ALDA : Cognitive Assistant for Legal Document Analytics

Karuna P. Joshi, Aditi Gupta, Sudip Mittal, Claudia Pearce, Anupam Joshi and Tim Finin
University of Maryland, Baltimore County, Baltimore, MD 21250, USA
Email: {kjoshi1, adigupta, smittal1, cpearce, joshi, finin}@umbc.edu

Abstract

In recent times, there has been an exponential growth in digitization of legal documents such as case records, contracts, terms of services, regulations, privacy documents and compliance guidelines. Courts have been digitizing their archived cases and also making it available for e-discovery. On the other hand, businesses are now maintaining large data sets of legal contracts that they have signed with their employees, customers and contractors. Large public sector organizations are often bound by complex legal legislation and statutes. Hence, there is a need of a cognitive assistant to analyze and reason over these legal rules and help people make decisions. Today the process of monitoring an ever increasing dataset of legal contracts and ensuring regulations and compliance is still very manual and labour intensive. This can prove to be a bottleneck in the smooth functioning of an enterprise. Automating these digital workflows is quite hard because the information is available as text documents but it is not represented in a machine understandable way. With the advancements in cognitive assistance technologies, it is now possible to analyze these digitized legal documents efficiently. In this paper, we discuss ALDA, a legal cognitive assistant to analyze digital legal documents. We also present some of the preliminary results we have obtained by analyzing legal documents using techniques such as semantic web, text mining and graph analysis.

Introduction

There has been an exponential growth in use of digitized legal documents. The majority of services on the Internet have associated legal documents such as Terms of Services, Privacy Policies and Service Level agreements. A large corpus of court cases, judgments and compliance/regulations are now digitally available for e-discovery. Large government and public sector organizations also face challenges in managing the legal legislations and statues that govern their day-to-day working. Furthermore, companies have to adhere to a variety of compliance and regulatory policies for many of these contracts, which are also increasingly digitally available. Managing and monitoring an ever increasing dataset of legal contracts, regulations and compliance is still a very manual and labour intensive job and can be a bottleneck in the smooth functioning of the enterprise.

Our research aims at building a Legal Question and Answer (LQnA) system that will be built upon large scale document analytics of legal documents using various techniques such as machine learning and text mining. We are working to transform legal databases from textual databases to graph-based datasets using Semantic Web technologies. Our long term goal is to develop a system that for any given action or question, can highlight all the statutes, 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 monitor for compliance.

This research will produce fundamental advances in areas of cognitive assistance technology. It will be beneficial to the public sector organizations, legal community as well as the business community who will be able to use this to significantly reduce the time needed to manage their legal contracts and compliance and regulatory needs. There are multiple User perspectives to using the ALDA system. From a lawyer's perspective, this application acts as a cog that searches, reasons and collates information from various legal documents, potentially draws inferences and presents it as a single interface. From the perspective of an end user in a government agency who broadly has a query if a planned action is legally compliant or not, the ALDA system can advise whether the action fits organizational policies / legal constraints or if it is an ambiguous area where the user should seek legal advice. For example a user may have a query like "Which service will not share personally identifiable information with third parties?". We envision that our system would be able to take this query in natural language as input, parse the information, link entities and extract the information from the corresponding knowledge graph. The information extracted would then be displayed to the user via the web-based system in form of text, graphs or tables.

Literature Review

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

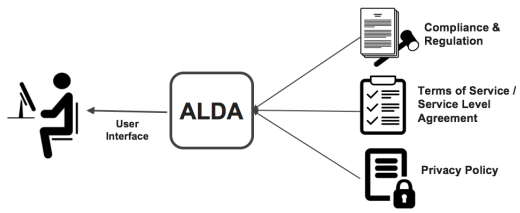

Figure 1: ALDA: Automating Legal Document Analysis

reveals patterns not visible from targeted manual review. This is critical not only for the data but also for the policies followed by service consumers or providers. The handling of heterogeneous policies is usually not present in a closed and/or centralized environment, but is an issue in the open cloud. One possible approach to this issue is to employ Semantic Web techniques for modeling and reasoning about services related information. We have used this approach for automating Cloud service level agreements (Joshi, Yesha, and Finin 2014; Mittal et al. 2015; 2016; Gupta et al. 2016). In one of our prior works, we described a new integrated methodology for the lifecycle of IT services delivered on the cloud, and demonstrate how it can be used to represent and reason about services and service requirements and so automate service acquisition and consumption from the cloud. We have divided the IT service lifecycle into five phases: requirements, discovery, negotiation, composition, and consumption. We detail each phase and describe the ontologies that we have developed to represent the concepts and relationships for each phase. We have described the five phases in detail along with the associated metrics in (Joshi, Yesha, and Finin 2014).

In (Rusu et al. 2007) the authors suggest an approach to extract subject-predicate-object triplets. They generate Parse Trees from English sentences and extract triplets from the parse trees. In (Etzioni et al. 2005) developed the KNOWITALL system to automate the process of extracting large collections of facts from the Web in an unsupervised, domain-independent, and scalable manner. They used Pattern Learning to address this challenge.

Researchers have explored the automated techniques for extracting permissions and obligations from legal documents using text mining and semantic techniques (Breaux and Anton 2005). (Kagal and Finin 2004) proposed a semantic web based policy framework to model conversation specifications and policies using obligations and permissions.

Technical Challenges and Approach

There are multiple computational challenges in automated analysis of legal documents which are often long and unstructured documents containing domain specific terminology. This section describes our technical challenges in building a cognitive assistant for analyzing legal documents. Following are the main technical challenges that we aim to address via ALDA:

- *Develop Ontological representation for legal documents.* We aim to compile a repository of variety of legal documents that outline the compliance and regulatory laws. We will analyze these documents, and with input from our legal consultants, start the process of developing ontologies to represent the facts and rules contained in these legal documents, thereby reducing the manual effort currently required for eDiscovery. As a first step towards this aim, we have significantly automated the process of managing and monitoring cloud Service Level Agreements (SLA) using semantic web technologies such as OWL, RDF and SPARQL.
- *Extract information from Legal Documents to create Knowledge-bases.* The next step will be to analyze the legal data corpus to extract facts and rules contained in them, represent them in the ontologies created, and build a knowledge-base of these facts and rules. This knowledge-base will be the foundation of our LQnA system. The system would reason over the facts and information in this knowledge-base to answer legal questions. In our system, we use text mining and natural language processing to extract relevant information from legal documents. We have built a system to automatically extract information such as definitions and measures from legal documents such as Service Level Agreements. In addition, we also used Modal and Deontic logic based grammatical rules to extract obligation and permission rules from these documents.
- *Develop techniques to analyze cross referencing among multiple documents.* In many legal situations, multiple disparate documents contain the facts and rules that are pertinent to answering a question. We will develop innovative techniques to represent and reason over legal documents that cross reference another set of documents and/or sections within the document. We will build on our prior work in cross document entity extraction and correlation to find these related legal elements, represent them as named sub-graphs in our knowledge-base and relate them by reifying on the subgraphs.
- *Use Deep learning techniques to extract semantically similar legal entities and terms.* Legal documents contain domain specific terminology, which may vary from one document to another. We aim to build a framework to extract semantically similar terms and entities across legal documents using word embeddings. We would train our word embedding models specifically for the legal document database, in addition to other public datasources such as Wikipedia. Obtaining semantic similarity for legal terminology would greatly enhance automation and generalization of legal document analysis.
- *Create compliance policies based on legal documents.* The next step after extracting rules and policies from legal documents is to create a system for compliance regulation. Components of consumer reports about cloud computing services such as encryption, sharing policy, etc. can be helpful for measuring and regulating compliance.
- *User's perspectives and understanding of digital legal documents.* Another essential component of building such a cognitive assistance technology is to understand the

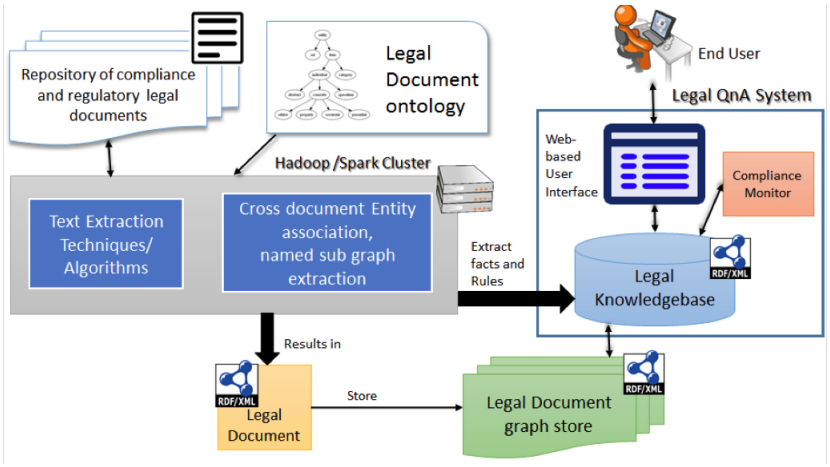

Figure 2: System Architecture of the proposed system to analyze legal documents.

user's perspective of digital legal documents which users accept / sign in their everyday life. We would be conducting large-scale user surveys and lab experiments for this task with the objective of understanding user awareness about their rights and obligations while using various services such as e-commerce, social media and cloud computing. The results of this study will also help us to evaluate the usability and accuracy of the automated system we are building for analysis of digital legal documents.

Preliminary Results

As a preliminary approach, we began developing techniques to automate the analysis and monitoring of legal documents such as Service Level Agreements (SLAs) of Cloud based services. Cloud services are increasingly being adopted by organizations to fulfill their IT needs because of their promise of cost savings, high availability and device and platform independence. The legal contracts between the cloud consumer and providers, like SLAs, terms of service document, privacy policy, etc., are used to define broadly the service data, delivery mode, service agent details, performance measures/metrics and cost of the service including penalty terms, if any. These documents are currently managed as text documents and so large manual effort is required to manage them as well as to map these to the main service performance indicators. With increasing use of cloud applications and services, the volume of such Cloud SLA documents is increasing exponentially. We have utilized semantic technologies in conjunction with text mining, natural language processing and machine learning for our research. Our preliminary analysis of legal documents such as Service Level Agreements, Privacy policies and Terms of Service documents for Cloud services have shown promising results. We have already built a prototype system for parsing, analyzing and reasoning over these documents.

Ontology Development and Knowledge Extraction

We have developed a framework to automate the acquisition and consumption of cloud based services and as part of that framework have developed a detailed ontology in OWL to represent cloud SLAs which is available in the public domain and illustrated in Figure 3. We have developed simple programs to extract cloud metrics from publicly available cloud SLAs and have published our preliminary results in (Joshi, Yesha, and Finin 2014). We have also conducted a study of all compliance and regulatory standards that are applicable to cloud services and have identified how multiple regulations can apply to a single cloud service.

To automatically extract key SLA definitions and measures from the legal terms of service documents, we began by retrieving publicly available SLAs or customer agreement documents that are posted by cloud providers on their website. We passed these documents through an Extractor module which used Pattern based rules like Stanford PoS Tagger [1] and CMU Link Parser [2] to automatically extract key term definitions and metrics from the document. The output generated from the Extractor was used by the Assessor module to evaluate it against our cloud SLA ontology. Once we identified the SLA definitions and terms, we saved it as a RDF graph which is machine understandable and hence can be used to automate the monitoring of SLA compliance of the service. Details can be found in (Mittal et al. 2015; 2016).

Extracting Permissions and Obligations

In our preliminary work, we have used text mining techniques to extract deontic rules from cloud SLA documents (Gupta et al. 2016). There are four basic types of deontic expressions: Permissions, Dispensations, Obligations and Prohibitions. Below are some of the grammatical rules used to extract Deontic expressions from legal documents using part-of-speech tagging and grammar analysis: After identifying the deontic expressions, the next two steps were: Categorize each deontic expression as either being a permission, prohibition, obligation or a dispensation. and Identify

[1] http://nlp.stanford.edu/software/tagger.shtml
[2] http://www.link.cs.cmu.edu/link/

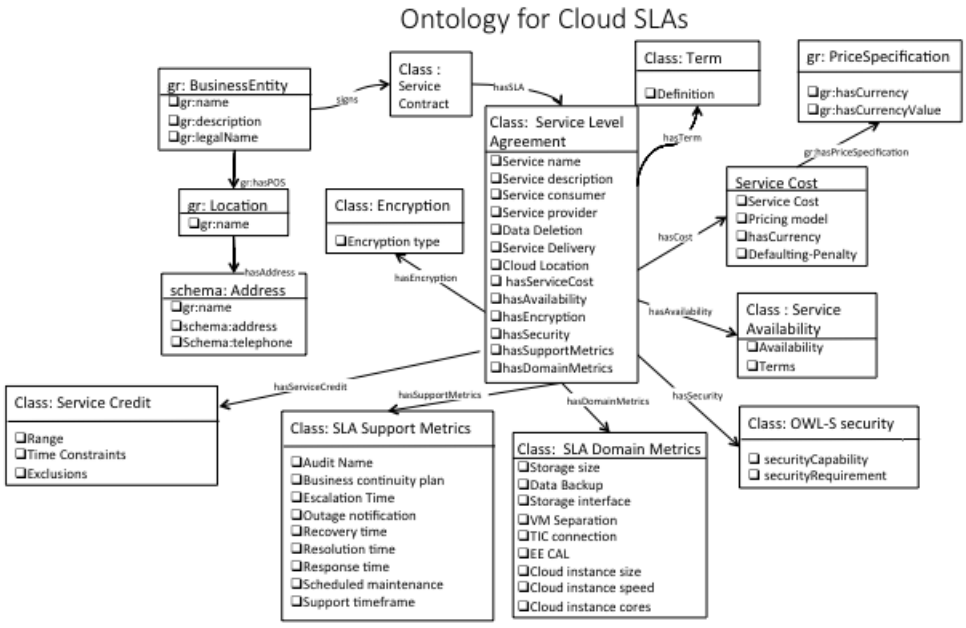

Figure 3: Ontology describing a Cloud Service Level Agreement (SLA)

the Actors for each of the deontic expression.

Positive Modalities:
<Noun / Pronoun> <modal verb> <verb>
Negative Modalities:
<Noun / Pronoun> <modal verb> <negation> <verb>

These are some of obligations and permissions extracted in our previous work:

- Obligation (Actor: Customer)
  "You must follow the procedure described herein within seven days of the end of the Claimed Outage."
- Obligation (Actor: Service Provider)
  "For each 30 continuous minute period of Qualifying Outage Minutes for a Service in a Measurement Period, [Service Provider] shall provide an SLA Credit of ..."

In case of Cloud SLAs we identified there are two main actors, the Customer or the Service Provider (service Itself). Identifying the actors is one of the important components for the rules in SLAs. For example identifying how many obligations and permissions are bound to a customer is useful information for customers in understanding what their rights and obligations are, and also in making a decision which service provider to choose.

Discussion

We presented our project ALDA, a legal cognitive assistance technology to analyze digital legal documents. This project will lead to development of a question and answer system built upon a comprehensive legal database. This research will be beneficial to the public sector organizations, legal community, the business community and the consumers at large. It will greatly benefit the legal community by enabling automatic review of a large set of legal documents in a very short time and with little manual effort. It will help to understand which compliance and regulatory standards will affect their work and decisions. The project will benefit the business community to understand which regulatory standard could affect their products and under what conditions.

References

Breaux, T. D., and Anton, A. I. 2005. Analyzing goal semantics for rights, permissions, and obligations. In *RE'05: Proceedings of the 13th IEEE International Requirements Engineering Conference (RE'05)*, 177–186. IEEE Computer Society.

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

Gupta, A.; Mittal, S.; Joshi, K. P.; Pearce, C.; and Joshi, A. 2016. Streamlining Management of Multiple Cloud Services. In *IEEE International Conference on Cloud Computing*, 8. IEEE Computer Society.

Joshi, K. P.; Yesha, Y.; and Finin, T. 2014. Automating cloud services life cycle through semantic technologies. *Services Computing, IEEE Transactions on* 7(1):109–122.

Kagal, L., and Finin, T. 2004. Modeling conversation policies using permissions and obligations. In *AAMAS 2004 Workshop on Agent Communication (AC2004)*.

Mittal, S.; Joshi, K. P.; Pearce, C.; and Joshi, A. 2015. Parallelizing natural language techniques for knowledge extraction from cloud service level agreements. In *Big Data (Big Data), 2015 IEEE International Conference on*, 2831–2833. IEEE.

Mittal, S.; Joshi, K. P.; Pearce, C.; and Joshi, A. 2016. Automatic extraction of metrics from slas for cloud service management. In *2016 IEEE International Conference on Cloud Engineering (IC2E 2016)*.

Rusu, D.; Dali, L.; Fortuna, B.; Grobelnik, M.; and Mladenic, D. 2007. Triplet extraction from sentences. In *Proceedings of the 10th International Multiconference" Information Society-IS*, 8–12.