Automated extraction of Information Elements from HIPAA Privacy Policies - Can we do away with annotation?

Richa Sharma and Vivek Joshi
TCS Research (Tata Research, Development and Design Center)
email: sricha@gmail.com, vivek.joshi3@tcs.com

Abstract

Compliance to policies, regulations, and laws is increasingly becoming important for software systems with increased pervasiveness of these systems in our daily life. These documents describe stakeholders' rights and obligations, in complex legal language. Manually analyzing these documents for extracting rights and obligations is an arduous and error-prone task. Earlier efforts to automatically analyze these documents suffer from the limitation of the need of manually annotated documents. In this paper, we propose human language technology based automated approach that does not require annotated documents for extracting information elements from regulatory documents. We present our preliminary investigation of the proposed approach on HIPAA privacy rules.

1 Introduction

Globalization of organizations is increasingly making it imperative for them to maintain adherence to policies, regulations, and laws for their business processes are now spanning across several geographies. These regulations and policies could be industry standards such ISO standards or government regulatory policies, or specific security and privacy policies. These documents lay down the rights and the obligations of the stakeholders involved along with the constraints under which rights and obligations hold valid. Rights, obligations and constraints – constitute the important *information elements* that must be identified and interpreted clearly for ensuring compliance to policies and regulations. However, owing to the legal nature of regulatory documents, organizations need to spend a lot in seeking expert advice, regulations review to ensure compliance at their end [1].

Analyzing regulatory documents manually to extract such information elements is both time and effort consuming, and could be error-prone too. Hence, approaches aimed at automatic extraction of information elements have been explored earlier [1], [2], [3] and [4]. However, most of these approaches require regulatory documents to be annotated for the present information elements. Annotation is also time-consuming, and often requires seeking expert advice. Our work in this paper is motivated by this intriguing question – can we do away with the annotation for extracting information elements from regulatory documents. We are of the view that the way experts analyze regulatory documents can be automated provided the documents are well-structured. Our preliminary investigation reveals that a good knowledge of structure of the document accompanied by semantic analysis of the document can support extraction of rights and obligations from unannotated regulatory documents. We have chosen to investigate HIPAA privacy rules from §164.520 in this work for two reasons, namely: earlier studies for these privacy rules exist for comparison [3], [6], and recent surge in web and mobile applications has aroused interest in privacy and security policies' study.

The rest of the paper is organized is: section 2 presents a brief overview of the related work done towards automated analysis of regulatory documents. We discuss in detail our approach, results and limitations in section 3. We finally conclude with future work in section 4.

2 Related Work

The legal nature and complex way of writing policies and regulatory documents makes verifying business process compliance a challenging task as discussed by Hashmi et al. [5]. Therefore, there has been a lot of interest in automating the process of compliance verification.

Some of the approaches consider logical approaches for compliance validation. Wael and Luigi propose UML-based Governance Extraction Model that validates logical expressions of enterprise rules against regulatory policies [7]. Kerrigan and Law propose first order predicate calculus based compliance assistance system [8]. While logical models provide sound validation mechanism, such models require human intervention or manually writing logical expressions from available regulatory documents. Oltramari et al. [9] have proposed ontology-based framework for representing annotated privacy policies where annotations are meant to indicate issues critical to users and/or legal experts.

Kiyavitskaya et al. [2] have proposed Gaius T tool based on annotated regulatory documents where annotations describe actors, rights, obligations etc. as suggested in [6]. The annotated documents are then parsed to deconstruct a rule statement to identify its components and constraints. Nair, Levacher and Stephenson [1] use handcrafted features for supervised classification to detect if regulatory statements represent obligation requirements or not, and then compliance entity extraction task determines to whom the detected requirements belong to. Engiel, Leite and Mylopoulos [3] have proposed modeling tool, NomosT for semi-automatic generation of law models from legal documents. NomosT supports identifying requirements from these generated law models. Papanikolaou [4] present in their work a tool for compliance validation in cloud. The tool processes semantically annotated regulation text to extract information with regards to cloud services from this legal text in order to ensure compliance against the agreed upon rules and regulations.

From the discussion of existing work for extracting information elements or requirements from policies and regulation documents, we find that automated processing of documents poses challenges because of complex nature of regulatory texts, and therefore annotation-based solutions have been explored so far. However, after annotating the documents, the only challenge that automated documents processing tools are left with is that of parsing. The regulatory documents are highly structured, and we feel that the well structured nature of these documents can be harnessed for automated processing. We propose our approach based on this observation, as discussed in the following section.

3 Proposed Approach

Our approach of extraction information elements from regulations comprises of two main steps, namely: (a) structural analysis, and (b) semantic analysis. We first present a brief overview of information elements present in regulation before discussing steps in our methodology.

3.1 Information Elements

The information elements that are important from the perspective of compliance validation are as follows as defined in [6]:

Right

A right is an action that a stakeholder is conditionally permitted to perform. Right describe what a stakeholder is eligible to do. For example - following is a statement of a covered entity's right as illustrated in §164.520 of HIPAA regulations:

A covered entity may provide the notice required by this section to an individual by e-mail.

Obligation

An obligation is an action that a stakeholder is conditionally required to perform. Obligation is an obligatory statement that a stakeholder must perform or is required to perform. Following is an example of an covered entity's obligation from §164.520 of HIPAA regulations:

The covered entity must provide a notice that is written in plain language and that contains the elements required by this paragraph.

Constraint

A constraint phrase is the part of a rights/obligation statement that describes a single pre-condition. For instance, in the obligation statement above, the phrases: *that is written in plain language* and *that contains the elements required by this paragraph* represent constraint on the *notice*.

In all of the above definitions, *stakeholder* is an entity that has been afforded rights and/or obligations in the regulatory documents.

3.2 Our methodology

Our methodology builds on the patterns suggested by authors in [6] for annotating the HIPAA policies. We have further added more patterns to the ones suggested in [6]. We arrived at these patterns after thorough manual analysis of HIPAA regulations. Our experience with related work on other documents served as a guide to identifying these additional patterns for rights and obligations used in our study as listed in Table 1:

Information Element	Pattern
Right	Has a/the right to
	Reserves a/the right to
	Retains a/the right to
	May
	Is permitted to
Obligation	Must
	Shall/will
	Is required to
	May not

Table 1: Patterns for rights and obligations

(a) Structural Analysis

As discussed in section 2, highly structured nature of regulatory documents can be harnessed for automated analysis of these documents, so first step in our approach is to conduct structural analysis of the text. A *major problem* with regulatory text is that most of the text is organized in the form of lists. Some list points are complete in themselves, constituting one paragraph such as §164.520(a)(1). On the contrary, some list points are complex, containing further sub-lists. For example: §164.520(a)(2) contains three sub-lists and sub-sub lists. Having studied structure of the privacy rules text of HIPAA regulations, we observe that combing each sub-list point to formulate a *paragraph* at first level of list (in §164.520, the first list level is designated by list points (a), (b) etc.) can enable further automated processing using patterns. The automated semantic processing takes each such constructed *paragraph* as input. To illustrate our proposed approach, let us consider excerpt from point (2) of §164.520(a):

> *(2) Exception for group of health plans.*
> *(i) An individual enrolled in… notice:*
> *(A) From the group health plan…or HMO; or*
> *(B) From the health insurance … health plan*
> *(ii) A group health plan…must:*
> *(A) Maintain a notice…section;*
> *(B) Provide such notice…health plan.*
> *(iii) A group health plan …under this section.*

These list points are processed algorithmically to formulate following five paragraphs in accordance to the list structure present:

(2) Exception for group of health plans. (i) An individual enrolled in… notice: (A) From the group health plan…or HMO; or

(2) Exception for group of health plans. (i) An individual enrolled in… notice: (B) From the health insurance … health plan

(2) Exception for group of health plans. (ii) A group health plan…must: (A) Maintain a notice…section;

(2) Exception for group of health plans. (ii) A group health plan…must: (B) Provide such notice…health plan

(2) Exception for group of health plans. (iii) A group health plan …under this section.

Such constructed paragraphs form the unit of processing for the next step of semantic analysis as discussed below. For instance, considering paragraphs at first level of list indicated by (a), (b) etc., §164.520(a) yields in a total of seven paragraphs.

(a) Semantic Analysis

In this step, each statement of the paragraphs is processed individually. The rights and obligations patterns presented in table 1 are used to extract corresponding rights and obligations phrases. In addition to these patterns, we further make use of constraint patterns to extract constraints. Table 2 illustrates the constraint patterns used in our study:

Information Element	**Pattern**
Constraint	<That is/verb-phrase ..>
	<enrolled..>
	<If/whether..>
	<with respect to..>
	<as defined..>
	<under .. section/paragraph.>
	<when ..>
	<required by.. section/paragraph>

Table 2: Patterns for constraints

Semantic analysis requires knowledge of the entities whose rights and obligations are to be extracted. For privacy rules of HIPAA in §164.520, there are 9 entities for which rights and obligations can be extracted. These entities are: *covered entity, individual, health plan, group health plan, health insurance issurer, covered health care provider, health care provider, health medical officer*, and *inmate*. We process each statement from its beginning, going left-to-right towards the end of the statement. A right or obligation is extracted by delimiting it between an entity and pattern for constraint (dropping the second delimiter after applying pattern), thus giving rise to following extraction pattern for rights/obligations:

<entity><rights/obligations pattern><constraint pattern >

Let us consider the paragraph in §164.520(a)(1), which is a simple paragraph with two statements:

S1: *Right to Notice.*

S2: *Except as provided … protected health information.*

S1 does not contain any pattern, and hence it is dropped from further processing, whereas S2 is processed for extracting the information elements:

Except as provided by paragraph (a)(2) or (3) of this section, an individual ***has a right*** *to adequate notice of the uses and disclosures of protected health information* ***that*** *may be made by the covered entity, and of the individual's rights and the covered entity's legal duties* ***with respect to*** *protected health information.*

This paragraph discusses right of an individual, extracted following the pattern for rights/obligations where the statement shows to contain the pattern:

<individual><has a right to..><that may..>

After applying the pattern for rights/obligations, the second delimiter <constraint> pattern is dropped to yield the right of the individual as:

an individual ***has a right*** *to adequate notice of the uses and disclosures of protected health information*

Our approach, thus, identify the entity who has been afforded the right or obligation. In addition, we get the following constraint phrase from this paragraph:

that *may be made by the covered entity, and of the individual's rights and the covered entity's legal duties* ***with respect to*** *protected health information.*

This phrase is further processed to find if it contains any more right/obligation or constraint. This remaining phrase yields in following two constraints in further processing:

C1: ***that*** *may be made by the covered entity, and of the individual's rights and the covered entity's legal duties*
C2: ***with respect to*** *protected health information*

The example from paragraph §164.520(a)(1) is a fairly simple example – complications arise with constructed paragraphs where possibility of duplication may arise. To illustrate these complexities and how we have overcome those, let us consider first two constructed paragraphs from §164.520(a)(2):

(2) Exception for group of health plans. (i) An ***individual enrolled*** *in a* ***group health plan has a right*** *to notice: (A) From the group health plan,* ***if,*** *and to the extent …or HMO; or*

(2) Exception for group of health plans. (i) An ***individual enrolled*** *in a* ***group health plan has a right*** *to notice: (B) From the health insurance issurer or HMO* ***with respect to*** *… health plan.*

In both of the above mentioned paragraphs, the first statement - *Exception for group of health plans,* is not further processed as it does not contain any relevant pattern. Rest of the statements in both the paragraphs is processed where following challenges are to met as:

1. For the statement - *An individual enrolled in a group health plan has a right to notice: (A) From the group health plan, if, and to the extent …or HMO; or*, it is difficult to associate the right to notice to either ***individual*** or ***group health plan*** as both of these are the entities to be considered in our processing. This challenge is overcome by considering the longer (in terms of length of words) right/obligation phrase as the finally extracted right/obligation assuming the shorter part would already be subsumed by the longer phrase. Similar argument holds for statement in second paragraph. Thus, the above two paragraphs yield two rights phrases, each for an *individual*:

   An individual enrolled in a group health plan has a right to notice From the group health plan, and

   An individual enrolled in a group health plan has a right to notice From the health insurance issuer or HMO.

2. Another challenge observed while processing these two paragraphs is that the constraint - *enrolled in a group health plan* is extracted twice. This challenge is fixed by removing duplicates, and thus counting this constraint only once. Similar challenge may arise with duplicate rights/obligations where such duplicates are removed to avoid any confusion.

In addition to rights/obligations and constraints extraction, we have also extracted cross-references using regular expressions for cross references. Following sub-section summarizes observation from our preliminary study on §164.520 of HIPAA.

3.3 Results

We present our preliminary results for the HIPAA privacy rules from §164.520. Following the methodology as discussed in section 3.2, we observe that our results are comparable to manual analysis study of the same article carried out in [6] and annotation based Gaius T tool [2], as presented in table 3 below:

System	Rights	Obligations	Constraints	Cross - Ref
Manual Analysis [6]	9	17	54	37
Gaius – T [2]	12	15	5	31
Our Approach	12	19	53	27

Table 3: Information Elements Extracted from §164.520

Our approach has been able to extract comparable counts of rights and obligations as compared with the ones obtained in manual analysis and by Gaius T tool. The number of constraints obtained by our approach is quite close to what has been obtained manually though Gaius T tool could extract only 5 constraints – much less than the manual counts of 54 for constraints. These observations are quite encouraging in terms of being close to manually identified information elements, indicating that annotation step may possibly be removed for rights/obligations extraction using human language technology. However, this is only a preliminary study and needs further exploration.

3.4 Limitations

Our approach relies on the presence of a well-formed and well-structured document. We do see limitation in our approach for the documents that are not well-organized. Currently, our approach suffers from the limitation of the structure of the statement as well, though we plan to overcome this limitation in future by parsing to correctly identify association between actors and their actions. An example of such a statement is present in §164.520(c)(3)(ii):

Provision of electronic notice by the ***covered entity will satisfy*** *the provision requirements paragraph (c) of this section* ***when*** *timely made in accordance with paragraph (c)(1) or (2) of this section.*

Here, the action ‘will satisfy’ is associated with ‘provision of electronic notice’ and not with the ‘covered entity’ yielding in incorrect obligation - *covered entity will satisfy the provision requirements paragraph (c) of this section.*

4 Conclusion

In this paper, we have presented our approach of extracting information elements, viz. rights, obligations, and constraints from HIPAA privacy rules in §164.520. The goal of our work was to find whether annotation of the policy text is really necessary or it can be avoided using human language technology since annotation is expensive in terms of time and effort, and is also subjective. Our preliminary study indicates that it is possible to do away with annotation with careful study of structure of the document. We further intend to improve upon our proposed approach in future.

References

[1] R. Nair, K. Levacher, and M. Stephenson, “Towards Automated Extraction of Business Constraints from Unstructured Regulatory Text”, *27th International Conference on Computational Linguistics: System Demonstrations*, pp. 157-160, 2018.

[2] N. Kiyavitskaya et al., “Automating extraction of rights and obligations for regulatory compliance.”, In: Li Q., Spaccapietra S., Yu E., Olivé A. (eds) *Conceptual Modeling - ER 2008*. Lecture Notes in Computer Science, vol 5231, 2008, Springer, Berlin, Heidelberg.

[3] P. Engiel, J. C. S. d. P. Leite, and J. Mylopoulos, “A tool-supported compliance process for software systems,” *11th International Conference on Research Challenges in Information Science (RCIS)*, Brighton, pp. 66-76, 2017.

[4] N. Papanikolaou, “Natural Language Processing of Rules and Regulations for Compliance in the Cloud”, In: Meersman R. et al. (eds) *On the Move to Meaningful Internet Systems: OTM 2012*. Lecture Notes in Computer Science, vol 7566, 2012, Springer, Berlin, Heidelberg.

[5] M. Hashmi, G. Governatori, H. P. Lam, and M. T. Wynn. Are we done with business process compliance: state of the art and challenges ahead. *Knowledge and Information Systems*, 57(1): 79-133, 2018.

[6] T. D. Breaux, M. W. Vail, and A. I. Anton, "Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with Regulations", *14th IEEE International Requirements Engineering Conference,* Minneapolis, pp. 49-58, 2006.

[7] W. Hassan and L. Logrippo, "Governance Requirements Extraction Model for Legal Compliance Validation," *Second International Workshop on Requirements Engineering and Law*, Atlanta, GA, pp. 7-12, 2009.

[8] S. Kerrigan and K. H. Law., "Logic-based regulation compliance-assistance", *9th international conference on Artificial intelligence and law (ICAIL '03),* pp. 126-135, 2003.

[9] A. Oltramari et al., "PrivOnto: A Semantic Framework for the Analysis of Privacy Policies", *Semantic Web*, 9: 1-19, 2017.