Towards Automated Extraction of Business Constraints from Unstructured Regulatory Text

Rahul Nair and **Killian Levacher** and **Martin Stephenson**
IBM Research - Ireland
rahul.nair@ie.ibm.com
killian.levacher@ibm.com
martin_stephenson@ie.ibm.com

Abstract

Large organizations spend considerable resources in reviewing regulations and ensuring that their business processes are compliant with the law. To make compliance workflows more efficient and responsive, we present a system for machine-driven annotations of legal documents. A set of natural language processing pipelines are designed and aimed at addressing some key questions in this domain: (a) is this (new) regulation relevant for me? (b) what set of requirements does this law impose?, and (c) what is the regulatory intent of a law? The system is currently undergoing user trials within our organization.

1 Setting

Large organizations spend considerable resources on reviewing regulations and ensuring that their operations, policies and procedures are compliant with the law. There has been a rapid growth in the number of regulations globally with more than 12,000 enacted/pending in 2016 compared to roughly 4,000 in 2008 (Compliance and Risks, 2016). In response, organizations are looking to make their compliance processes more responsive and efficient.

To better understand the existing workflows, we undertook a series of 14 interviews with compliance experts within our organization. Each expert is responsible for a certain class of legal requirements within a particular jurisdiction. For example, one of the experts was responsible for labeling requirements in North America, another for battery regulations in Chile etc.

Most experts have access to services that provide periodic briefs on regulatory changes. Once a new regulation is received, a legal review is conducted to broadly classify the document into the business categories that it most likely impacts. Many documents require translation before this legal review, since they originate from non-English speaking countries. Based on the labels, a more detailed review is carried out by each specific department. The review results in a list of requirements that need to be addressed. Compliance experts then map the requirements to current policies and evaluate if changes are needed for compliance. Changes recommended are handled by implementation teams.

The main pain points uncovered during the interviews where (a) too much (irrelevant) information from services that alert the experts on regulations, (b) lengthy and time consuming processes to determine whether specific products are relevant to the legislation (e.g. translation), (c) diverse set of regulations, some short (2 pages) and some long ($\sim$800 pages) with no way to prioritize them. The experts indicated interest in tools to help them "get from the law to 'action required' status" quickly, and distill requirements to "likely to impact us".

Natural language processing (NLP) applications in this domain are not new. Previous efforts have shown obligation extraction using semantic annotations (Kiyavitskaya et al., 2008), use of deep question answering architecture to evaluate compliance (Pasetto et al., 2013) and perform entity extraction using a domain ontology (Sapkota et al., 2012). We limit our review due to space restrictions.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 157–160
Santa Fe, New Mexico, USA, August 20-26, 2018.

In this paper, we present a system to extract business constraints from regulatory text[1]. We design three NLP pipelines aimed at addressing some of the challenges identified in the interviews. The pipelines address the questions (a) what set of requirements does a law impose (Section 2.2)? (b) what is the regulatory scope/intent of a law (Section 2.3)?, and (c) is this (new) regulation relevant for me (Section 2.4)?

2 Models

We focus on global regulations in the import-export area, which consists of laws related to batteries, labeling, electronic waste/product take back, emissions, energy efficiency, chemical and environmental legislation. Figure 1 shows the three pipelines. These are described next with the data used to train the system.

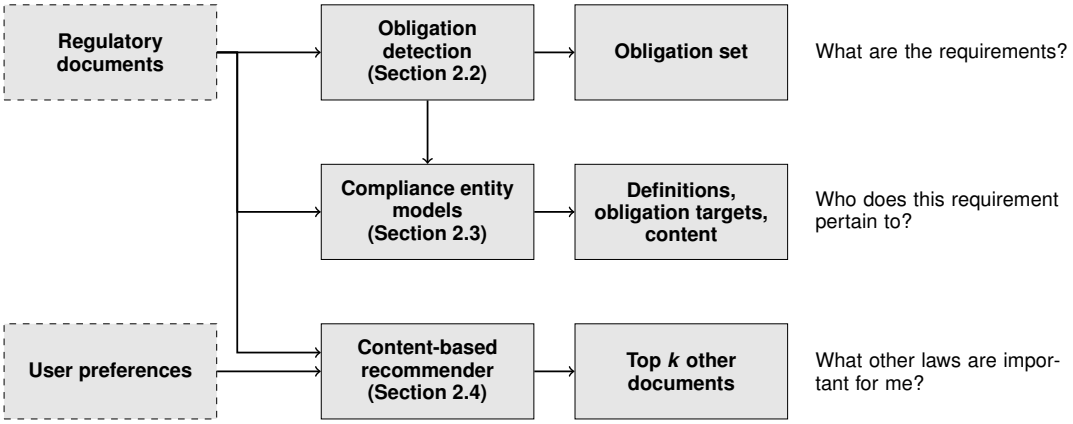

Figure 1: Overview of pipelines

2.1 Data

Our analysis is based on two primary data sources, as well as one semi-automatically annotated dataset. The primary source is an internal dataset, containing several thousand regulations from 168 jurisdictions, curated by legal experts since 2008. The dataset contains translations into English where necessary, along with manually generated summaries and product classifications. A second primary source is a set of requirements that have been manually extracted from these laws that impact one specific division within IBM. While this represents positive examples, we sampled from unrelated documents to generate negative samples for our classification task in Section 2.2.

Based on these primary sources, using a semi-automatic annotation process, the set of 129,313 obligations was parsed with a set of patterns. Patterns were manually extracted from a sample of obligations. The annotation process generated regulatory entity annotations (see Section 2.3) iteratively generating a corpus of 100,831 annotations covering 54,632 obligations.

2.2 Obligation detection

The *obligation detection* task seeks to classify input sentences into binary classes representing requirements or not.

Classical approaches involve using handcrafted features following by a supervised classification model. Features could include one hot encoding of words, distributed word representations, or TF-IDF vectors. More recently, deep learning architectures, such as LSTM's and CNN's have been shown to give good results. We experimented with these approaches. Of all the features, character n-grams performed the best along with TF-IDF vectors. Features derived from distributed word representations did the worst. Table 1 shows accuracy scores for the top performing pipelines. We chose to implement the random forest model with TF-IDF features and bigrams of character n-grams ($n = 3$). Character

[1]A video of the system can be viewed at `https://www.youtube.com/watch?v=Xt9j0qb_yT0`

n-grams improved classification scores over simply using TF-IDF vectors presumably by accounting for spelling variations.

Model	Features	Mean Acc.	Min Acc.	Max Acc.
Random forest	TFIDF-bigrams	0.930	0.912	0.943
Linear SVM - L1	TFIDF-bigrams	0.920	0.908	0.943
Random forest	TFIDF	0.919	0.885	0.941
Passive-Aggressive	TFIDF-bigrams	0.911	0.889	0.939

Table 1: Accuracy of top-4 pipelines for the obligation classification task (5-fold cross validation)

In our corpus, we found that obligation clauses tend to be longer than general descriptive text, on average roughly $\sim 50\%$. The system gathers user feedback on obligations clauses to improve classification scores over time.

2.3 Compliance Entity Extraction

The *compliance entity extraction* task seeks to determine to whom specific requirements pertain to. We distinguish two broad classes of entities (figure 2), (a) definition entities - within clauses that represent stakeholders or specific equipment, and (b) obligation entities - that are the legislative target of a specific clause. Within each class of entities, we further extract *targets* that is the specific entity, and *content* which describes the target. Taken together, the definitions and obligation targets, allows the system to determine if a specific obligation has a material impact on a business.

Extracting definitions is challenging for several reasons. Definition entities can be located in a dedicated section of laws, or scattered across different sections. This makes it difficult to assess the relevancy of a document. Definitions may be inconsistent across documents or even sections. For example, *manufacturers of batteries* in one document might be referring specifically to *manufacturers of zinc-carbon batteries*, while another document might be referring to *manufacturers of batteries located in the European Union*.

Extraction of obligation targets within clauses is challenging using traditional information retrieval approaches. These approaches can, at best, only return whether business entities are mentioned or not. This is usually insufficient to determine the legislative target. For example, consider the case of *manufacturers of batteries* in the following clause *"Distributors of lithium batteries should provide consumers with recycling services depending upon the recycling requirement stated by manufacturers of batteries"* where it is not the target.

The model consists of a perceptron algorithm trained using as an input, the sentences previously classified as an obligation (section 2.2). For each of the tokens in these obligation sentences, the entity extraction model is provided with, as an input, features consisting of each token's original string, lemma, part of speech, lower case string and shape (whether the token is a number, abbreviation, legal article number etc.). The algorithm was trained using a mini-batch approach with the semi-automatic annotated dataset described in Section 2.1 and produces labels in IOB format for each of the four possible annotation outputs described in figure 2. The model achieves a token-level accuracy of 0.95 across labels. An example of the four entities extracted by the model are shown in Figure 2.

"The OB importer and manufacturer of mercury-added button cell batteries must have OB the documents confirming the quantity of exported products. By DF mercury-added button cell batteries, this document refers specifically to DF non-rechargeable electrochemical batteries, which use a zinc electrodes DF in an alkaline electrolyte and retain a constant voltage of 1.35 Volts during discharge. Any OB equipment containing mercury-added button cell batteries should not be disposed in regular landfills due to their toxicity. It is the responsibility of the OB agent selling mercury based battery equipment to OB provide a suitable disposal alternative."

Figure 2: Clauses annotated by the entities model showing obligation OB targets and OB content along with definition DF targets and DF content

2.4 Content-based legislation recommender

Given (a) a corpus of regulatory documents and (b) user preferences for regulatory topics, the *recommendation* task seeks to determine the top k documents that best match user preferences. The purpose of this task is to aid compliance experts in regulation discovery.

The approach is driven by a simple domain insight gleaned during one of our interviews. Large legislative initiatives happen infrequently and result in large framework type documents. Over time, legislative bodies then issue amendments to fix issues with law or close loop holes. Compliance experts in turn reference a handful of these 'framework' type regulations often.

We therefore define a user profile as a set of documents, called a *user library*. For each document in the corpus, we generate a feature vector X based on TD-IDF vectors along with additional hand crafted features of one hot encoding of jurisdictions. Based on positive samples from the user library, and negative samples randomly sampled to have balanced classes, we train a user-specific linear SVM $y = f_u(X)$ to determine the separating hyperplane for a user. Given a set of new documents, the recommendation procedure sorts them by distance to the hyperplane and reports the top k documents.

We do not present a formal evaluation of this pipeline, since user trials are ongoing. However, we report on preliminary experiments with two 'framework' agreements, the REACH legislation (chemical restriction laws) passed in 2006, and the WEEE directives (waste electronics laws) passed in 2003 both in the EU. Using a set of documents that represent a hypothetical user library (the regulation along with guidance and explainer documents), the system recommended all the amendments to these laws within the top 20 recommended documents.

3 Architecture

The implemented system does a web crawl of authoritative sources for 8 jurisdictions and continuously updates the corpus. The platform relies on a mix of open and proprietary components to implement these pipelines. It is deployed for internal use on a kubernetes cluster on IBM Cloud, and scales easily on demand. The system is undergoing user trials with a panel of compliance experts within IBM. Feedback on annotation quality, document recommendation value and other user focused metrics are being gathered as part of this. Early feedback suggests improved ways to present information on the extracted entities and their definitions across documents.

4 Challenges and future work

Several technical challenges remain. Parsing of text from some document formats is unreliable, notably PDFs. Legislative documents come in varied formats, and occasionally are multi-lingual. Sentence structures of obligations are complex and it is unclear if pipelines, such as those presented here, readily transfer across various regulatory domains. Lastly, obligation clauses are open to interpretation.[2]

References

Compliance and Risks. 2016. Global growth of regulations. `https://www.complianceandrisks.com//public/growth_of_regulations_jan_2016.pdf`. Accessed: 2018-05-08.

Nadzeya Kiyavitskaya, Nicola Zeni, Travis D Breaux, Annie I Antón, James R Cordy, Luisa Mich, and John Mylopoulos. 2008. Automating the extraction of rights and obligations for regulatory compliance. In *International Conference on Conceptual Modeling*, pages 154–168. Springer.

Davide Pasetto, Hubertus Franke, Weihong Qian, Zhili Guo, Honglei Guo, Dongxu Duan, Yuan Ni, Yingxin Pan, Shenghua Bao, Feng Cao, et al. 2013. Rts-an integrated analytic solution for managing regulation changes and their impact on business compliance. In *Proceedings of the ACM International Conference on Computing Frontiers*, page 24. ACM.

Krishna Sapkota, Arantza Aldea, Muhammad Younas, David A Duce, and Rene Banares-Alcantara. 2012. Extracting meaningful entities from regulatory text: Towards automating regulatory compliance. In *Requirements Engineering and Law (RELAW), 2012 Fifth International Workshop on*, pages 29–32. IEEE.

[2]The authors thank Léa Deleris and Yufang Hou for reviews of an earlier draft.