Semantic Annotation of Finance Regulatory Text using Multilabel Classification

Kartik Asooja[1], Georgeta Bordea[1], Gabriela Vulcu[1], Leona O'Brien[2], Angelina Espinoza[2], Elie Abi-Lahoud[2], Paul Buitelaar[1], and Tom Butler[2]

[1]Insight Centre, National University of Ireland, Galway
firstname.lastname@insight-centre.org
[2] Governance Risk and Compliance Technology Centre (GRCTC)
University College Cork, Ireland

Abstract. The Financial Industry is experiencing continual and complex regulatory changes, on a global scale, making regulatory compliance a critical challenge. Semi-automated solutions based on semantic technologies are called for to assist subject matter experts in identifying, clasifying and making sense of these regulatory changes. Previous work on fine-grained semantic annotation of regulatory text focused on general regulatory concepts, but domain-specific concepts play an important role in compliance verification in a specific area of interest. This paper describes an approach for semantic annotation of regulations in the financial domain, more specifically for Anti-Money Laundering legislation, that uses multilabel classification to populate a formal ontology. Evaluation results show that contextual features provide important cues for fine-grained classification of regulatory documents.

1 Introduction

Regulatory compliance is a daunting challenge in many industries, but none more so than the financial industry, that is undergoing major regulatory change across the globe in the aftermath of the 2008 Financial Crisis. The US Dodd-Frank Wall Street Reform and Consumer Protection Act[1] (hereinafter Dodd-Frank), and the EU Single Supervisory Mechanism are resulting in numerous and staggered changes in regulations. Take for example the on going implementation of Dodd-Frank in the US that resulted in a significant increase of regulatory provisions: in 2010, before Dodd-Frank, the US Code of Federal Regulations Titles 12 and 17 combined had a total of 343 restrictions [9]. As of end 2012, the amount of restrictions increased to 3172 and the implementation of Dodd-Frank is barely halfway complete [9]. Current solutions for regulatory compliance rely on expensive and time-consuming manual labour to analyse regulatory documents and provide guidance on them. The main objective of this work is to leverage semantic technologies to assist Subject Matter Experts (SMEs) in making sense of the wide and complex spectrum of financial services regulations, by providing

[1] Pub.L. 111203, H.R. 4173

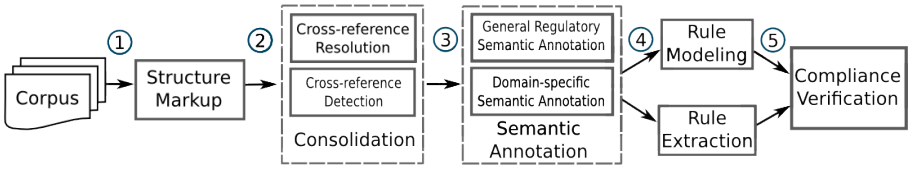

Fig. 1. Pipeline for Semi-Automated Compliance Verification

intelligent querying over regulations for advanced search and retrieval using semantic annotation. Additionally, semantic annotation is an important step for identifying relevant text units for automated compliance verification. Based on related work, the stages of a semi-automated pipeline for regulatory compliance verification are presented in Figure 1.

Although more and more regulatory documents are available in XML/RDF, the formal structure of PDF/HTML documents has to be marked up to allow further processing, as can be seen in the first step. This is usually done by hand, but automated approaches based on hand-crafted extraction patterns are proposed as well [2]. Regulatory text makes use of a complex structure of cross-references [12]. Cross-references may refer to elements within the document itself but also to other elements from external regulatory documents. These have to be consolidated in the second stage, using either hand-crafted extraction patterns [2, 15], or supervised learning [16]. The next stage deals with semantic annotation of text units, which requires a semantic framework consisting of general regulatory semantics and domain-specific semantics [4], separating prescriptive (deontic) knowledge from world knowledge to enable fine-grained annotation. The work proposed here is positioned at this stage, introducing a classification approach for automatic semantic annotation with domain-specific semantics. The fourth step is related to the conversion of regulatory text to a standardised format for normative rules [3]. This can be achieved either through rule modelling, using an appropriate interface [13], or through rule extraction using a combination of gazetteers and hand-crafted extraction patterns [19]. Solutions based on deep parsing are also proposed [11], but existing parsers are not well adapted to regulatory text. Finally, compliance can be automatically checked using an appropriate logical framework [8].

More specifically, we focus on a semantic annotation use case based on the Anti-Money Laundering (AML) domain, an area of the financial industry that is under intense scrutiny. Because an ontology for Anti-Money Laundering is not readily available, domain experts are involved to construct one. The goal is to semantically annotate relevant text units that fall under areas such as Customer Due Diligence (CDD), Customer Identification and Verification (CID&V), Enforcement, Monitoring, and Reporting. This work is structured as follows. First, we discuss related work in Section 2, then we present an overview of the overall semantic annotation approach in Section 3, followed by a description of the AML classification approach in Section 4. The proposed approach is evaluated in Section 5, concluding the paper in Section 6.

2 Related Work

An important step for enabling intelligent querying over regulatory documents, and eventually automatic compliance verification, is semantic annotation of paragraphs in regulatory documents. Existing statistical learning methods that classify text given a set of classes are successfully used for automatic annotation, especially for general regulatory concepts. A more recent attempt to classify regulatory text [5,11] analyses regulatory text for automatically translating regulations into rules. They propose an approach to classify paragraphs into several classes from the general regulatory domain. These include definitions, types of references (i.e., cross-document and intra-document) and regulation types (e.g., Prohibition, Obligation, Permission). Mencia et. al. [10] consider a more fine-grained classification based on a large number of regulatory concepts. They focus on comparing different algorithms for multilabel classification, including multilabel multiclass perceptron (MMP), multilabel pairwise perceptron (MLPP) and binary relevance (BR). The main difference with our work is that in their work semantic annotation is primarily intended for retrieval and search of a large repository of regulatory documents. For this reason, they analyse a large dataset based on Eur-Lex, which provides direct access to the Official Journal of the European Union and EU Law including EU treaties, directives, regulations, decisions, consolidated legislation, EU case-law, preparatory acts and international agreements. Classification is done at document level, using annotations provided by subject matter experts from EUROVOC[2]. Francesconi E. et. al. [7] focus as well on classifying general regulation types such as Prohibition and Exception, but follow a multi-class approach at a more granular level, considering that each paragraph can be associated with multiple annotations. In our work, we consider a semantic framework that combines general regulatory ontologies and domain-specific ontologies, using separate ontologies for general regulatory concepts and domain-specific provision types. The novel aspect of our work is that related research is mostly focussed on general regulatory concepts, rather than on domain-specific concepts.

3 Semantic Annotation Approach

The approach proposed in this work is based on manual creation and automatic population of an ontology with high-level regulatory concepts as well as domain-specific concepts. To achieve this, we combine input from legal SMEs with results extracted using text analysis techniques. Our main goal is to semantically annotate regulatory documents for automated compliance verification, aiming to provide the capability to answer questions such as:

- List the Obligations in a regulatory text
- List the Prohibitions in a regulatory text

[2] http://eurovoc.europa.eu/

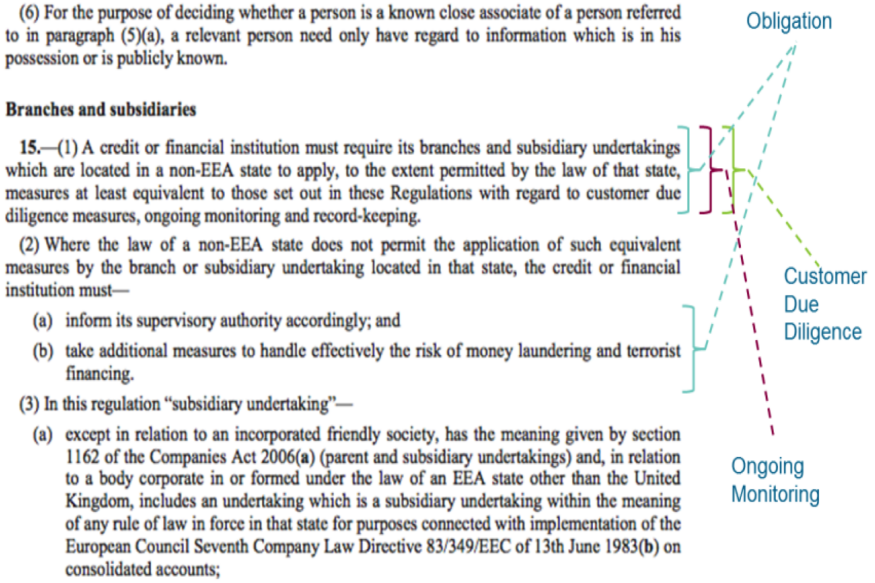

Fig. 2. Semantic annotation of a text unit from UK AML Regs 2007

– List provisions related to regulatory area (e.g., Anti-Money Laundering) refined by provision (e.g., Customer Due Diligence, Reporting) and further refined by modality (e.g., Obligation, Prohibition)

Take for example the text provided in Figure 2, that was extracted from The Money Laundering Regulations 2007 (UK)[3] (hereinafter UK AML Regs 2007). To answer these types of queries, the appropriate domain-specific provision and modality types have to be identified for each text unit. Note that while provision types are in general mutually exclusive, multiple AML provisions can be associated with one text unit. For example, Section15(1) is related to two AML previsions, customer due diligence and ongoing monitoring.

Figure 3 illustrates the four stages of constructing a knowledge base for Regulatory Change Management (RCM), combining several manual techniques with semi-automatic technologies as follows. First, in the ontology engineering phase, SMEs use a Regulatory Controlled Natural Language to restitute in an unambiguous way the regulatory intent of legal texts into a rulebook. The output of this stage is then used by Semantic Technologies Experts (STEs) to create a family of formal ontologies. The stack of ontologies created are generically referred as Financial Industry Regulatory Ontology (FIRO). This phase is supported by the Semantics of Business Vocabulary and Rules[4] as described in [1]. There are four ontologies in the stack, dealing with different aspects of a regulatory document.

[3] S.I No. 2175 of 2007
http://www.legislation.gov.uk/uksi/2007/2157/pdfs/uksi_20072157_en.pdf

[4] Semantics of Business Vocabulary and Business Rules: http://www.omg.org/spec/SBVR/1.2

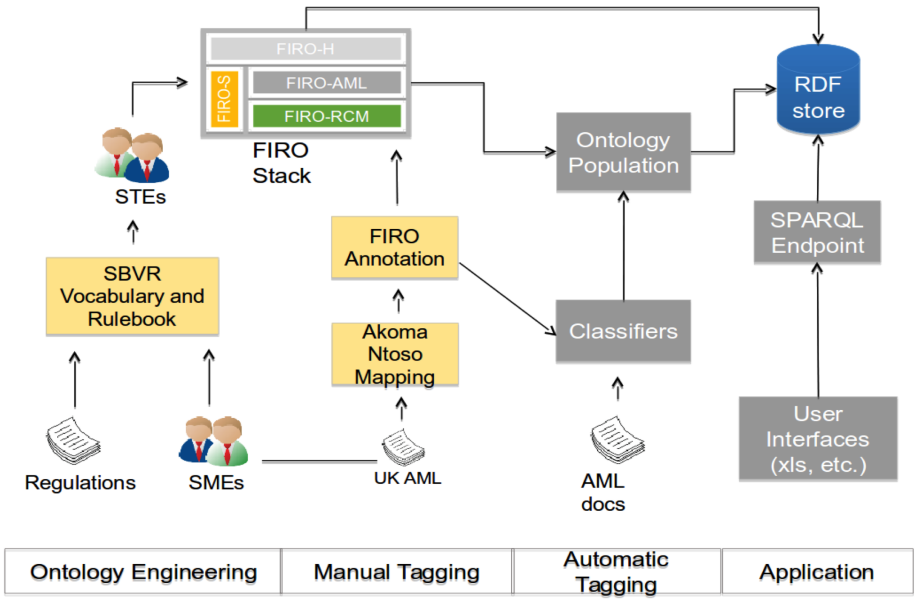

Fig. 3. Stages of the Semantic Annotation Approach for Regulatory Change Management (RCM)

The Financial Industry Regulatory Ontology (FIRO) is an ontology model composed of relevant and interlinked ontologies in the financial industry regulatory domain. These interlinked ontologies are also called modules. FIRO captures regulatory imperatives from legislation and regulatory texts, transforming them into formal semantics, and is specified using the Description Logic-based ontology language: OWLDL [18].

The FIRO semantic framework is composed of four modular ontologies FIRO-H, FIRO-S, FIRO-[Domain] (in our case FIRO-AML), and FIRO-Op[Operational] (in our case FIRO-RCM). The FIRO-H ontology describes high-level concepts and their relationships that are applicable across the regulatory domain. This includes concepts, such as Obligation, Prohibition, Exemption or Sanction. The FIRO-S ontology models the formal structure of parliamentary, legislative and judiciary documents. For this purpose, the Akoma Ntoso Standard [14] is used as the main source for defining this ontology. FIRO-[Domain] describes domain-specific concepts and their relationships. The work presented in this work makes use of the FIRO-AML ontology for Anti-Money Laundering regulations. Finally, FIRO-Op[Operational] merges all the three previous FIRO ontologies, in order to support a particular purpose or task in the compliance verification process.

The second stage, manual tagging, is concerned with the creation of training data for the next automatic stages of the approach. FIRO-H concept labels are referred in this work as modalities, while FIRO-AML concept labels are referred as AML provision types. Each textual unit is annotated with corresponding modality and provision. In the case of modalities we consider that each text unit

refers to a single modality type, while for AML provision types a textual unit can refer to multiple types at the same time. This calls for two different classification approaches, a standard multinomial approach and a multilabel classification approach respectively. In the third stage of the RCM approach, automatic tagging, regulatory documents are automatically tagged using supervised classification. Both classifiers are trained using manually tagged data by legal SMEs from the previous step. Each type of annotation requires a separate training set to train the corresponding classifier. We rely on open source machine learning libraries, including Weka, Meka, Mulan, and Mallet. Tags assigned by the classifiers are then used to automatically populate the ontology, or more exactly, to tag regulatory text with concepts from the ontology. However, in this paper, we focus only on the multilabel classification leading to domain-specific provision tagging in regulatory documents. In the fourth and final phase, the content of the resulting knowledge base is made accessible via a SPARQL[5] endpoint to enable direct querying. At the moment we consider two approaches for the front end, one based on a Graphical User Interface and another one using spreadsheets, a more familiar environment for the financial industry, that give dynamic views on automatically annotated documents using Filters and Pivots.

4 Multilabel Classification of AML Provisions

Multilabel classification is required in order to identify AML provision types. A text unit can fall under multiple AML provisions at the same time, therefore multinomial classification is not suitable in this scenario. Multilabel classification associates a set of target labels to each test sample, instead of just one label as typically done in multinomial classification. In general, existing approaches for multilabel classification are either problem transformation methods or algorithm adaptation methods. Problem transformation methods translate the multilabel classification problem into one or more binary classification or regression problems, while algorithm adaptation methods modify specific classification algorithms to handle multilabel data [17].

To model the multilabel classification problem for identifying AML provision types, we used the one vs. the rest strategy, also called Binary Relevance (BR) approach, training a separate off-the-shelf binary classifier (e.g., SVM) per class/label. Basically, we create N instances of a binary classifier to solve N independent classification problems, where N is the number of labels in the problem. To predict the set of labels for a sample, each classifier is consulted for its corresponding label, and then all the positives are collected from the classifiers to form the resultant set [17]. This model is implemented using the open source libraries MEKA and MULAN, which provide open source implementations of methods for multilabel classification and evaluation[6][7]. Both libraries are based

[5] http://en.wikipedia.org/wiki/SPARQL
[6] MEKA: http://meka.sourceforge.net/
[7] MULAN: http://mulan.sourceforge.net

on the WEKA machine learning toolkit[8]. Basic text pre-processing is applied including stemming, lower casing and stopword removal.

Manual Annotation To create training data for supervised classification, SMEs annotate AML regulatory documents with provision types. The segmentation of legal text into meaningful textual units for classification is done by relying on the XML version of the documents. This includes tags that define different types of paragraphs, the only issue is selecting the right level of granularity for the RCM application. Regulatory documents are segmented into smaller textual units using the P2 tags from UK XML format[9] for all UK Legislation. In our case, the decision on the appropriate level of text granularity is done with the help of SMEs. SMEs manually annotate regulatory text based on FIRO concepts, following a standard methodology based on GATE [6]. GATE is an open-source framework for natural language processing that provides a graphical environment for development and annotation. The user friendly interface allows SMEs to mark up documents more easily, using an annotation schema that covers concepts of interest from FIRO-H and FIRO-AML. Annotations are stored in a machine readable format, therefore a parser can be used to automatically convert the training data in the format required by the machine learning framework.

Provision type	**Manually identified features**
CDD	due diligence measures; beneficial owner; public function; business relationship; ownership and control
CID&V	customer identity; reliable source documents; account opening; risk-based assessment; high-risk customers
Enforcement	requires information; prosecution; penalties; fail to comply; civil offence
Monitoring	ongoing monitoring; customer relationships; suspicious activity; transaction types; suspicious transactions
Reporting	reports; internal audits; suspicious activity report; internal reporting; annual report

Table 1. Manually identified cue-phrases by SMEs for the five main AML provision types

Feature Engineering We experimented with different features for classification including n-grams (uni/bi/tri grams), manually identified cue phrases, modal verbs (must, should etc.) and contextual features. We also performed automatic feature selection to find the best feature subset in the dataset. Table 1 lists several examples of n-gram features for five different provision types manually identified by SMEs. The basic features described above take into consideration

[8] WEKA: http://www.cs.waikato.ac.nz/ml/weka/
[9] http://www.legislation.gov.uk/uksi/2007/2157/made/data.xml

information contained in the text unit itself, but the provision addressed in a text unit can be derived from the surrounding text unit as well. Standard sequential classification models such as CRF or HMM are specifically designed to account for sequential dependencies, but these algorithms are not readily available in a multilabel classification setting. To account for this type of information we consider contextual features, to account for sequential dependencies. Succeeding and preceding n-grams are extracted from surrounding text unit. Each textual unit, i.e,. a P2 tagged text unit, in the document is transformed into a feature vector that consists of n-grams from the text unit itself, as well as n-grams from the succeeding and preceding text units. The classifier is allowed to differentiate contextual features from the others by adding the strings "_previous" and "_next" as subscript.

AML Provision Type	**No. of Instances**
Customer Due Diligence	67
Customer Identification and Verification	20
Enforcement	99
Monitoring	12
Reporting	14

Table 2. AML provision classes considered for classification and the number of manually annotated instances in the UK AML Regs 2007

5 Evaluation

This section presents the evaluation of our approach for automatically tagging regulatory documents using supervised classification. The approach is evaluated on the Anti-Money Laundering regulatory domain, but the approach can be applied to other regulatory domains as well if an appropriate ontology of provision types is available.

Dataset Overview For our experiments, we use the UK AML Regs 2007, which is annotated by SMEs with FIRO-H and FIRO-AML ontology labels. We relied on the XML version of the document for the text segmentation in smaller textual units given by P2 tags. Legal SMEs tag these text units with the corresponding provision types. To evaluate, we use cross validation with 8 folds on the UK AML Regs 2007 dataset. For our current experiments, we limit ourselves to the five most frequent Provision classes, based on the number of annotated examples available in the dataset. Table 2 shows the number of instances available in the dataset, i.e., the UK AML Regs 2007, for each Provision label. The number of available instances varies considerably between provision types, with a larger number of text units of type Enforcement and CDD, and a much smaller number of Reporting and Monitoring text units due to the focus of the Regulations.

Approach	Micro-average F-measure	Macro-average F-measure
Baseline (unigrams + bigrams)	0.7610	0.6164
Baseline + Manual Features	0.7571	0.6187
Baseline + Context Features	0.8066	**0.7011**
All Features	**0.8127**	0.6961

Table 3. Overall averaged results for classification of AML provision types

AML Provision	Baseline	Baseline +Manual	Baseline +Context	All Features
CDD	0.7676	0.7653	0.8409	**0.8475**
CID&V	0.4000	0.4238	**0.5224**	0.4924
Enforcement	0.8612	0.8508	0.8486	**0.8670**
Monitoring	0.7133	0.7133	**0.8038**	**0.8038**
Reporting	0.3400	0.3400	**0.4900**	0.4700

Table 4. Detailed results for each AML provision type based on averaged F-Measure

Evaluation measures Different evaluation measures are typically used when computing evaluation scores for classification. When multiple class labels are considered, averaging the evaluation measures for each class can give an overview of the results. There are two types of evaluation measures when averaging results, micro-averaged and macro-averaged results [17]. Macro-averaging refers to the unweighted average of the precision/recall taken separately for each class, therefore it is an average over classes. The micro-average, on the contrary, is an average over instances, therefore classes which have many instances are given more importance. Macro-averaging computes a simple average over classes while micro-averaging pools per-document decisions across classes, and then computes an effectiveness measure on the pooled contingency table. Differences between the two methods can be considerable because macro-averaging gives equal weight to each class, whereas micro-averaging gives equal weight to each per-document classification decision. Because the F1 measure ignores true negatives and its magnitude is mostly determined by the number of true positives, large classes dominate small classes in micro-averaging. Micro-averaged results are therefore a measure of effectiveness on the large classes of the test collection. To get a sense of effectiveness on small classes, one should compute macro-averaged results. Therefore, for our evaluation, we compute both micro and macro F-measure.

Experimental Results Table 3 shows the results for AML provision classification with cross validation over the UK AML Regs 2007 dataset using different feature combinations. We consider a simple baseline approach that uses the unigrams and bigrams from the target text unit as features. There is a clear improvement in both micro-averaged and macro-averaged F-measure by considering contextual features. This indicates that preceding and succeeding paragraphs provide important information about the AML provision type as-

sociated with a text unit. However, the best results are obtained by combining all the features. Table 4 presents detailed results for each AML provision class, when using different feature combinations. With the Exception of the Enforcement class, context features improve significantly the F-measure scores for all the classes. On the other hand, the combination of all the features does not improve the scores for the CID&V and Reporting classes. These differences can be explained by variations in the number of training instances for different classes.

System output To illustrate the results provided by the complete semantic annotation pipeline, we provide as an example a query and the results obtained by querying the knowledge base. Automatically annotated text units with provision and modality types are used to populate FIRO ontologies with RDF instances. These RDF instances are loaded in an RDF repository and made available through a SPARQL endpoint service, to allow exploitation in different application scenarios. For instance, these results can be directly used to construct spreadsheet reports that can be further analysed by SMEs. At the same time, they can provide valuable input for downstream applications that allow rule modelling or rule extraction for compliance verification. Figure 4 shows a snapshot of the hosted SPARQL endpoint. It shows a query that extracts all the provisions with their modality types. Figure 5 shows the provisions returned by the endpoint, with links to their location in the original document and the full free text.

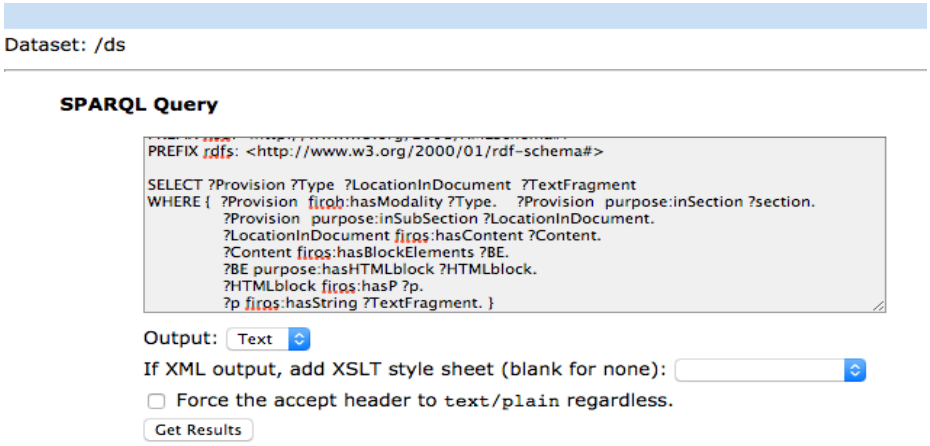

Fig. 4. SPARQL Endpoint for querying the automatically annotated UK AML Regs 2007

AML Provision	Modality	Location	Text
CID&V	Obligation	section16-sub6	In this regulation, "financial conglomerate" and "third-country financial conglomerate" have the meanings given by regulations 1(2) and 7(1) respectively of the Financial Conglomerates and Other Financial Groups Regulations 2004 .
Enforcement	Obligation	section37-sub3	An officer may exercise powers under this regulation only if the information sought to be obtained as a result is reasonably required in connection with the exercise by the designated authority for whom he acts of its functions under these Regulations.
CDD	Prohibition	section16-sub3	A credit or financial institution carrying on business in the United Kingdom must not set up an anonymous account or an anonymous passbook for any new or existing customer.

Table 5. Sample results of the SPARQL query shown in Figure 4

6 Conclusions

This paper describes the implementation of an approach to semantic tagging of regulatory documents for the purpose of regulatory compliance verification in the financial industry. The main challenge when adopting a supervised approach is the creation of the training dataset that requires extensive involvement of SMEs. The training dataset used in our experiment is relatively small, although comparable with a similar dataset used in related work [11]. The small size of the annotated Anti-Money Laundering documents indicates that there is still place to improve the classification performance by using a larger training set. Early results are promising, however, the proposed approach has limitations when training data is not abundant. Therefore, as a future step, we aim to collect more training data by designing a user interface for interactive data curation by legal SMEs. The plan is to further improve the feature set by consulting legal SMEs. Another challenge is that legislation structure is not always available in standardised, machine readable formats. The segmentation of legislation documents into meaningful units for classification is not a trivial task, when documents are available in PDF or HTML format alone. In this demo, we make the assumption that we can rely on specific XML tags following the UK XML structure for legislation.

Acknowledgement

This work has been funded in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT).

References

1. E. Abi-Lahoud, L. OBrien, and T. Butler. On the road to regulatory ontologies. In P. Casanovas, U. Pagallo, M. Palmirani, and G. Sartor, editors, *AI Approaches to the Complexity of Legal Systems*, volume 8929 of *Lecture Notes in Computer Science*, pages 188–201. Springer Berlin Heidelberg, 2014.

2. M. Adedjouma, M. Sabetzadeh, and L. C. Briand. Automated detection and resolution of legal cross references: Approach and a study of luxembourg's legislation. In *Requirements Engineering Conference (RE), 2014 IEEE 22nd International*, pages 63–72. IEEE, 2014.
3. T. Athan, H. Boley, G. Governatori, M. Palmirani, A. Paschke, and A. Wyner. Oasis legalruleml. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 3–12. ACM, 2013.
4. T. Beach, Y. Rezgui, H. Li, and T. Kasim. A rule-based semantic approach for automated regulatory compliance in the construction sector. *Expert Systems with Applications*, 2015.
5. A. Buabuchachart, K. Metcalf, N. Charness, and L. Morgenstern. Classification of regulatory paragraphs by discourse structure, reference structure, and regulation type. 2013.
6. H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, et al. Developing language processing components with gate version 6 (a user guide).
7. E. Francesconi and A. Passerini. Automatic classification of provisions in legislative texts. *Artificial Intelligence and Law*, 15(1):1–17, 2007.
8. L. Giordano, A. Martelli, and D. T. Dupré. Temporal deontic action logic for the verification of compliance to norms in asp. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 53–62. ACM, 2013.
9. P. McLaughlin, R. Greene, and Dodd-Frank. What it does and why its flawed. *Mercatus Center, George Mason University*, 2013.
10. E. L. Mencía and J. Fürnkranz. *Efficient multilabel classification algorithms for large-scale problems in the legal domain*, volume 6036. Springer, 2010.
11. L. Morgenstern. Toward automated international law compliance monitoring (tailcm). Technical report, Intelligence Advanced Research Projects Activity (IARPA), 2014.
12. M. Palmirani and L. Cervone. Measuring the complexity of the legal order over time. In *AI Approaches to the Complexity of Legal Systems*, pages 82–99. Springer, 2014.
13. M. Palmirani, L. Cervone, O. Bujor, and M. Chiappetta. Rawe: An editor for rule markup of legal texts. In *RuleML (2)*, 2013.
14. M. Palmirani and F. Vitali. Akoma-ntoso for legal documents. In *Legislative XML for the Semantic Web*, pages 75–100. Springer, 2011.
15. P. Spinosa, G. Giardiello, M. Cherubini, S. Marchi, G. Venturi, and S. Montemagni. Nlp-based metadata extraction for legal text consolidation. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 40–49. ACM, 2009.
16. O. T. Tran, M. Le Nguyen, and A. Shimazu. Reference resolution in legal texts. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 101–110. ACM, 2013.
17. G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2010.
18. C. Welty, D. L. McGuinness, and M. K. Smith. Owl web ontology language guide. *W3C recommendation, W3C (February 2004) http://www. w3. org/TR/2004/REC-owl-guide-20040210*, 2004.
19. A. Wyner and W. Peters. On rule extraction from regulations. In *JURIX*, volume 11, pages 113–122, 2011.