An Information Infrastructure for Government Regulations

Gloria T. Lau, Shawn Kerrigan, Kincho H. Law
Department of Civil & Environmental Engineering
Stanford University
{glau, kerrigan, law}@stanford.edu

Abstract
The complexity and diversity of government regulations make understanding the regulations a non-trivial task. One of the issues is the existence of multiple sources of regulations and interpretive guides which are often independent of governing bodies. In this work, we propose an information infrastructure for regulation management and analysis, which includes a document repository and tools for similarity analysis and compliance assistance. Our corpus currently includes accessibility and environmental regulations, as well as selective supplementary documents from the Federal government and private organizations. A shallow parser is developed to consolidate different regulations into a unified XML format, which is well suited for handling semi-structured data such as legal documents. Important features, such as concepts, measurements, definitions and so on, are extracted and incorporated into the corpus by using handcrafted rules and text mining tools.

Information Retrieval (IR) techniques are employed to compare and locate similar or related provisions in different regulatory documents. Structural and referential information from regulations are used to further refine the similarity analysis. Compliance check is performed using a reasoning tool based on First Order Predicate Calculus (FOPC) logic. The compliance assistance system guides users through provisions using a question and answer interface. Examples of an e-rulemaking scenario for a rights-of-way draft and a compliance check procedure with a used oil regulation are shown to demonstrate current capabilities of the prototype system.

1 Introduction

Government regulations should ideally be understandable and retrievable with ease by practitioners as well as the general public. In reality, regulations are voluminous, heavily cross-referenced and often ambiguous. Multiple sources of regulations, for instance, from the Federal, State and local governments, amend and complement and potentially conflict with one another. There are many reference guides, that are published independent of governing bodies, attempting to help the public to better understand and comply with the regulations. The regulations, amending provisions and interpretive manuals together create a massive volume of semi-structured documents with potentially similar content but possible differences in format, terminology and context. An example of such complexity and conflict is shown in Figure 1 on design requirements of a curb ramp, where the Federal regulation focuses on wheelchair traversal, which is in conflict with the California regulation [1] (this provision is from the 1998 version) focusing on the visually impaired when using a cane [2].

ADA Accessibility Guidelines 4.7.2: Slope
Slopes of curb ramps shall comply with 4.8.2. The slope shall be measured as shown in Fig. 11. Transitions from ramps to walks, gutters, or streets shall be **flush** and **free of abrupt changes**. Maximum slopes of adjoining gutters, road surface immediately adjacent to the curb ramp, or accessible route shall not exceed 1:20.

California Building Code 1127B.5.5: Beveled lip
The lower end of each curb ramp shall have **a ½ inch (13mm) lip beveled at 45 degrees** as a detectable way-finding edge for persons with visual impairments.

Figure 1: Two conflicting provisions

In this work, we present a system that combines text mining and knowledge management techniques to help better manage, understand and analyze regulatory documents. The example domains include accessibility and environmental regulations. This paper first presents the development of a legal corpus with multiple sources of regulatory documents consolidated into a unified format. Extraction of important features, e.g., concepts, measurements and so on, is described in Section 2. Section 3 discusses the ongoing work on applying Information Retrieval (IR) and structural matching techniques to perform a similarity analysis between provisions, and preliminary results are drawn from the application on the e-rulemaking process. A regulation compliance assistance system follows in Section 4, where First Order Predicate Calculus (FOPC) logic sentences are implemented to help users to perform compliance check in a question and answer style. A brief summary and discussion on future works are given in Section 5.

2 Development of a Repository with Feature Extraction

In order to develop a prototypic system, this work focuses on accessibility and environmental regulations. For accessibility regulations, our corpus currently includes two Federal documents: the Americans with Disabilities Act Accessibility Guidelines (ADAAG) and the Uniform Federal Accessibility Standards (UFAS). In addition, Chapter 11 of the International Building Code, titled Accessibility, is included to reflect the similarity and dissimilarity between federal and private agency mandated regulations. Related sections from the British Standard BS8300 and the Scottish Technical Standards are also included to show the differences between American and European regulations. For environmental regulations, we currently covers US Code of Federal Regulations Title 40 (40 CFR): Protection of the Environment, along with selected supplementary and supportive documents that focus on regulations covering hazardous waste and the management of used oil.

Presently, regulatory documents are available in Hypertext Markup Language (HTML), Portable Document Format (PDF) or hardcopy. To ease the development of document analysis tools, we have chosen the eXtensible Markup Language (XML) as a unified format to represent regulations in our corpus because of XML's capability to handle semi-structured data. Figure 2a shows a schematic of our repository development process. A shallow parser is first developed to consolidate documents into XML format, as well as to extract feature information as discussed below. The hierarchical structure of regulations, as shown in Figure 2b, is preserved by properly structuring provisions as XML elements. For instance, Section 262.12(a) is a provision in Section 262.12, and thus is structured to be a child node of the XML element of Section 262.12.

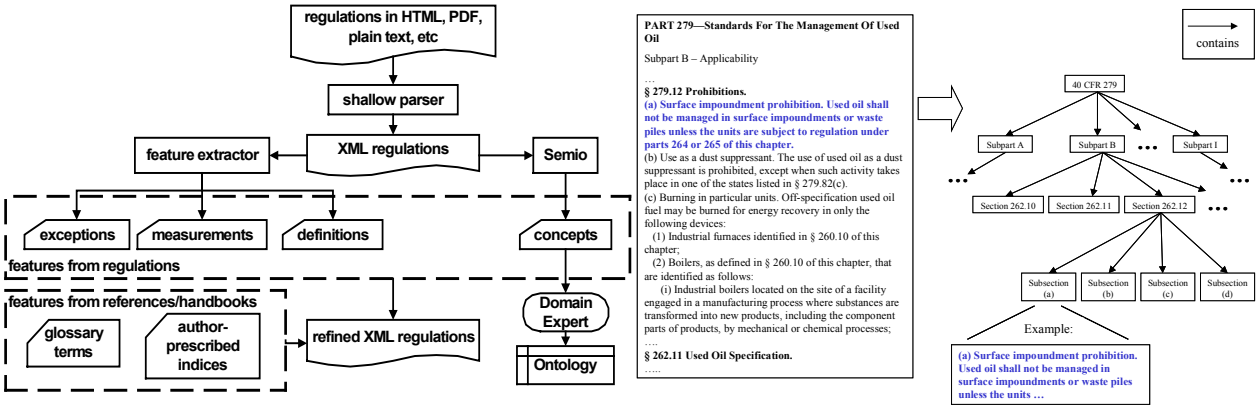

Figure 2: (a) Repository development with feature extraction, (b) Decomposition of reg into a XML tree

The example shown in Figure 1, where two provisions are in direct conflict, clearly demonstrates the need for a comparison system that pulls together related sections in regulations. It further amplifies the importance of conceptual information, such as key phrases in the corpus (e.g., "free of abrupt changes"), as well as domain-specific information such as measurements (e.g., ½ inch lip), for deep comparisons between provisions. However, traditional textual comparison techniques that employ simple term

matching, such as the Vector model [7], lack conceptual understanding of documents. They also suffer from the inflexibility to incorporate domain-specific information. Therefore, our comparison system, which is discussed in Section 3, combines conceptual information with domain knowledge. To enable this deeper comparison, the repository is refined with the extraction of features.

The process of feature extraction identifies the important features from the corpus that signal similarity or relatedness. Concept extraction is performed with the help of the software tool Semio Tagger [8], which is also used for a semi-automated concept ontology generation as shown in Figure 3a to help document retrieval. An ontology is developed based on the list of concepts extracted, and provisions are classified according to the ontology. For other features such as measurements and references, handcrafted rules are implemented to automatically match them in provisions [5]. The corpus of documents is refined with the extracted features tagged as additional XML elements in provisions where they appear. Figure 3b shows excerpts from a provision and its refined XML version that includes several features such as concept, index term and measurement.

Original Section 4.6.3 from the UFAS

4.6.3 Parking Spaces

Parking spaces for disabled people shall be **at least 96 in** (2440 mm) wide and … shall be part of an **accessible route** to the building or facility entrance and shall comply with **4.3** … **EXCEPTION**: If accessible parking spaces for vans ...

Refined Section 4.6.3 in XML format

```
<regElement name="ufas.4.6.3" title="parking spaces">
  <concept name="accessible route" num="1" />
  <indexTerm name="accessible circulation route" num="1" />
  <measurement unit="inch" size="96" quantifier="min" />
  <reference name="ufas.4.3" num="1" />
  …
  <regText> Parking spaces for disabled people ... </regText>
  <exception> If accessible parking spaces ... </exception>
</regElement>
```

Figure 3: (a) Ontology for accessibility regulations, (b) Example of XML structures and extracted features

3 Automated Extraction of Related Provisions

As is discussed above, related provisions are extracted by comparing regulations based on conceptual information as well as domain knowledge. In addition, specific structures of legal documents, such as the tree hierarchy of regulations in Figure 2b and the referential structure in Figure 3b, also represent useful information in locating related provisions. We employ a combination of IR techniques and document structure analysis to extract related provisions based on a similarity measure, which is defined as a similarity score between 0 and 1. Since typical regulations are massive in size, we take a provision as the unit of comparison, such as a comparison between Section 4.13.9 and Section 12.5.4.2. We first compute a base score between two sections by matching extracted features; the scoring scheme for each of the features essentially reflects how much resemblance can be inferred between the two sections based on that particular feature. For instance, concept matching is done similar to the index term matching in the vector model [7], where the degree of similarity of documents is evaluated as the correlation between their index term vectors. Using this vector model, we take the cosine similarity between the two concept vectors as the similarity score based on a concept match. Scoring schemes for other features are developed using the same idea.

The base score is subsequently refined by utilizing the tree structure of regulations. The parent, siblings and children (the immediate neighbors) of the interested sections are compared to include similarities between the interested sections that are not previously accounted for based on a direct comparison. In other words, similarities between the immediate neighbors imply similarity between the interested pair. The referential structure of regulations is handled in a similar manner, based on the assumption that

similar sections often reference each other. Therefore, after successive score refinements, similarities from both near-tree neighbors and references are identified, and related provisions are retrieved based on the resulting scores. Results obtained from the comparisons between different regulations are briefly illustrated in Figure 4 and described in [4]. Figure 4 shows two provisions focusing on door hardware that are identified as similar by our system. Due to the differences in American and British terminologies ("door hardware" versus "door furniture"), a simple concept comparison, i.e., the base score, cannot identify the match between them. However, similarities in neighboring nodes, in particular the parent and siblings, revealed a higher similarity between Section 4.13.9 of UFAS and Section 12.5.4.2 of BS8300.

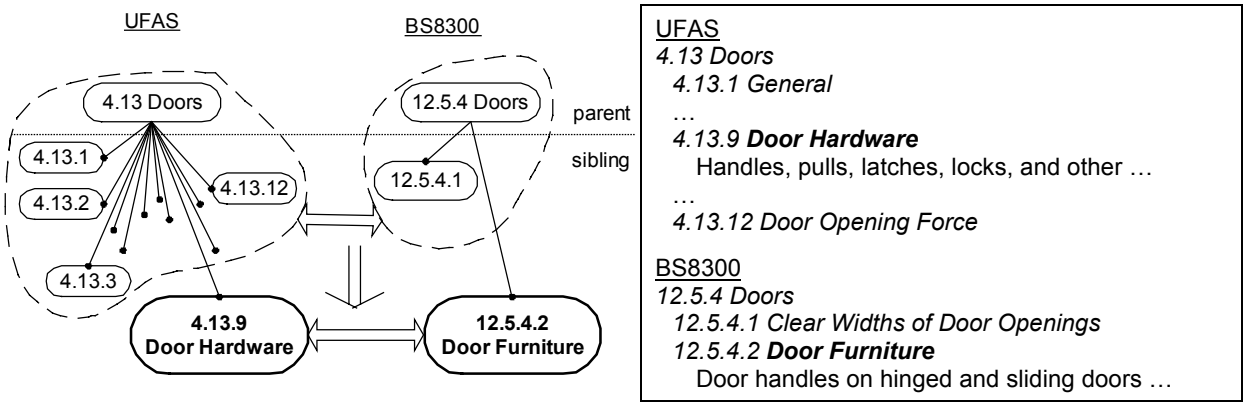

Figure 4: Example of a similarity analysis between American and British regulations

Besides the intended application on comparisons between regulatory documents, the prototype can also be applied to other domains as well, such as electronic rulemaking. Similarity analysis is performed on a recent e-rulemaking scenario on a newly drafted chapter for the ADAAG on rights-of-way access. Over a period of 4 months, the ADA Board received over 1400 public comments which total around 10 Megabytes in size for this 15-page draft. Based on the review of these public comments, the Board revises the proposed rules. The process of e-rulemaking generates a huge amount of data, i.e., the public comments, that needs to be reviewed and analyzed together with the drafted rules.

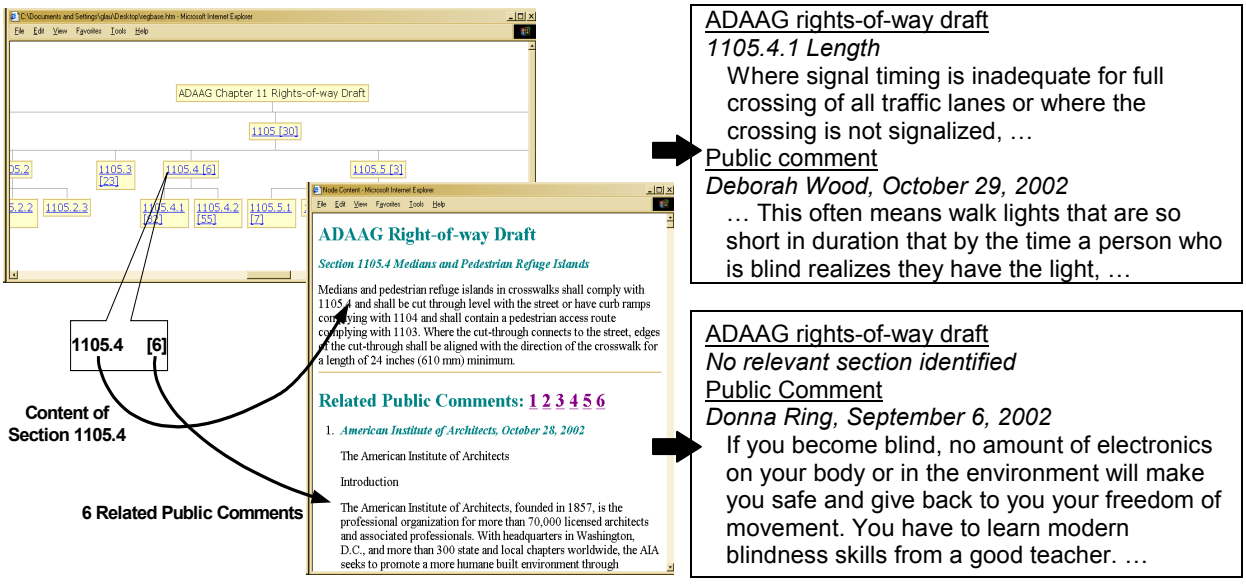

Figure 5: Application on e-rulemaking

We applied our system on this domain by comparing the drafted rules with the associated public comments. Figure 5 shows the generated output, where the drafted regulation appears in its natural tree

structure with each node representing sections in the draft. Next to the section number on the node is a bracketed number that shows the number of related public comments identified. Users, e.g., potential rule-makers and interested public parties, can follow the link to view the content of the selected section along with its retrieved relevant public comments. This prototype shows how a regulatory comparison system can be immensely useful in an e-rulemaking situation where one needs to review drafted rules based on a large pool of public comments. For instance, a typical pair of drafted section and its identified public comment, where both discussed about inadequate signal timing for pedestrian crossing of traffic lanes, is shown on the upper right of Figure 5. An interesting result is shown on the lower right, where a public comment is not latched with any drafted section by our system. Indeed, this reviewer commented on how a visually impaired person should practice "modern blindness skills from a good teacher" instead of relying on electronic devices, which is clearly not an issue covered by the draft.

4 Compliance Assistance using a Question and Answer System

After locating the provisions related to a certain project or user interest, there is still the question of compliance with the provisions and their implicit references to others. For compliance assistance, we add logic and control processing metadata to our regulation framework [3]. Regulation logic metadata represents a rule or concept from a regulation using FOPC logic sentences. These logic sentences are used to represent the rules that must be followed for an entity to be in compliance with the regulations. User interface logic metadata uses FOPC logic sentences to represent compliance questions and a list of possible user answers to those questions. In addition to regulation and user interface logic metadata, control processing metadata is implemented as well to provide information about what provisions of a regulation need to be checked for compliance. Each type of logic or control processing metadata can be associated with any regulation provision in the document. We employ Otter, a publicly available FOPC theorem prover developed at the Argonne National Laboratory, for logic check [6]. For the purpose of demonstration, a used oil regulation (40 CFR 279) has been manually tagged with regulation logic metadata, with user-interface logic metadata, and with control processing metadata.

A web interface asks users questions based on information in the XML logic metadata. Users may select a response from a menu of possible answers, including "Yes", "No" and "I don't know" options, where the "I don't know" option forks the compliance process along all possible answers. When the system completes a check against the regulation provisions or detects a conflict between the user's answers and the regulation, it displays a summary of the question-and-answer history as well as the compliance results. The use of and the results produced by the system are illustrated in Figure 6. The logs of the compliance session allow users to maintain a detailed compliance record which is useful for record keeping or when the regulations are to be revisited in the future.

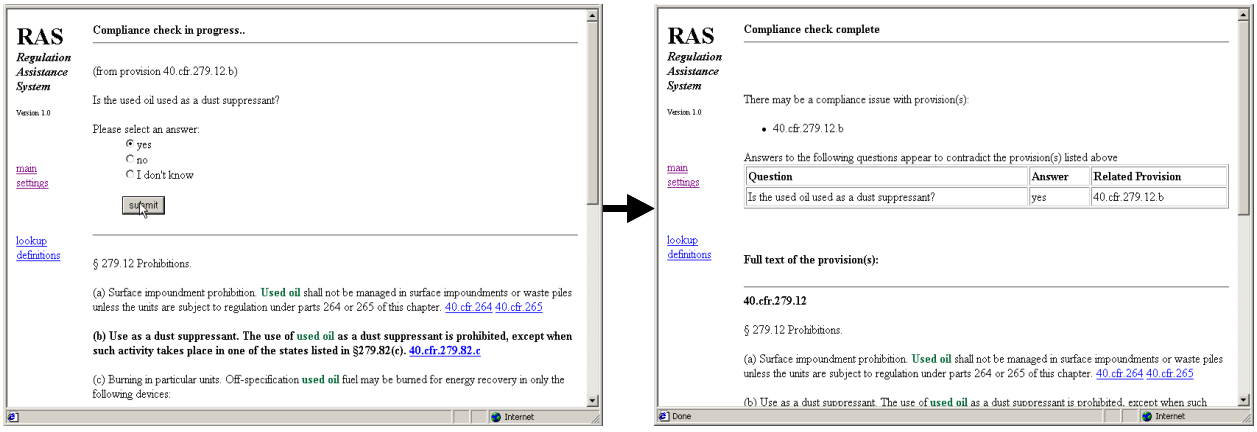

Figure 6: Example compliance-checking session

5 Conclusions and Future Tasks

In this paper, we present the development of a legal corpus, its associated similarity analysis, and a compliance assistance framework. A regulation repository is developed using XML as the standard, and our prototype includes several accessibility regulations as well as environmental regulations and supplementary documents. Tools have been developed for extracting feature information which include concepts, measurements, definitions and so on.

Similarity analysis, which combines IR techniques with corpus-specific document structure information, is shown to provide a reliable measure of relatedness between pairs of provisions. Potential application of our system on the e-rulemaking process is shown to help identify relevant drafted provisions and public comments. An interactive compliance assistance tool is developed by incorporating FOPC logic sentences and control elements to the XML structure. The compliance assistance system guides users through provisions and its implicit references as well as logging the answers for future reference. Limitations of our system include mismatches between provisions that use same phrases with different meanings in similarity analysis, and scalability issues that involve vocabulary consolidation in logic implementation for compliance check.

The goal of this project is to develop an information infrastructure to aid regulation management and *understanding* in e-government. Due to the existence of multiple sources of regulations and the potential conflicts between them, conflict identification becomes the natural next step to a complete regulatory document analysis. We plan to study the formal representation derived from structured texts to perform an automated analysis of overlaps, completeness and conflicts.

6 Acknowledgements

This research project is sponsored by the National Science Foundation, Contract Numbers EIA-9983368 and EIA-0085998. The authors would like to acknowledge an equipment grant from Intel Corporation. We would also like to acknowledge the support by Semio Corporation in providing the software for this research.

7 References

[1] *California Building Code (CBC)*. California Building Standards Commission, 1998.

[2] Gibbens, M.P. *California Disabled Accessibility Guidebook 2000*. Builder's Book, Canoga Park, CA, 2000.

[3] Kerrigan, S. and Law, K. Logic-Based Regulation Compliance-Assistance. in *Proceedings of the Ninth International Conference on Artificial Intelligence and Law (ICAIL 2003)* (Edinburgh, Scotland, 2003), 126-135.

[4] Lau, G., Law, K. and Wiederhold, G. A Framework for Regulation Comparison with Application to Accessibility Codes. in *Proceedings of The National Conference on Digital Government Research* (Boston, MA, 2003), 251-254.

[5] Lau, G., Law, K. and Wiederhold, G. Similarity Analysis on Government Regulations. in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, DC, 2003), 111-117.

[6] McCune, W.W. *Otter 3.0 Reference Manual and Guide*. Mathematics and Computer Science Division, Argonne National Laboratory, 1994.

[7] Salton, G. *The smart retrieval system - experiments in automatic document processing*. Prentice Hall, Englewood Cliffs, NJ, 1971.

[8] *Semio Tagger*. Semio Corporation, 2002. http://www.semio.com.