Cognitive Assistance in Government and Public Sector Applications
AAAI Technical Report FS-17-02

Cognitive Assistance for Automating the Analysis of the Federal Acquisition Regulations System

Srishty Saha, Karuna P. Joshi, Renee Frank
University of Maryland, Baltimore County, MD-21250, USA
Email:{srishty1, kjoshi1, rmfrank}@umbc.edu

Abstract

Government regulations are critical to understanding how to do business with a government entity and receive other benefits. However, government regulations are also notoriously long and organized in ways that can be confusing for novice users. Developing cognitive assistance tools that remove some of the burden from human users is of potential benefit to a variety of users. The volume of data found in United States federal government regulation suggests a multiple-step approach to process the data into machine-readable text, create an automated legal knowledge base capturing various facts and rules, and eventually building a legal question and answer system to acquire understanding from various regulations and provisions. Our work discussed in this paper represents our initial efforts to build a framework for Federal Acquisition Regulations System (Title 48, Code of Federal Regulations) in order to create an efficient legal knowledge base representing relationships between various legal elements, semantically similar terminologies, deontic expressions and cross-referenced legal facts and rules.

Introduction

People from various walks of life have an interest in United State government regulations. Individuals and organizations seeking seeking to sell goods or services to the government or obtain grants from the government need to find and comprehend the meaning of relevant regulations. Government employees themselves need to ensure that they have considered and are following all of the regulations governing their assigned tasks. Members of the public interested in government activities may wish to review regulations in order to guide their advocacy in support of their favorite cause.

While the Code of Federal Regulations[1] is available in electronic form on a variety of free and pay-wall sites, its 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 search capability against the text of the code can be helpful, but is dependent on choosing ideal or nearly ideal search terms. 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 data 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.

Given these challenges to finding the relevant sections of the regulations, we ask what automated analysis tools might be useful to aid human review and answer user questions about the text. Machine learning tools that are designed to find relationships between words seem to be a promising set to try against the regulatory code text. Much larger sections of the code can be processed quickly and those results returned to the human user without that user experiencing the fatigue and frustration that accompanies a less-precise keyword search.

The rules involving doing business with the federal government can be complicated. Having the ability to discern the rules that govern the acquisition of goods and services by gathering the sections that describe those rules is helpful to both providers and federal customers. Understanding how acquisitions progress from “requests for proposal (RFP)” to “contract award” and sometimes, “appeal” of an award helps prospective providers participate in trade with the government. Questions that might be asked about the federal acquisition process include: “How many days at a minimum must an RFP be posted/open/available for offers;” “What is the maximum number of days that an RFP may stay posted;” “How is a bidder notified of contract award;” or “How does a vendor appeal an award decision?”

In this paper, we intend to automate the analysis of legal documents like the Federal Acquisition Regulations System (Title 48, CFR)[2] by creating a framework which captures all possible vital legal elements and their relationships. 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 base. Our preliminary analysis of Federal Acquisition Regulations System, described in section 4, showed that Information Retrieval techniques and Deep Learning have shown promising results in extracting vital legal elements and resolving context disambiguation.

[1] https://www.ecfr.gov/cgi-bin/ECFR?page=browse

[2] https://www.ecfr.gov/cgi-bin/text-idx?/Title48

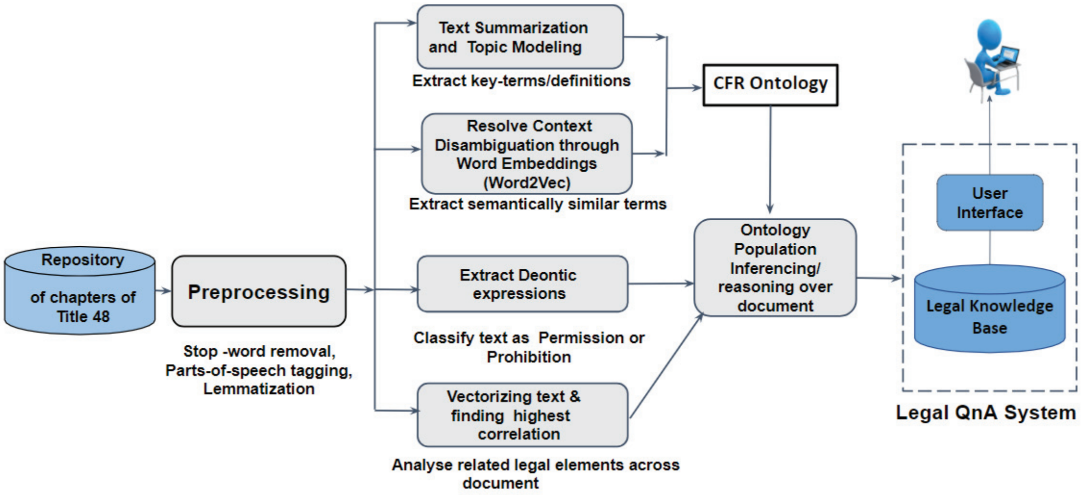
Text Summarization and Topic Modeling
Extract key-terms/definitions
Resolve Context Disambiguation through Word Embeddings (Word2Vec)
Extract semantically similar terms
CFR Ontology
Repository of chapters of Title 48
Preprocessing
Stop -word removal, Parts-of-speech tagging, Lemmatization
Extract Deontic expressions
Classify text as Permission or Prohibition
Vectorizing text & finding highest correlation
Analyse related legal elements across document
Ontology Population Inferencing/ reasoning over document
User Interface
Legal Knowledge Base
Legal QnA System

Figure 1: Proposed system architecture

Related Work

The Code of Federal Regulations (CFRs) 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. Representing long and complex legal documents like CFRs in an organizational structure has been an active area of research.In (Lau et al. 2004), they created a repository in XML format for several accessibility regulations as well as environmental regulations based upon CFRs in a tree hierarchical structure for extracting feature information such as definitions and measurements. But, it did not capture the mismatches between provisions that use the same phrases with different meanings in similarity analysis and did not capture relationships between various chapters and sections.

We aim to improve and automate the analysis and retrieval of relevant information from CFRs through semantic web technologies, Natural Language Processing (NLP), Information Retrieval and Deep Learning techniques. Our previous work [(Joshi et al. 2016); (Gupta et al. 2016); (Mittal et al. 2015)] focused on legal documents like Service Level Agreements (SLAs) of Cloud services where we had developed a semantically-rich ontology to capture key elements of cloud SLAs for modeling and reasoning about services related information. We extracted key SLA definitions and measures from these documents using pattern-based rules using the Stanford PoS Tagger[3] and CMU Link Parser[4]. As CFR titles such as the Federal Acquisition Regulations System(Title 48, CFR) are much longer and complex documents than Service Level Agreements, we need to improve and refine our existing approach for automating legal document text analytics. This paper describes our technical approach towards developing an efficient legal knowledge base which will capture all possible facts and rules of the Federal Acquisition Regulations System. This legal knowledge base will eventually contribute to building a legal question and answering system.

Methodology

This section describes various modules for our proposed framework for automating the analysis of the Federal Acquisition Regulation System (FARS). Figure 1 shows our proposed architecture. First, we created a repository of various chapters of Title 48 describing the Federal Acquisition Regulations System within the Code of Federal Regulations (CFRs) in machine readable form. Legal documents like CFRs are often long and semi-structured data containing tables and figures. The text portion from these documents is extracted using ElementTree python library[5]. Then, we preprocessed the extracted text using NLP techniques such as conversion to lowercase, removal of stop words, lemmatization and parts of speech tagging. For our analysis, we do not remove certain stop-words like “should” or “must” from the corpus as these might semantically refer to concepts like “prohibition”, “permission” or “authorization” rule which could be useful in resolving the issue of context disambiguation.

The next step is to build a framework for creating a legal knowledge base. Following are the modules which we aim to address:

- *Extract Key-terms to create an ontological representation of legal knowledge base.* In order to create an ontology for legal documents like CFRs, we need to first extract key-terms and definitions from the legal document. Figure 2 explains the process of extracting key-terms from the legal document. The Federal Acquisition Regulations System has 99 chapters, each chapter has various subparts

[3] http://nlp.stanford.edu/software/tagger.shtml

[4] http://www.link.cs.cmu.edu/link/

[5] https://docs.python.org/2/library/xml.etree.elementtree.html

and for each of those subparts, there are various sections and sub-sections. Extracting legally vital terms from such a long text document is a labor- intensive process. In order to automate the process of finding key-terms and definitions, we intend to use the concept of text summarization (Nallapati et al. 2016) and topic modeling (Blei, Ng, and Jordan 2003). For each section of each subpart for all chapters of Title 48, we first summarized the text using TensorFlow text summarization model[6]. Then, we implemented Latent Dirichlet Allocation model to perform Topic Modeling on the summarized text to extract top *k* topics from the section. These top *k* topics will form a set of vital key-terms for our ontology. After creating sets of vital key-terms and definitions, with the input from our legal expert and extracted vital key-terms from the corpus, we will create an ontology representing facts and rules contained in the Federal Acquisition Regulations System. We will use semantic web technologies such as OWL and RDF to manage the legal ontology.

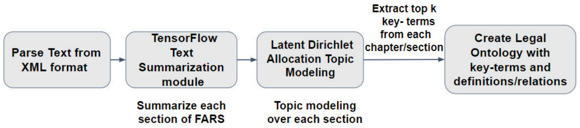
Parse Text from XML format
TensorFlow Text Summarization module
Latent Dirichlet Allocation Topic Modeling
Extract top k key- terms from each chapter/section
Create Legal Ontology with key-terms and definitions/relations
Summarize each section of FARS
Topic modeling over each section

Figure 2: Methodology to extract vital key-terms

- *Context Disambiguation and Word Vectorization phase*. Legal documents like CFRs contain contextually similar terms often leading to the issue of context disambiguation. We have implemented TensorFlow's Word2Vec deep learning architecture (Abadi 2016) and (Mikolov 2013) 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 contextually similar words. The results from word embedding will be used to populate and reason over the ontologies and creating knowledge graphs. Obtaining analogies and semantically similar words using deep learning architecture will help users in understanding variation in terminologies across various chapters of Title 48.
- *Extract related legal elements across various documents*. In CFRs, there are disparate sections containing facts and rules that might be applicable to answering a question. We intend to find related legal elements across various chapters and sections through vectorization and similarity metrics. We will first vectorize each section using the term frequency inverse document frequency metric (*tfidf*) and find correlation between vectorized sections using similarity metrics like Cosine similarity or Pearson Correlation Coefficient (Huang. 2008). This module is helpful in retrieving the most possibly relevant answers to a question.

Key-Terms
affiliate
architect
acquisiton
claim
pricing
federal-agency
procurement
solicitation
taxpayer
contract

Table 1: Key Terms extracted from the Federal Acquisition Regulations System

- *Extracting deontic expression from legal text.* In our previous work, we used text mining techniques to extract deontic rules from cloud SLA documents (Gupta et al. 2016). We will use this technique to extract deontic expressions from the Federal Acquisition Regulations System. For this initial phase, we will classify the deontic expression into basic two categories: Permissions and Prohibitions. This is helpful in answering generalized questions like "What are the responsibilities of a contracting officer?"- Answers to such questions should include Do's (permissions) and Don'ts (prohibitions) of a contracting officer.

Our research aims at building a legal question and answer system to analyze legal documents like the Code of Federal Regulations based on information extracted from the legal knowledge base.

Preliminary Results

In this section, we describe the preliminary results of our approach in automating legal document text analytics. We have used the Information Retrieval, Natural Language Processing and Deep Learning techniques to implement our framework. We have developed a system for parsing, preprocessing, extracting key terms and semantically similar terms, analyzing and reasoning over documents.

Extraction of Key Terms

Extracting vital key-terms manually from long-text documents like the Federal Acquisition Regulations System requires legal expertise and is a labor intensive and time consuming process. In order to automate this, we implemented TensorFlow's text summarization model for summarizing each section. Then implemented Latent Dirichlet Allocation (LDA) model for topic modeling to extract key-terms. Table 1 describes vital keywords extracted from few chapters of Federal Acquisition Regulations.

[6] https://research.googleblog.com/2016/08/text-summarization-with-tensorflow.html

Capturing semantically similar terms

We implemented TensorFlow Word2Vec word-embedding deep learning architecture to capture semantically similar terms across various chapters and sections. For our initial experiment, we provided some of the keywords identified by our legal expert about the Federal Acquisition Regulations System (Title 48, CFRs) to our learned model. The word embedding model has shown promising results. For example, for a query keyword like "rule", the words extracted from the model are "guidelines", "terms" and "regulations" all of which are semantically similar to each other. Table 2 describes results of our initial experiment on word embedding model.

Query Keyword	Analogous words
rules	terms, preamble, guidelines, regulations
publication	document, findings, survey, certification
compliance	comply, conformance, imposes, meets
claim	case, settlement, confirmatory, appealed

Table 2: Initial results of word-embedding model

Results from extracting vital key-terms and semantically similar terms will be used to create an ontological representation of a legal knowledge base used for reasoning over various chapters of Federal Acquisition Regulations System.

Discussion

We presented our research project on building a cognitive assistant for automatic analysis of the Federal Acquisition Regulations System (Title 48, CFR). We aim to build an ontological representation for a legal knowledge base describing vital facts and rules across Title 48. The long term goal of this project is to build a Legal Question and Answer system (Legal QnA) for various enterprises, federal customers and providers. It will be useful for the legal community as it provides an automatic way to analyze long and complex legal documents with less time and labor.

References

Abadi, M. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation, journal of machine learning research.

Gupta, A.; Mittal, S.; Joshi, K. P.; Pearce, C.; and Joshi, A. 2016. Streamlining Management of Multiple Cloud Services. In Proceedings, IEEE International Conference on Cloud Computing.

Huang., A. 2008. Similarity measures for text document clustering. Proceedings of the sixth new zealand computer science research student conference.

Joshi, K. P.; Gupta, A.; Mittal, S.; Pearce, C.; Joshi, A.; and Finin, T. 2016. Semantic Approach to Automating Management of Big Data Privacy Policies. IEEE BigData.

Lau, G. T.; Kerrigan, S.; Law, K. H.; and Wiederhold, G. 2004. An e-government information architecture for regulation analysis and compliance assistance.

Mikolov, T. 2013. Efficient estimation of word representations in vector space, https://arxiv preprint arxiv:1301.3781.

Mittal, S.; Joshi, K. P.; Pearce, C.; and Joshi, A. 2015. Parallelizing Natural Language Techniques for Knowledge Extraction from Cloud Service Level Agreements, IEEE BigData.

Nallapati, R.; Zhou, B.; dos santos, C. N.; Gulcehre, C.; and Xiang, B. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond, https://arxiv.org/abs/1602.06023.