology is a challenge as there are no existing ontologies similar to DBpedia or Freebase for legal documents as ground truth. We have used Title 48, CFR representing the Federal Acquisition Regulations System to conduct experiments.
In this subsection, we use 3 step approach to build a legal ontological-knowledge graph:

- Extraction and validation of key-entities and attributes.
- Extraction of semantically similar terminologies and ontology populations.
- Extraction of relations between key-entities.

### *3.2.1 Extraction and Validation of key-entities of knowledge graph*

The titles of CFRs have various chapters, parts, sub-parts, sections and subsections of varied number of sentences representing facts and rules. In order to extract important key-entities for legal ontology, we summarized the text and performed topic modelling on the summarized text. Text summarization is done to capture only vital information from lengthy paragraphs of section and sub-sections of a Title of CFRs. To preserve the vital information while performing the summarization, we performed TensorFlow extractive text summarization model [46] on each paragraph of whole document. Extractive text summarization takes words and words phrases to create summary which in turn does not least to information loss. After summarizing the text, we implemented Latent Dirichlet Allocation [49] Topic Modelling on the summarized text to extract top $k$ topics from each section. For example, if a summarized section A has 10 sentences then 5 topics will be extracted whereas another summarized section B having 50 sentences will have 15 extracted topics. The validation of considering topics as key-entities for legal ontology has been done with the help of legal dictionary [54] and our legal expert. We also extracted definitions of those key-entities from the document. In our previous work, we

developed a topic descriptor system for extracting definitions of topic of importance using pattern based rules like Stanford POS Tagger [43] and CMU Link Parser [44] from Service Level Agreements of cloud based-services [2]. The extracted topics from each section of Title 48 form a set of vital key-terms which contribute towards forming legal entities for ontology.
After creating set of vital key-entities of Title 48 of CFRs, we extracted definitions or description of those key-entities from the document. In our previous work, we developed a topic descriptor system for extracting definitions of topic of importance using pattern based rules like Stanford POS Tagger [43] and CMU Link Parser [44] from Service Level Agreements of cloud based-services [13][14]. We used same system to extract description of entities from each Title of CFRs. Also for each entity, we obtained associated section number, Part number, Chapter number and Title number i.e, location where entity has been mentioned in the document. Description and location of each of the entities will form set of attributes for each entity of legal knowledge graph. Section 4 describes the result.

**Algorithm 1** Entity Extraction

Let T be the hashmap where keys of hashmap store the location of paragraph and the values of hashmap store array of topics extracted from Latent Dirichlet Allocation Topic Modelling.

```
input ← Pre-processed raw text
For Title_number in each Title of CFRs:
  For Chapter_number in each Chapter within Title:
    For Part_number in each Part within Chapter:
      For Subpart_number in each Subpart within Part:
        For Section_number in each Section within Subpart:
          For Sub-section_number in Sub-section within section:
            Array A ← Topics extracted from each sub-section
            T [Sub-section_number] =A    */where T is Hashmap/*
For each key in Hashmap T:
  T[key]=A(1:k)
  */Select top k topics from values of each key where value of k varies with length of paragraph in each sub-section/*
```

### *3.2.2 Extraction of semantically similar terminologies and ontology populations*

In Code of Federal Acquisitions, there are various chapters which 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 CFRs as “findings”” or “document”. In order to resolve context disambiguation, we used TensorFlow Word2Vec deep learning architecture [48] [49] to generate 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 analogous words. To build and train the skip-gram model, several parameters need to be decided, which are batch size, number of skips and skip-window. The skip-window represents the number of words to be considered at left and right of the target word. And num-of-skips represents the number of output words will be picked in the span of a single word in a (input, output) tuples. We used the set of target words, we are interested in to evaluate the similarity on, every certain steps, the model is evaluated by looking 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 set of words in the skip-window which are 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 *softmax* function [] 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 fully connected layer of deep-neural network architecture after every epoch.

$$P(w) = \frac{\exp^{w}}{\sum\limits_{i=1}^{K} \exp^{(}\, w_i)} \tag{1}$$

The equation 1 describes the probability calculation using Softmax function[]. In order to, maximize the likelihood of probability $P(w)$, we apply logarithmic function to $P(w)$. The equation 2 describes the maximization of probability $P(w)$.

$$Maximized\ P(w) = \log \frac{\exp^{w}}{\sum\limits_{i=1}^{K} \exp^{(}\, w_i)} \tag{2}$$

The program stops after 100000 steps, and the loss and and similar words result will be optimized. So for each of the entities, we obtained semantically similar terminologies and used it for 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 will create an ontology representation of facts and rules contained in the Code of Federal Acquisition. This ontology will contribute in building legal knowledge graph for answering legal questions. Figure 2 explains the methodology of entity creation for legal knowledge graph

#### *3.2.3 Extraction of relations between key-entities*

In order to extract relations between key-entities, using text mining techniques we first extracted the description of each entities from the raw text (as explained in Section 3.2.1). We applied Stanford POS tagger on the raw text

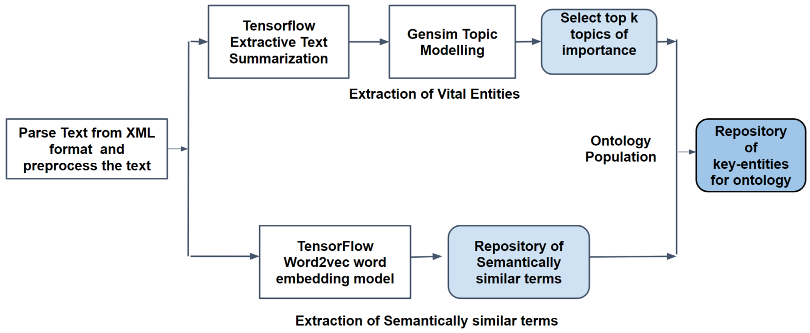
Tensorflow Extractive Text Summarization
Gensim Topic Modelling
Select top k topics of importance
Extraction of Vital Entities
Parse Text from XML format and preprocess the text
Ontology Population
Repository of key-entities for ontology
TensorFlow Word2vec word embedding model
Repository of Semantically similar terms
Extraction of Semantically similar terms

Fig. 2: Methodology to extract vital key-entities and semantic similar terminologies for legal ontology

to generate Subject-Object-Predicate rule. The list of key-entities extracted were used as subjects or objects and associated Actions or Predicates were considered as relations. We extracted all associated relations of entities. In order to establish relationship between two entities, we calculated the frequency of occurrence of entity-relation in the text. The most frequent entity-relation occurrence were considered for knowledge graph. As there are no knowledge graphs like DBpedia and Freebase for Code of Federal Regulations for purpose of ground truth and validation of results, we validated our results through legal dictionary[54] and with the help of our legal expert. Section 4.1 explains the results of legal knowledge graph.

### 3.3 Identification and classification of Deontic Expression

In previous section of this paper, we have extracted vital components of Federal Acquisition Regulations System (Title 48, CFR )such as key-entities and definitions,contextually similar terms. Now, we intend to classify the extracted components such as definitions of key-terms and sections into basic deontic expressions. This method is applicable in answering to questions like ”Why is Federal Agency signatures to required to assist XYZ officer”, the answer to such questions should clearly specify four deontic expressions, i.e, Permission (Do’s), Prohibitions(Don’ts), Obligations(mandatory Do’s) and Dispensation (NonMandatory conditions). We have classified sentences into Permissions, Obligations and Prohibitions. In our previous work, we used text mining techniques to extract deontic rules from cloud SLA documents [2][14]. We have used similar techniques to classify sentences into Permissions and Prohibitions, i.e, we implemented the Stanford POS tagger [43] for each of the sentence of in the document comprising of vital components. Next we formulated grammatical rules based on the POS tags to obtain rules in the form of permissions and prohibition. Section 4.2 describes the results for classification of text into deontic expressions. The following are the grammar rules we used to classify text into deontic expression:

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

Here are few examples for Deontic Expression in Title 48, CFR:

- **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, SubChpater 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. 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, SubChpater B, Chapter 2, Title 48]

## 4 Results

In this section, we describe the results of our framework for automating legal document text analytics. For the analysis, we have experimented on Title 48 of Code of Federal Regulations (CFRs) which describes about the Federal Aquisition Regulations System. Title 48 of CFRs has total of 99 chapters and 9999 parts. We have developed a system for parsing, preprocessing, extraction of key-entities and definitions, capturing semantically similar terminologies, extraction of relation between key-entities, classifying facts and rules as deontic

| 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 controls 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 terminate contracts 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. |

Table 1: Extracted vital key-entities and description from Title 48, CFR

expression and analyzing and reasoning over documents. We have validated our results with the help of our legal expert and from the legal dictionary [54].

### 4.1 Building of ontological legal knowledge graph

In this section, we extracted vital key-entities and definitions, semantically similar terms and and relations between key-entities for ontology population and reasoning over it. As explained in Section 3.2.1 , in order to create legal ontology, we used TensorFlow text summarization to summarize each section and Latent Dirichlet Allocation model to extract top $k$ topics from each sub-section. These top $k$ topics will form set of key-entities for ontology. Subsequently, by using text mining techniques, we extracted definition/description for each key-term. Table 1 shows some of the extracted key-

| Query Keywords | 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 |

Table 2: Semantically similar words extracted from word-embedding model

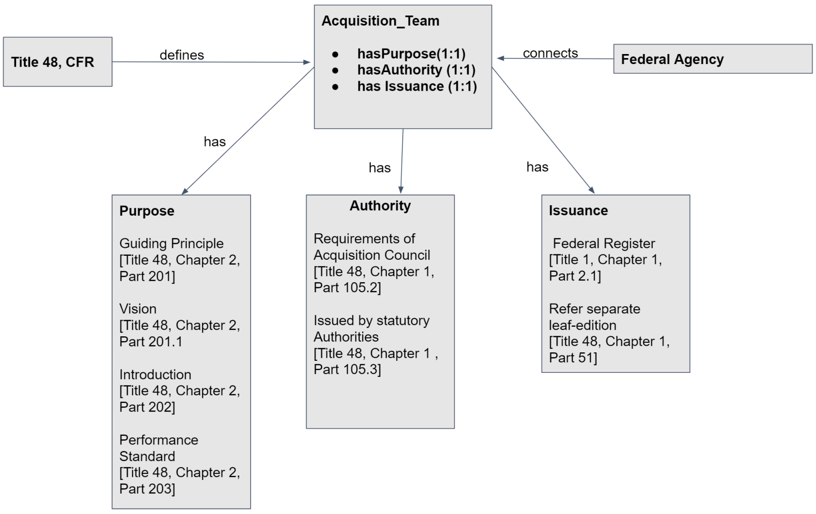
Acquisition_Team
● hasPurpose(1:1)
● hasAuthority (1:1)
● has Issuance (1:1)
Title 48, CFR
defines
connects
Federal Agency
has
has
has
Purpose
Guiding Principle
[Title 48, Chapter 2, Part 201]
Vision
[Title 48, Chapter 2, Part 201.1
Introduction
[Title 48, Chapter 2, Part 202]
Performance Standard
[Title 48, Chapter 2, Part 203]
Authority
Requirements of Acquisition Council
[Title 48, Chapter 1, Part 105.2]
Issued by statutory Authorities
[Title 48, Chapter 1 , Part 105.3]
Issuance
Federal Register
[Title 1, Chapter 1, Part 2.1]
Refer separate leaf-edition
[Title 48, Chapter 1, Part 51]

Fig. 3: Partial view of ontology of Title 48, CFR

terms and definitions. For extracting semantically terms, TensoFlow's word embedding word2vec model have shown some promising results. For example, for a query keyword like "acquisition", some of the words extracted from the model are "procurement" and "purchase" all of which are semantically similar to each other. Table 2 shows some of the semantically similar terms. We will use the extracted vital keys and semantic relationship between the key-terms for for ontology development. The results of this section were validated by our legal expert. Figure 3 describes the partial view of the legal ontology of

Title 48, CFR. We used the extracted key entities and semantic relationships between these key-entities for building the CFR ontology. It consists of main class of Aquisition_team that has three sub-classes of Purpose, Authority and Issuance. In the ontology, we also capture the chapter and part numbers under which the key-terms and rules are listed. This enables us to also note the provenance of the compliance policy.

### 4.2 Classifying rules into deontic expressions

Using the grammar based rules as mentioned above, we extracted deontic expressions from text document of Federal Acquisition Regulations System (Title 48, CFR) and classified each sentences into one of the deontic modal logics (mentioned above).
We used the following modal verbs for extraction 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

For our experiments, we tested our approach on all chapters of Title 48, CFR chapters. In total, 9,084 deontic expressions were extracted. With the help from our legal expert, we classified each sentences into 4 categories: Permission; Prohibition; Obligations and Dispensation. The Table 3 and Figure 4 describes the results of classification of sentences into deontic expressions.

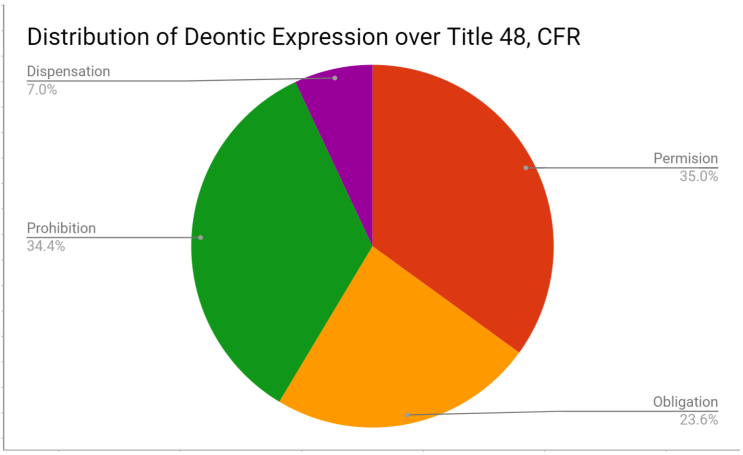

Fig. 4: Distribution of Deontic Expression over Title 48, CFR

| Deontic Expression | Total number of extracted sentences |
|---|---|
| Permission | 3178 |
| Obligation | 2145 |
| Prohibition | 3125 |
| Dispensation | 636 |

Table 3: Results for extraction of deontic expression

## 5 Conclusion and Future work

Currently legal documents like Code of federal regulations are presented and analyzed as text documents. The Code of Federal Regulations (CFRs) [9] is a long and complex document. The analysis and retrieval of relevant information across various titles and chapters manually is a complex and time consuming process. In this paper, we presented an approach towards automating the analysis of legal documents through building an efficient legal knowledge base contributing towards legal question and answer. We focused on the Federal Acquisition Regulations System (Title 48, CFR) for our research. We developed techniques to automate the extraction of important key-terms/ definitions, semantically similar terms for ontological representation of legal knowledge base. In addition to this, we classified text into deontic expressions as Permisssion Prohibition, Obligation and Dispensation. using pattern base rules. Above mentioned three modules will help in developing a building legal knowledge base. This semantically rich legal knowledge base will be a part of legal questions and answer system.As part of our ongoing work, with the help of our legal expert and by using extracted vital key-entities and semantically similar terms of CFRs, we are in the process of creating semantically rich legal ontology for all the titles of Code for Federal acquisition . This ontology will eventually be a vital part of our legal knowledge base.The long terms goal is to build an efficient and automated legal question and answer system.

## 6 Acknowledgement

We would like to thank you 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. This research was partially supported by a DoD supplement to the NSF award 1439663 : NSF I/UCRC Center for Hybrid Multicore Productivity Research (CHMPR).

## References

[1] A. Hendre and K. P. Joshi, A semantic approach to cloud security and compliance, in 2015 IEEE 8th International Conference on Cloud Computing (CLOUD). IEEE, 2015, pp. 10811084.
[2] ALDA: https://ebiquity.umbc.edu/project/html/id/105/ALDA-Automated-Legal-Document-Analytics [3] Minimize Cross Referencing,

http://www.plainlanguage.gov/howto/guidelines/FederalPLGuidelines/writeNoXRefs.cfm
[4] K. P. Joshi, Y. Yesha, and T. Finin, Automating cloud services life cycle through semantic technologies, Services Computing, IEEE Transactions on, vol. 7, no. 1, pp. 109122, 2014.
[5] K. P. Joshi and C. Pearce, Automating cloud service level agreements using semantic technologies, in CLaw Workshop, IEEE International Conference on Cloud Engineering (IC2E). IEEE Computer Society
[6] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, and D. Mladenic, Triplet extraction from sentences, in Proceedings of the 10th International Multiconference Information SocietyIS, 2007, pp. 812. http://plato.stanford.edu/entries/logicdeontic/
[7] O. Etzioni, M. Cafarella, D. Downey, A.M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, Unsupervised namedentity extraction from the web: An experimental study, Artificial intelligence, vol. 165, no. 1, pp. 91134, 2005.
[8] F. Ciravegna, 2, an adaptive algorithm for information extraction from webrelated texts, in In Proceedings of the IJCAI2001 Workshop on Adaptive Text Extraction and Mining. Citeseer, 2001.
[9] S. Soderland, Learning to extract textbased information from the world wide web. in KDD, vol. 97, 1997, pp. 251254.
[10] P. Cimiano, S. Staab, and J. Tane, "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 EuropeanConference on Principles and Practice of Knowledge Discovery in Databases, 2003.
[11] J. Cowie and W. Lehnert, Information extraction, Communications of the ACM, vol. 39, no. 1, pp. 8091, 1996.
[12] K. Barker and N. Cornacchia, Using noun phrase heads to extract document keyphrases, in Advances in Artificial Intelligence. Springer, 2000, pp. 4052
[13] K. Joshi and T. Finin, Ontology for cloud services SLA. Available: http://ebiquity.umbc.edu/resource/html/id/344
[14] K. Joshi, Ontology for services on the cloud. [Online]. A vailable: http://ebiquity.umbc.edu/resource/html/id/318/ OntologyforServicesontheCloud
[15] The stanford parser: A statistical parser. [Online]. Available: http://nlp.stanford.edu/software/lexparser. shtml
[16] Link grammar. [Online]. Available: http://www.link.cs.cmu.edu/link/
[17] Resource description framework (rdf). [Online]. Available: http:
//www.w3.org/RDF/ [18] Owl web ontology language. [Online]. Available: http://www.w3.org/ TR/owlfeatures/
[19] Sparql 1.1 overview. [Online]. Available: http://www.w3.org/TR/ sparql11overview/
[20] Mark Kerzner, Text Analytics, Big Data and Law, [Online]. Available: https://bol.bna.com/textanalyticsbigdataandlaw/
[21] General Architecture for Text Engineering (GATE) Project, https://gate.ac.uk/
[22] Quoc V. Le, Tomas Mikolov, Distributed Representations of Sentences and Documents, Proceedings of the 31 st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W and CP volume 32
[23] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz, Multidocument summarization by sentence extraction, In Proceedings of the 2000 NAACLANLPWorkshop on Automatic summarization Volume 4 (NAACLANLPAutoSum '00), Vol. 4. Association for Computational Linguistics, Stroudsburg, PA, USA, 4048.
[24] L. Thorne McCarty, Deep semantic interpretations of legal texts, In Proceedings of the 11th international conference on Artificial intelligence and law (ICAIL '07)
[25] John O. McGinnis and Russell G. Pearce, The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services, Fordham Law Review, VOL. 82, NO. 6, May 2014
[26] Making The Most Of Document Analytics, By Rand Ghayad, Paul Hinton, Mark Sarro and Michael Cragg, The Brattle Group Inc ., and David Cohen, Reed Smith , http://www.law360.com/articles/730189/makingthemostofdocumentanalytics, last retrieved 2/6/16
[27] Bureau of Justice Statistics, http://www.bjs.gov/index.cfm?ty=tp xsaqtid=30

[28] Travis D. Breaux, Matthew W. Vail, Annie I. Antn. Towards Compliance: Extracting Rights and Obligations to Align Requirements with Regulations. In Proc. IEEE 14th International Requirements Engineering Conference (RE'06), Minneapolis, Minnesota, pp. 4958, Sep. 2006

[29] Travis D. Breaux and Annie I. Antn. Analyzing Goal Semantics for Rights, Permissions and Obligations. In Proc. IEEE 13th International Requirements Engineering Conference (RE'05), Paris, France pp. 177186, Aug. 2005

[30] Chase Cardmember Agreement Link: https://www.chase.com/content/feed/public/creditcards/cma/Chase/COL00056.pdf

[31] Sudip Mittal, Karuna Pande Joshi , Claudia Pearce, and Anupam Joshi, Automatic Extraction of Metrics from SLAs for Cloud Service Management, In proceedings of IEEE International Conference on Cloud Engineering (IC2E 2016), April 2016

[32] Sudip Mittal, Karuna Joshi , Claudia Pearce, and Anupam Joshi, "Parallelizing Natural Language Techniques for Knowledge Extraction from Cloud Service Level Agreements.", poster at IEEE International conference on BigData, October, 2015

[33] Veselin Stoyanov, James Mayfield, Tan Xu, Douglas W. Oard, Dawn Lawrie, Tim Oates and Tim Finin, A ContextAware Approach to Entity Linking, Joint Workshop on Automatic Knowledge Base Construction and Webscale Knowledge Extraction, NAACLHLT, Montreal, June 2012.

[34] Tim Finin, Paul McNamee, Dawn Lawrie, James Mayfield and Craig Harman, Hot Stuff at Cold Start: HLTCOE participation at TAC 2014, 7th Text Analysis Conf., National Institute of Standards and Technology, Nov. 2014.

[35] Tim Finin, Dawn Lawrie, Paul McNamee, James Mayfield, Doug Oard, Ning Gao, Nanyun Peng, YiuChang Lin, Josh MacKin and Tim Dowd, HLTCOE Participation in TAC KBP 2015: Cold Start and TEDL, 8th Text Analysis Conference, NIST, November, 2015.

[36] Jennifer Sleeman and Tim Finin, Type Prediction for Efficient Coreference Resolution in Heterogeneous Semantic Graphs, Seventh IEEE Int. Conf. on Semantic Computing, Sept. 2013.

[37] James Mayfield, Paul McNamee, Craig Harman, Tim Finin and Dawn Lawrie, Kelvin: Extracting Knowledge from Large Text Collections, AAAI Fall Symposium on Natural Language Access to Big Data, Nov. 2014.

[38] Varish Mulwad, Tim Finin and Anupam Joshi, A Domain Independent Framework for Extracting Linked Semantic Data from Tables, in Search Computing Broadening Web Search, Stefano Ceri and Marco Brambilla (eds.), LNCS vol. 7538, Springer, 2012.

[39] Nikhil Puranik, A Specialist Approach for Classification of Column Data, MS thesis, Computer Science, University of Maryland, Baltimore County, 2012.

[40] Varish Mulwad, Tim Finin and Anupam Joshi, Semantic Message Passing for Generating Linked Data from Tables, 12th Int. Semantic Web Conf., Sydney, 2013

[41] Code of Federal Regulations [Online]. https://www.ecfr.gov/cgi-bin/ecfr?page=browse.

[42] Federal Acquisition Regulations System. [online]. available. https://www.ecfr.gov/cgi-bin/textidx?tpl=/ecfrbrowse/Title48.

[43] Stanford POS Tagger. [online] http://nlp.stanford.edu/lexparser.shtml.

[44] CarnegieMellonUniversityLinkgrammarparser. [online]. available: http://www.link.cs.cmu.edu/link/

[45] Element python library. [online]. available. https://docs.python.org/2/library /xml.etree.elementtree.html.

[46] TensorFlow Text Summarization. https://research.googleblog.com/2016/08/textsummarizationwith-tensorow.html.

[47] Tomas Mikolov et. al. Efcient estimation of word representations in vector space,https://arxiv preprint arxiv:1301.3781, 2013.

[48] Martin Abadi et. al. Tensorow: Large-scale machine learning on heterogeneous distributed systems, 2016.

[49] Gensim for Topic Modelling https://radimrehurek.com/gensim/ [50] O. Lassila, R. Swick and others, Resource Description Framework (RDF) Model and Syntax Specification, WWW Consortium, 1999.

[51] D. McGuinness, F. Van Harmelen, et al., OWL web ontology language overview,

W3C recommendation, World Wide Web Consortium, 2004.
[52] T. D. Breaux and A. I. Anton, Analyzing goal semantics for rights, permissions, and obligations, in RE05: Proceedings of the 13th IEEE International Requirements Engineering Conference (RE05), IEEE Computer Society, August 2005, pp. 177186
[52] T. D. Breaux, M. W. Vail, and A. I. Anton, Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations, in RE06: Proceedings of the 14th IEEE International Requirements Engineering Conference (RE06), IEEE Society Press, September 2006
[53] N. Kiyavitskaya, N. Zeni, T. D. Breaux, A. I. Anton, J. Cordy, L. Mich, and J. Mylopoulos, Automating the extraction of rights and obligations for regulatory compliance, in ER08: Proceedings of the 27th International Conference on Conceptual Modeling (ER08), Springer-Verlag, October 2008.
[54] Legal Dictionary, https://dictionary.law.com/