Better Regulatory Compliance with Active Participation of Domain Experts

Sagar Sunkle, Deepali Kholkar, and Vinay Kulkarni,
Tata Consultancy Services Research,
54B, Hadapsar Industrial Estate, Pune,
India, 411013

Abstract

Regulatory compliance is at once a vital concern for modern enterprises, while also being extremely cost and effort intensive. The document-oriented and expert-reliant nature of regulatory compliance calls for natural language processing (NLP) and machine learning (ML) solutions to this problem. Most current NLP and ML approaches still mandate considerable involvement of the domain experts. We present an approach for regulatory rule identification using key entities and their mentions in a given domain in an active learning manner. We also propose how to arrive at formal specifications of regulatory rules thus identified. Our approach encourages active participation of the domain expert while reducing latent burdens on her as in other approaches.

1 Introduction

Modern enterprises face an onslaught of regulations in all of their dealings in most geographies [Alberth *et al.*, 2012]. Government and regulatory bodies tighten up existing regulations and routinely come up with new regulations aimed at limiting the moral hazard and the collective cost of support measures [Kaminski and Robu, 2015]. Transparency and accountability by enterprises and protection of customer are the dominant themes of this regime [English and Hammond, 2014]. Regulatory compliance is therefore a prime concern of higher management of enterprises and its enactment a vital requirement [Cau, 2014].

In spite of the stakes involved, enterprises fall behind compliance due to two aspects of compliance enactment- regulations are predominantly document-driven and the compliance is largely domain expert-reliant. The complexity of the legal texts and the continued reliance on the domain experts at various stages of compliance are the key reasons behind this.

The complexity of legal natural language (NL) texts compels the existing approaches to come up with targeted solutions [Sunkle *et al.*, 2016a; 2016b]. In particular, the NLP approaches use steps that include a) utilizing structural characteristics of legal NL texts by identifying sections of text at varying granularity from phrases to chapters and annotating cross references as in [Zeni *et al.*, 2015], b) manually transforming statements in legal NL texts to simplified forms such as *restricted natural language statements* as in [Breaux and Antón, 2005] for easier processing, and c) identifying language patterns like *juridical natural language constructs* as in [van Engers *et al.*, 2004]. ML approaches on the other hand have so far focused mainly on classifying the sentences/paragraphs from the legal texts into different kinds of provisions as in [Biagioli *et al.*, 2005] and [de Maat and Winkels, 2008], with underlying learning techniques that require training sets labeled by the domain expert, which is a cost and time intensive effort.

In most of these approaches, domain experts have to create intermediate artifacts specific to given approaches from the NL texts, or partake in creating labeled set of varying granularities of regulations. In contrast, we take a stance that it is possible to involve the domain expert in a manner that reduces additional burdens of creating intermediate artifacts while positing her as a teacher to a system that does the heavy lifting.

Unlike other ML approaches, our approach requires only a small, labeled training set from which an active learner learns to classify rules by interacting with the domain expert. Although we use standard active learning strategies, the key difference lies in our use of feature representation based on the *domain model* of regulations along with a *dictionary* of domain entities [Sunkle *et al.*, 2016a]. Our specific contributions are:

1. We present details of how informed active learning enables the domain expert to interact meaningfully with the system that learns to identify rules.
2. To obtain formal rule specification from identified rules, we propose two broad lines of investigation applicable to legal NL texts.

Our approach differs from the current state of the art in two ways. First, it focuses on regulatory rules at a generic level. It does not require knowledge of the textual structure of regulations, simplification of regulations into simpler NL forms, or annotations segregating the legal NL text into different provisions. Second, it uses a method of building a domain model (Section 3) to inform the learning process (Section 4), which is independent of a specific domain or a specific NL. Both the domain model building and rule identification is carried out in interactive sessions with the domain expert. She is relieved from any other task and only has to engage in terms

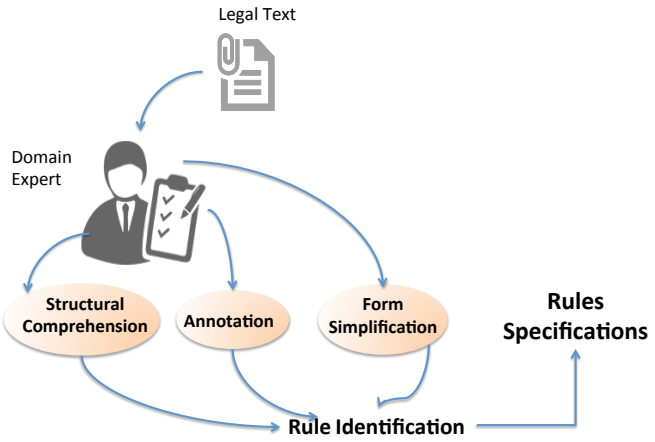

Figure 1: Role of Domain Expert(s) in Current Approaches

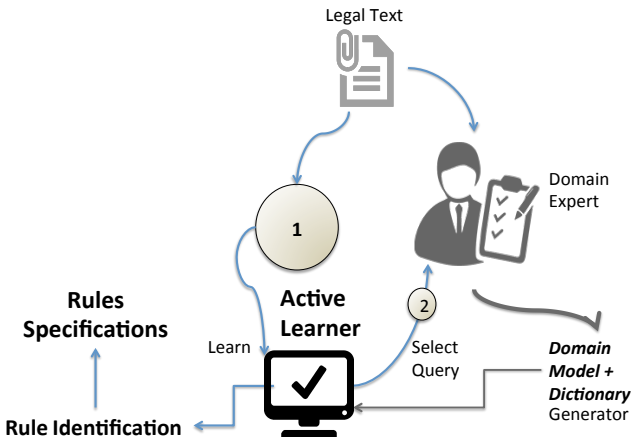

Figure 2: Role of Domain Expert(s) in Informed Active Learning

of responses to queries posed by the domain model generator and the active learner. We demonstrate this (Section 5) using Reserve Bank of India's latest Know Your Customer[1] (KYC) regulations.

A more detailed account of our domain model generation method is presented in [Sunkle *et al.*, 2016a], whereas the use of active learning for rule identification is discussed at length in [Sunkle *et al.*, 2016b]. This paper focuses specifically on the interactions of the domain expert with the domain model generator and the active learner.

We begin in the next section by motivating the use of active learning and presenting the technical overview of our approach.

2 Motivation and Technical Overview

Reliance on Domain Experts in Legal Context Existing NLP and ML solutions mandate considerable involvement of the domain expert. Many of these approaches require the domain experts to identify structural arrangements like chapters, sections, paragraphs, etc., specific to legal texts [van Engers *et al.*, 2004] (structural comprehension), to annotate legal texts to identify provisions [Biagioli *et al.*, 2005], conditional and other expressions [Wyner and Peters, 2011], parsing patterns [de Maat and Winkels, 2008], and various other aspects specific to given approaches (annotation), and to simplify complex legal sentences and making them amenable to analyses [Breaux and Antón, 2005; Kiyavitskaya *et al.*, 2008; Zeni *et al.*, 2015] (form simplification).

Legal texts are inherently more complex than other NL texts. It is not unusual for them to contain extremely long sentences with complex clauses. Another feature of them is the use of a number of lists representing characteristics of norms and their applicability in specific conditions with modalities [Kiyavitskaya *et al.*, 2008]. Several cross references may occur throughout the text whereby various details of a norm are scattered throughout different chapters/ sections/subsections [Wyner and Peters, 2011]. The updated regulation texts may contain amendments in terms of changes to the definitions of norms over time in terms of exceptions, and variety of repeals and amendments are often placed in supplementary annexes [Boella *et al.*, 2012]. From NL processing point of view, there are other bottlenecks like the mixed usage of active and passive voice in the legal texts, which NL-driven approaches may have to take care of. Different regulations differ in terms of their structures, ways in which provisions are specified and amended, and the NL they are encoded in. Due to these variations, even if targeted NLP and ML solutions were obtained, their applicability remains potentially restricted.

In a nutshell, using existing NLP and ML approaches; the domain experts must create several artifacts to encode their knowledge. They have to identify the specific regulations. Then using the artifacts generated earlier, they need to specify the regulations in given rules specification language [Sunkle *et al.*, 2016b]. This is shown in Figure 1.

We propose to NOT consider the syntactical specifics of NL texts for rule identification. Instead, we gather key entities and their mentions from the text and use them to make the rule identification task easier for the learner to learn. Once rules are identified, we propose that an active semantic parser transforms NL rule sentences into logical specification. It is to this representation, that we defer the treatment of segmentation of rules and their compositions etc. This paper only focuses on domain expert's role till rule identification.

Our Approach In contrast to the current approaches, we rely on an ML system to do the *heavy lifting* of identifying rules [Sunkle *et al.*, 2016b]. This is illustrated in Figure 2. The learning process of the active learner is driven by the domain model and the dictionary that the generator has produced making the learner *informed* [Sunkle *et al.*, 2016b].

The domain model generation thus becomes the key activity performed *before* teaching the learner to identify the rules. We make no assumptions about the legal text as in other approaches described above, rather we use two peculiarities prevalent in any legal text that it often contains definitions of key entities referred in the regulations and throughout the text of regulations there is a proliferation of *instances* and *refer-*

[1] https://rbi.org.in/scripts/BS_ViewMasCirculardetails.aspx?id=9848

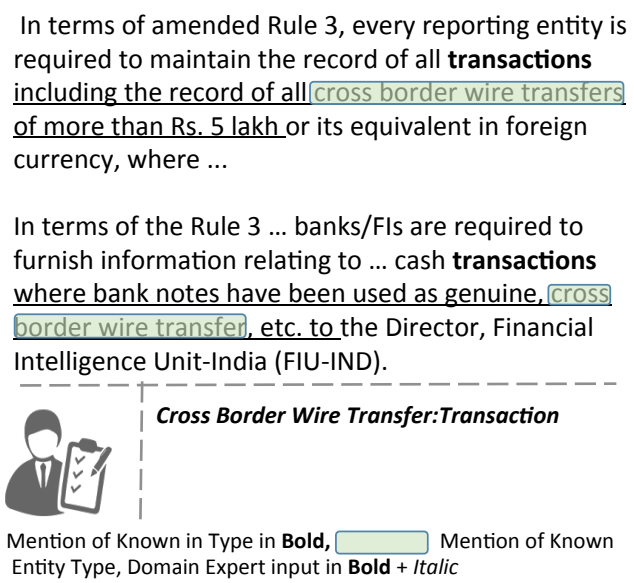

Figure 3: Clusters Identifying Mentions of Known Entity Type *Transaction*

ences of key domain entities and relations, which we collectively refer to as *mentions*.

The domain expert provides *seed* entities and their mentions from the definition section to the domain model and dictionary generator. The other activity that she needs to participate in is in responding to the queries of the active learner indicated by (2) in Figure 2.

The active learner learns by processing sentences from the legal text shown by (1) in Figure 2, which is shown in a much larger circle to indicate an unlabeled pool. (1) shown as a smaller circle in Figure 2 compared to (2) is a labeled training set of sentences, the classes of which the learner queries the domain expert.

It can be seen that in our approach, the domain expert is relieved of creating other artifacts for identifying and separating regulations from the rest of the text in the legal NL text. The learner learns not from the artifacts but from the sentences from the legal text directly without any structural comprehension, simplification, or annotations [Sunkle *et al.*, 2016b]. This is possible due to the feature representation used by the learner, which builds on the concepts captured in the domain model and their mentions in the dictionary [Sunkle *et al.*, 2016a].

Once the rules are identified as described in Section 4, we propose that it might be possible to use semantic parsing techniques to obtain logical specifications from the identified rules. If statistical techniques are used, then the effort needed to identify syntactical characteristics of regulations in each identified rule can also be potentially reduced.

3 Generating Domain Model and Dictionary

In our approach the domain expert provides seed domain model, i.e., entity types and relations immediately available

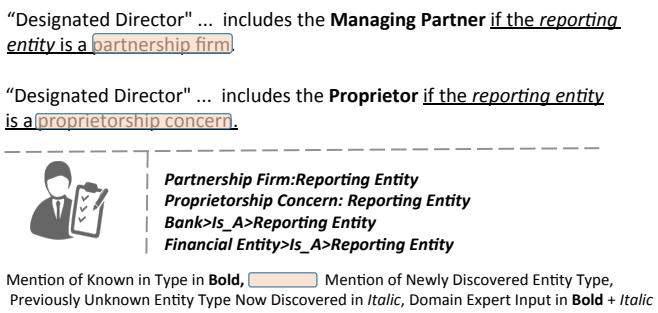

Figure 4: Clusters Identifying Mentions of Unknown Entity Type *Reporting Entity*

from the *definitions* section of the legal NL text.The domain model generator uses context-based clustering. The idea behind this technique is that the contexts, i.e., spans of texts, around the mentions of various domain entities are important and could be clustered to extract useful information, in our case, other entity types not covered in the definitions sections and mentions of all entity types so far known.

In the following, We refer to Indian KYC regulations. KYC regulations aim to prevent money laundering (ML) and financing of terrorism (FT). To kickstart clustering, we use as seeds, the entity types and mentions available in the *definitions* section. For instance, from the definitions section of KYC, we find the definitions of Customer, Designated Director, Document (that the customer has to submit to the reporting entity like a bank), and Transaction, which we take as seed entity types. This section in KYC also provides mentions, in terms of sub-type entities, instances, and synonyms, of these entity types. We find 12, 4, 33, and 15 mentions respectively of these entity types. These are generally very easy to spot. Interested readers are invited to look at the KYC definitions section at the link shared earlier. In most financial services regulations that we have encountered apart from KYC, such as MiFID II[2], we have found that definitions of key concept types are provided clearly along with their subtypes and terms with which they are referred to in the text.

Distributional Semantics through Context Clustering In order to find entity types that could be part of the domain model but not yet known, i.e., not in the definitions section specifically, we use mentions of entities that we have so far found. We use a hypothesis known as *distributional semantics*[Harris, 1968], which suggests that counting the contexts that two words share improves the chance of correctly guessing whether they express the same meaning, in other words, semantically similar expressions occur in similar contexts [Baroni and Lenci, 2010].

We cluster the contexts, i.e., n characters to the left and right of mentions of each entity type so far known, to suggest to the domain expert, what looks like other possible mentions. This is illustrated in Figures 3 and 4.

Interactions between Domain Expert and Domain Model Generator The domain expert either adds to the dic-

[2] http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32014L0065&from=EN

tionary, a new mention of a known entity type as in the case illustrated in Figure 3 or as in the case illustrated in Figure 4 has the option to add a new entity type along with the mention(s), if she recognizes that the mention(s) refers to different entity type not in the current set of known entity types. In Figure 3, we illustrate how clustering the contexts of mentions of the entity type Transaction reveals a mention *cross border wire transfer* which the domain expert deems to be of the same entity type Transaction. In Figure 4, we illustrate how clustering the contexts of mentions of the entity type Designated Officer reveals the mentions *partnership firm* and *proprietorship concern*, which the domain expert adds along with previously unknown type Reporting Entity of which they are mentions.

A simple input syntax enables the domain expert to indicate whether the mention is of known or unknown entity type, and add the unknown entity type and its relation to known entity types. This is shown in Figures 3 and 4. The syntax followed is ***mention:entity type*** for adding the mention, and ***entity type1<relation<entity type2*** for relating the newly added entity type to the existing entity types. Note that relations obtained from the definitions section of regulations and relations entered by the domain expert are not used in active learning but could be useful when authoring rules as proposed in Section 5.

4 Informed Active Learning

Our choice of active learning is motivated by the fact that unlike supervised learning, active learning is able to learn from very less number of labeled sentences, by querying the domain expert on possible classes of a sentence [Settles, 2009].

Interactions Between the Active Learner and the Domain Expert In our case, the features of the learner are represented in terms of domain model entities and their mentions from the dictionary. The domain expert interacts with the learner by providing correct classification via console-based input when the learner needs guidance over the classification that it has come up with.

Informed Active Learning Strategies We use two strategies for informed active learning, namely uncertainty and certainty sampling. In uncertainty sampling or *query by uncertainty*, the learner queries the learning instances for which the current hypothesis is least confident [Olsson, 2009]. The measure of confidence is the conditional probability estimate for a given category obtained from the underlying classifier model, in our case LogisticRegressionClassifier. The query strategy of certainty sampling is the opposite of uncertainty sampling. Instead of querying classes of instances where the learner is least confident, it queries for instances where it is most confident. The results of these strategies are indicative of the performance of the active learner.

Implementation Details We use the LingPipe[3] toolkit for processing text for context-based clustering to discover unknown entities as well as known and unknown mentions. In order to identify known mentions in the text, we use an implementation of approximate dictionary chunker ApproxDictionaryChunker in LingPipe. For clustering, we use SingleLinkClusterer from LingPipe, which implements standard single link agglomerative clustering, in which pairs of clusters are merged together in a bottom up fashion.

We use a specialized FeatureExtractor[4] from LingPipe called ChunkerFeatureExtractor. A feature extractor provides a method of converting generic input objects into feature vectors. A ChunkerFeatureExtractor implements a feature extractor for character sequences based on a specified chunker. This arrangement helps us in uniquely representing features in terms of entities and their mentions from the domain model and the dictionary respectively. By using ChunkerFeatureExtractor and supplementing it with the domain model and the dictionary via ApproxDictionaryChunker, it becomes possible to influence the features the learner will create as it learns from the sentences by querying the domain expert.

To implement an active learner for rule identification, we use LogisticRegressionClassifier from LingPipe. It is a scored classifier that provides conditional probability classifications of input objects. It uses an underlying logistic regression model and feature extractor which in our case is the ChunkerFeatureExtractor. We implement the prototypical active learning algorithm from [Olsson, 2009].

5 Results and Discussion

5.1 Applying Context Clustering to KYC Text

We copy pasted the text of KYC from the link shared earlier. We use LingPipe's IndoEuropeanSentenceModel to split the text into sentences. We obtained 525 sentences. From the *definition* section, we obtained 4 entities. We get 4 more entities and their mentions using context clustering [Sunkle *et al.*, 2016a]. We only specify 5 mentions of entities in Table 1 for the want of space. The column #S indicates number of sentences out of 525 where mentions of entities were found.

5.2 Applying Informed Active Learner

Out of 525 sentences, we use 400 sentences to teach the active learner in a 10-fold cross validation setup with 125 sentences held out to tune the learner. We annotated 10 sentences as denoting rules and 5 sentences as denoting non-rules before starting the learning sessions. In each session from thereon, the domain expert is queried by the learner for each sentence based on confidence values of the classification of that sentence.

To identify how the use of domain model and dictionary affect recall and precision, we show the feature representations a) when the domain model and the dictionary is used, b) when only a feature extractor based on n-gram tokenizer is used, and c) when the domain model and dictionary is used along with a feature extractor based on n-gram tokenizer. We include the extractor based on n-grams because n-gram extractors have been known to perform well, especially for text classification problem. For case (c), we treat the value of an interaction feature as the product of the values of the individual features.

[3] http://alias-i.com/lingpipe/index.html

[4] http://alias-i.com/lingpipe/docs/api/com/aliasi/util/FeatureExtractor.html

Table 1: KYC Entities and Mentions; #S: No. of Sentences where Entity Mention Occurs

Sr.	*Concepts*	*Mentions*	*#S*
1	Reporting Entity	all India financial institutions, local area banks, primary (urban) co-operative banks, scheduled commercial banks, state and central co-operative banks	14
2	Bank	bank	257
3	Account	client accounts, small accounts	5
4	Customer	foreign portfolio investors, politically exposed persons, artificial juridical person, association of persons, body of individuals	123
5	Document	certificate of incorporation, certificate/licence issued by the municipal authorities under Shop and Establishment Act, complete Income Tax Return, Licence/certificate of practice issued in the name of the proprietary concern by any professional body incorporated under a statute	128
6	Transaction	creating a legal person, cross-border wire transfer, deposits, withdrawal, fiduciary relationship	111
7	Risk Category	high, low, medium	23
8	Designated Director	managing Partner, managing director, managing trustee, whole-time director	2

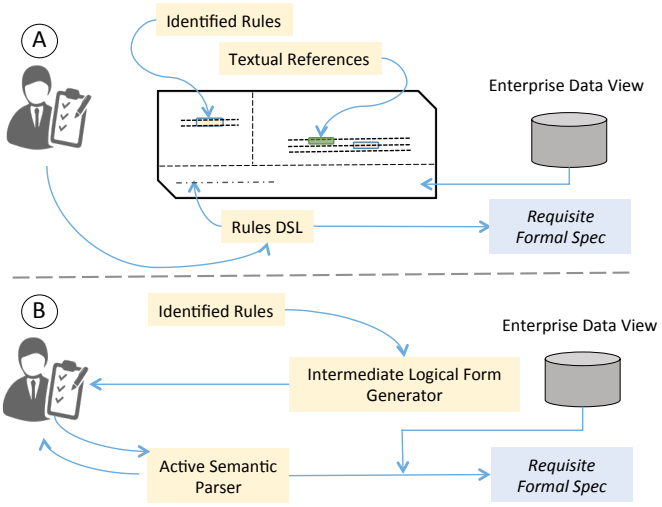

Figure 5: **A**: Rule Authoring Editor; **B**: Active Semantic Parser

① in Table 2 shows precision and recall values for active learning for the 3 extractors using uncertainty sampling. ② in Table 2 shows precision and recall values using certainty sampling. ③ in Table 2 shows representation of top valued features after the completion of 5 learning sessions. Both ① and ② in Table 2 show results for class of rule sentences. We do not show the results for the class of non-rules.

Uncertainty Sampling Results As seen in Table 2 ①, we found that when we used the domain model and the dictionary exclusively to represent features with uncertainty sampling, we obtained consistently higher recall than the other two extractors. On the other hand, using n-grams of lengths 3 to 5 exclusively, we obtained higher precision than the other two extractors as indicated in the second row of Table 2 ①. Recall represents retrieval coverage. Because the dictionary captures mentions of entities, the recall or coverage of dictionary extractor is comprehensive compared to n-grams generated by the learner, which are uninformative.

The extractor based on n-grams consistently has higher precision, which measures retrieval specificity but has correspondingly lower recall than the dictionary extractor as indicated in Table 2 ①. A look at the feature representation in first two columns of Table 2 ③ shows why these results may be attributed to capturing entities via dictionary of mentions against n-grams which do not make sense (*ther, nci, nanci, cash*, and so on). Note that in Table 2 ③, we only present top ranking features and row-wise they are unrelated.

The results in ① in Table 2 show that the combined extractor achieves recall of dictionary extractor and reaches the precision of n-gram extractor. A look at the features of this extractor in Table 2 ③ shows the combination of dictionary feature with n-grams. Note that symbols **+:** and ***:** merely separate the features from the dictionary and features from the n-gram extractors.

Certainty Sampling Results While our observations about better recall with the domain model and dictionary extractor, as well as better precision with the n-gram tokenizer extractor continue to hold with certainly sampling, we see higher precision and recalls as illustrated in Table 2 ② compared to uncertainty sampling counterparts in Table 2 ①.

Summary Our results show that informed active learner learns quickly and performs well to identify rules from the KYC text with the combined extractor and certainty sampling. Use of domain model and dictionary to inform features of active learning helps in comprehensive coverage of rules. Combining them with n-grams enable better precision as well. Using the combined extractor in certainty sampling setting enables even better performance due to early separation of decision boundary and setting of generative centers.

From the point of view of the domain expert, she is relieved of creating such intermediate artifacts while focusing only on the key domain entities and their mentions. Also, since rule identification is carried out by the active learner systems in an interactive manner with the domain expert, the burden of identifying rules from legal NL texts is also lessened for her.

5.3 From Rule Identification to Rule Specification

In the current state of the art, once the rules are identified, it is the domain expert who has to specify them in given rule specification language as shown previously in Figure 1. Two concerns have to be addressed for correct and exhaustive

Table 2: (1). Precision and Recall Values over 5 Runs of Active Learner using Uncertainty Sampling; (2). Precision and Recall Values over 5 Runs of Active Learner using Certainty Sampling; (3). Exemplar Feature Representations in 3 Extractors

Feature Representation	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*
Using ***domain model and dictionary***	0.80	1.0	0.78	1.0	0.78	1.0	0.75	1.0	0.75	1.0
Using ***ngrams (3-5)***	0.91	1.0	0.87	0.93	0.84	0.86	0.80	0.87	0.77	0.79
Using **both**	0.86	0.95	0.85	0.94	0.74	0.95	0.78	0.90	0.71	0.84

(1)

Feature Representation	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*
Using ***domain model and dictionary***	0.80	1.0	0.86	1.0	0.90	1.0	0.92	1.0	0.92	1.0
Using ***ngrams (3-5)***	0.91	1.0	0.89	0.97	0.93	0.98	0.96	0.98	0.92	0.99
Using **both**	0.86	0.95	0.90	0.85	0.93	0.87	0.90	0.95	0.94	0.94

(2)

Domain Model and Dictionary Features	NGram Features	Combined Features
DOCUMENT_proof of address	ther	+:TRANSACTION_transfers*:clu
DOCUMENT_documents	nci	+:DOCUMENT_proof of address*:pro
CUSTOMER_Politically Exposed Persons	nanci	+:CUSTOMER_non-face-to-face customers*:ust
TRANSACTION_transfers	cash	+:RISK CATEGORY_low*:ent

(3)

specification of regulatory rules. The rules should be specified in a language that enables report generation (as in practice [Sunkle *et al.*, 2015]) or formal checking (as in academia [Kharbili *et al.*, 2008; Racz *et al.*, 2011]). The enterprise data schema has to be taken into consideration for the facts in the rule specification. The facts need to reflect the schema from which they will be populated and against which the rules will be checked. We propose two lines of investigation in authoring the rules with the focus on reducing burden on domain experts, as it exists in current approaches.

Rule Authoring Editor This approach is illustrated in Figure 5 (A). Although it mirrors existing practical solutions to rule authoring [Koo, 2012; French Caldwell, 2013], the 3 key differences are that a) rules are identified with our active learner b) all reference sentences of all entity types and mentions are shown in a view per rule, and c) a rule domain-specific language is provided that enables the domain expert to write rules in familiar way from which requisite specifications are generated. This is unlike enterprise data tagging approaches followed in current industrial solutions [Pohlman, 2008], which often lack in coverage and accuracy of data tagged with regulations [Du *et al.*, 2013; Bartley *et al.*, 2010]. This approach is easier to implement on top of compliance frameworks which enable formal compliance checking of regulations against enterprise data as in [Sunkle *et al.*, 2015].

Active Semantic Parser It is possible to use some of the state of the art techniques in semantic parsing [Artzi and Zettlemoyer, 2013; Berant and Liang, 2014; Reddy *et al.*, 2014; Choi *et al.*, 2015], to obtain regulatory rules in logical forms. At the outset, this seems difficult for the same two reasons. Legal texts are considerably more complex than Freebase-style question answer pairs with a) business domain specific lexicon and b) context-dependent logical formulation. Also, the domain expert must expend considerable time and effort in creating sufficient number of mappings from each NL sentence to its logical form, before a given semantic parsing technique could be used reliably. We propose to tackle this through a technique employed in [Berant and Liang, 2014], whereby an intermediate logical form is generated, in our case based on the domain model and the dictionary as opposed to entity type signatures from Freebase as in [Berant and Liang, 2014]. Our proposal is that the domain expert uses these intermediate logical forms to provide handful of mappings to an *active semantic parser*, which learns in subsequent interactions to correctly map each regulatory rule to its desired logical form. This is illustrated in Figure 5 (B).

The active semantic parser is a more promising approach to rule specification but difficult to implement compared to rule authoring editor. It might be possible to implement the semantic parser itself in terms of a statistical machine translation (SMT) system. Given that the regular semantic parsers would generate logical forms of entire sentences, which are especially long in legal texts, an SMT system with a negative translation memory may be useful in discarding much of extraneous text in a rule sentence when translating it to a logical form. Such a negative translation memory could be obtained by clustering the legal text sentences to find which texts are most likely to be discarded by a human translator. We are currently exploring both the rule authoring editor and active semantic parsing options in our ongoing work.

6 Conclusion

In this paper, we presented an informed active learning approach to regulatory rule identification. Our approach reliably reduces the burden on the domain expert who simply interacts with our domain model generator and active learner systems. This is in contrast to existing NLP and ML approaches for rule identification in legal NL texts, which require considerable involvement of the domain expert. In our approach, the domain expert is not merely a labeler. Rather our approach enables automating the actual way a domain expert works with legal texts. We also proposed two future lines of investigations in which rules identified with our approach can be transformed to requisite rule specifications. We believe that our ongoing explorations have the potential to pave a better way toward largely automated regulatory compliance.

References

[Alberth *et al.*, 2012] Stephane Alberth, Bernhard Babel, Daniel Becker, Georg Kaltenbrunner, Thomas Poppensieker, Sebastian Schneider, Uwe Stegemann, and Torsten Wegner. Compliance and control 2.0: Unlocking potential through compliance and quality-control activities. *McKinsey Working Papers on Risk*, 33, 2012.

[Artzi and Zettlemoyer, 2013] Yoav Artzi and Luke Zettlemoyer. UW SPF: The University of Washington Semantic Parsing Framework, 2013.

[Baroni and Lenci, 2010] Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

[Bartley *et al.*, 2010] Jon W. Bartley, Y. Al Chen, and Eileen Zalkin Taylor. A Comparison of XBRL Filings to Corporate 10-Ks - Evidence from the Voluntary Filing Program. *Social Science Research Network*, Feb 2010.

[Berant and Liang, 2014] J. Berant and P. Liang. Semantic parsing via paraphrasing. In *Association for Computational Linguistics (ACL)*, 2014.

[Biagioli *et al.*, 2005] Carlo Biagioli, Enrico Francesconi, Andrea Passerini, Simonetta Montemagni, and Claudia Soria. Automatic Semantics Extraction In Law Documents. In Giovanni Sartor, editor, *ICAIL, June 6-11, 2005, Italy*, pages 133–140. ACM, 2005.

[Boella *et al.*, 2012] G. Boella, L. di Caro, L. Humphreys, L. Robaldo, and L. van der Torre. Nlp challenges for eunomos, a tool to build and manage legal knowledge, 2012.

[Breaux and Antón, 2005] Travis D. Breaux and Annie I. Antón. Deriving Semantic Models from Privacy Policies. In *6th POLICY Workshop, Sweden*, pages 67–76. IEEE Computer Society, 2005.

[Cau, 2014] David Cau. Governance, risk and compliance (GRC) software business needs and market trends, 2014.

[Choi *et al.*, 2015] Eunsol Choi, Tom Kwiatkowski, and Luke Zettlemoyer. Scalable semantic parsing with partial ontologies. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1311–1320. The Association for Computer Linguistics, 2015.

[de Maat and Winkels, 2008] Emile de Maat and Radboud Winkels. Automatic Classification of Sentences in Dutch Laws. In *Proceedings JURIX 2008*, pages 207–216, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[Du *et al.*, 2013] Hui Du, Miklos A. Vasarhelyi, and Xiaochuan Zheng. XBRL mandate: Thousands of filing errors and so what? *Journal of Information Systems*, 27(1):61–78, 2013.

[English and Hammond, 2014] Stacey English and Susannah Hammond. Cost of Compliance -2014, 2014.

[French Caldwell, 2013] John A. Wheeler French Caldwell. Magic quadrant for enterprise governance, risk and compliance platforms, 2013.

[Harris, 1968] Z. S. Harris. *Mathematical Structures of Language*. Wiley, NY, USA, 1968.

[Kaminski and Robu, 2015] Piotr Kaminski and Kate Robu. Compliance and Control 2.0: Emerging Best Practice Model. *McKinsey Working Papers on Risk*, 33, Oct 2015.

[Kharbili *et al.*, 2008] Marwane El Kharbili, Ana Karla A. de Medeiros, Sebastian Stein, and Wil M. P. van der Aalst. Business Process Compliance Checking: Current State and Future Challenges. In *MobIS*, volume 141 of *LNI*, pages 107–113. GI, 2008.

[Kiyavitskaya *et al.*, 2008] Nadzeya Kiyavitskaya, Nicola Zeni, Travis D. Breaux, Annie I. Antón, James R. Cordy, Luisa Mich, and John Mylopoulos. Automating the Extraction of Rights and Obligations for Regulatory Compliance. In Qing Li, Stefano Spaccapietra, Eric S. K. Yu, and Antoni Olivé, editors, *Conceptual Modeling - ER*, volume 5231 of *LNCS*, pages 154–168. Springer, 2008.

[Koo, 2012] Julia Koo. *What to look for in enterprise content management for compliance.* Wiley-IEEE Computer Society Pr, March 2012.

[Olsson, 2009] F. Olsson. A Literature Survey Of Active Machine Learning in the Context Of Natural Language Processing. Technical Report 06, Box 1263, SE-164 29 Kista, Sweden, April 2009.

[Pohlman, 2008] Marlin B. Pohlman. *Oracle identity management: governance, risk, and compliance architecture, third edition*. Auerbach Publications, Boston, MA, USA, 3rd edition, 2008.

[Racz *et al.*, 2011] Nicolas Racz, Edgar R. Weippl, and Riccardo Bonazzi. IT Governance, Risk & Compliance (GRC) Status Quo and Integration: An Explorative Industry Case Study. In *SERVICES 2011, USA, July 4-9, 2011*, pages 429–436. IEEE Computer Society, 2011.

[Reddy *et al.*, 2014] Siva Reddy, Mirella Lapata, and Mark Steedman. Large-scale semantic parsing without question-answer pairs. *Transactions of the Association for Computational Linguistics*, 2:377–392, 2014.

[Settles, 2009] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[Sunkle *et al.*, 2015] Sagar Sunkle, Deepali Kholkar, and Vinay Kulkarni. Toward better mapping between regulations and operations of enterprises using vocabularies and semantic similarity. *CSIMQ*, 5:39–60, 2015.

[Sunkle *et al.*, 2016a] Sagar Sunkle, Deepali Kholkar, and Vinay Kulkarni. Comparison and Synergy Between Fact-Orientation and Relation Extraction for Domain Model Generation in Regulatory Compliance. In *Conceptual Modeling - ER 2016, 35th International Conference, Gifu, Japan*, page Accepted, 2016.

[Sunkle *et al.*, 2016b] Sagar Sunkle, Deepali Kholkar, and Vinay Kulkarni. Informed Active Learning to Aid Domain Experts in Modeling Compliance. In *20th International IEEE Enterprise Distributed Object Computing Conference, ECOC 2016, Vienna, Austria*, page Accepted, 2016.

[van Engers *et al.*, 2004] Tom M. van Engers, Ron van Gog, and Kamal Sayah. A Case Study on Automated Norm Extraction. In T. Gordon, editor, *Legal Knowledge and Information Systems. Jurix 2004: The Seventeenth Annual Conference.*, Frontiers in Artificial Intelligence and Applications, pages 49–58, Amsterdam, 2004. IOS Press.

[Wyner and Peters, 2011] Adam Wyner and Wim Peters. On Rule Extraction from Regulations. In Katie Atkinson, editor, *Legal Knowledge and Information Systems - JURIX: Vienna, Austria*, volume 235 of *Frontiers in Artificial Intelligence and Applications*, pages 113–122. IOS Press, 2011.

[Zeni *et al.*, 2015] Nicola Zeni, Nadzeya Kiyavitskaya, Luisa Mich, James R. Cordy, and John Mylopoulos. GaiusT Supporting The Extraction Of Rights And Obligations For Regulatory Compliance. *Requir. Eng.*, 20(1):1–22, 2015.