Semantic-ART: A Framework for Semantic Annotation of Regulatory Text

Krishna Sapkota[1], Arantza Aldea[1], David A Duce[1], Muhammad Younas[1] and René Bañares-Alcántara[2]

[1]Department of Computing, Oxford Brookes University, Oxford, UK
{k.sapkota, aaldea, daduce, m.younas}@brookes.ac.uk

[2]Department of Engineering Science, University of Oxford, Oxford, UK
rene.banares@eng.ox.ac.uk

ABSTRACT
Converting regulatory texts to machine interpretable models can enhance the automation of compliance management (CM) processes. The process poses serious research challenges as the information to be extracted from the regulatory texts comes from different regulatory bodies and is in different formats. In this paper, we present the main problems that we have faced in this area and how we have tackled them. Our proposed framework, Semantic-*ART,* considers the use of semantic annotation (SA) techniques to extract the regulations automatically.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: General

General Terms
Design, Experimentation

Keywords
Ontology, Information Extraction, Semantic Annotation, Text Analysis, Regulation

1. INTRODUCTION
Businesses and organizations must comply with requirements and expectations such as regulations, policies, mandates and guidelines to meet public standards and to avoid hefty penalties. Organizations must show that they comply with all the regulations by creating a set of compliance tasks. A system that helps to extract regulations automatically, will help to speed up the Compliance Management (CM) process, but extracting information from the regulations faces many challenges as the regulatory bodies use different formats to describe regulations. An example of a regulation for the pharmaceutical industry is depicted in Figure 2. The characteristics of the regulatory documents can be considered as both pros and cons. The pros are: 1) They have similar document format such as chapter, section and paragraph 2) They have common components such as subject, action and obligation 3) They are concerned with a domain, and the terminology is related to the domain. The cons are: 1) It is hard to interpret the meaning of the regulation in plain text 2) It is hard to deal with different formats 3) Low accuracy in the extraction of the regulation.

In order to face the problems derived from the different document layouts, we propose a semi-automatic specification of the relevant documentation format used for the regulatory bodies. The user must then help to identify how the documentation is formatted so the information extraction tool can follow a standard structure. Once the format has been identified, the challenge associated with the actual extraction of the regulation can be tackled by using ontologies. The domain ontology supports the identification of the meaningful words within the document, and the regulatory ontology helps to extract a regulation and its meaning automatically. Finally, the extracted regulation can also be linked to compliance tasks. As a result of our research in this area, a framework called *Semantic-ART* has been developed.

SA has been investigated for knowledge and information extraction in several approaches. The Cerno framework [3] extracted rights and obligations semi-automatically using TXL (the light weight pattern-matching programming language). Compared to the Cerno, our framework employs deep information extraction (IE) techniques, which help in extracting information more accurately. The popular tools such as Amilcare, MnM, Armadilo and SemTag have shown very inspiring results in SA [5,6], but they are based on machine learning approaches, thus require a massive training corpus. The work by Mokhtari and Corby [4] extracted deep concepts and their temporal aspects using contextual relation dependency for automatic rule creation. We believe that list of relevant words generation from an ontology would provide an easy and more feasible approach as compared to [4]. Date and place of birth of a person were extracted in [1] and the domain ontology was populated. However, the rules were specifically designed for the purpose, which makes its adaptation to the other domains difficult. Compared to the above, the proposed Semantic-ART framework provides a more flexible and adaptable approach as the rules are based on the concepts described in the domain ontology.

2. THE FRAMEWORK
The Semantic-ART framework, depicted in Figure 1, has been developed as a part of a larger research work towards developing a semantic compliance management system. It includes a semi-automatic process to extract regulatory information from text and convert it into semantic models. The objective is to create a domain dependent schema semi-automatically. The schema specifies the hierarchical structure of the text, which helps in the identification of salient components, for example, a paragraph imposing some regulatory constraints. The process comprises the following steps.

a) Pre Processing: The texts are available in various formats such as PDF, DOC, HTML and XML. Instead of developing processors for each format, our approach is to convert them into a single processing format – i.e., a standard HTML format.

ESAIR'11, October 28, 2011, Glasgow, Scotland, UK.
ACM 1978-1-4503-0958-5/11/10.

b) Semi Automatic Schema Generation: The schema is the definition of the structure of the text in the document, where the document components (e.g. chapter, section, paragraph as in Figure 2) are identified. The *Features Reader* identifies the font features such as style, weight, family and color, and based on these features, the *Structure Predictor* infers the structure of the document. The feature scores computed for each text component and their distribution in the document are analyzed to predict the structure. Numbered list (e.g. 5, 5.11, etc), text introducers (e.g. chapter 1, section 2, etc) and empirical values are also considered in the process. The user verifies the suggested structure.

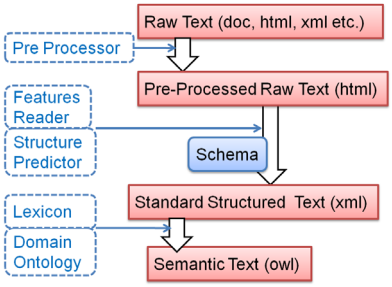

Figure 1. The Semantic-ART framework

c) Schema to XML Regulation: The schema, defining the overall structure of the document, is saved in an XML file for SA. Note that the previous two stages may be unnecessary if the regulators publish the documents in a standard format. However, this is not a common practice and those stages constitute an important part of the process.

d) Semantic Annotation: This is the core phase of the framework which extracts the regulatory constraints for the organizational processes. In the framework, a *regulation* is defined as a paragraph which imposes some restrictions to the processes. A *regulation* typically comprises at least one *statement* (a sentence in the regulation), which must have a *subject*, an *obligation* and an *action*. The *subject* is the entity on which the restriction applies; the *obligation* is the restriction type imposed; and the *action* is the process, which should be followed in order to comply with the *regulation*. The presence of an *obligation* in a paragraph helps to distinguish the *regulation* from the ordinary paragraph. Figure 2 illustrates an example of an annotated text component and the regulatory entities.

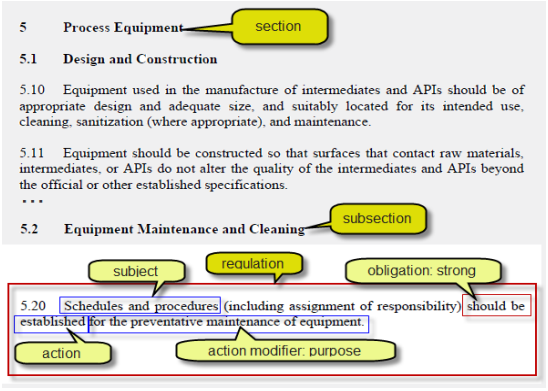

Figure 2. An excerpt from the Eudralex regulation

The text engineering platform, GATE [2], has been used to process the list of obligatory words (gazetteer) and their grammar (JAPE). For the *subjects* and *actions*, a domain ontology is used with the lexical ontology, WordNet. The domain ontology contains the relevant *subjects* and *actions*; WordNet helps to generate synsets (sets of words with similar meanings) of the root words of the ontological concepts. Missing concepts (i.e. *subjects* and *actions*) are identified by specifying the predicted grammar in JAPE. These are added in the domain ontology for the future annotation.

e) Semantic Regulation Generation: The extracted concepts (*subject* and *action*) and relations (*obligation*) are used to generate the retrieved regulation. Figure 3 shows how the regulation in Figure 2 was modeled as instances of the regulatory ontology concepts.

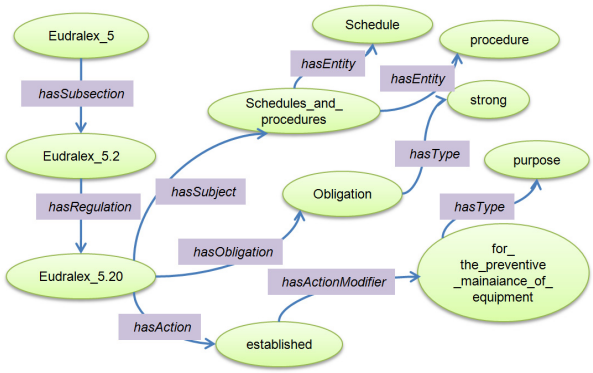

Figure 3. An excerpt from SemReg ontology modeling the concepts extracted from the regulatory text.

3. CONCLUSION

This paper presents the issues associated with extracting regulations from different documents and how to convert these regulations into machine interpretable models. We propose to tackle these problems by employing a domain ontology coupled with a structural schema for semantic annotation. The domain and regulatory ontologies have already been developing and we are in the initial stages of developing and testing our framework.

4. REFERENCES

1. Alani, H., Kim, S., Millard, D.E., et al. Automatic Ontology-Extraction from Web Documents. *IEEE Intelligent Systems 18*, 1 (2003), 14-21.
2. Cunningham, H., Maynard, D., Bontcheva, K., et al. Developing language processing components with GATE (a user guide). *University of Sheffield, Sheffield UK 5*, Gate 2 (2005).
3. Kiyavitskaya, N., Zeni, N., Cordy, J.R., and Mich, L. Cerno : Light-Weight Tool Support for Semantic Annotation of Textual Documents Introduction : The Semantic Annotation Challenge. *Data & Knowledge Engineering*, (2009), 1-39.
4. Mokhtari, N. and Corby, O. Modelling and automatic extracting of contextual semantic annotations. *Evaluation*, (2004).
5. Uren, V., Cimiano, P., Iria, J., et al. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web 4*, 1 (2006), 14-28.
6. Wimalasuriya, D.C. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science 36*, 3 (2010), 306-323.