An Obligations Extraction System for Heterogeneous Legal Documents: Building and Evaluating Data and Model

Maria Iacono, Laura Rossi, Paolo Dangelo, Andrea Tesei, Lorenzo De Mattei
Aptus.AI / Pisa, Italy
{maria,laura,paolo,andrea,lorenzo}@aptus.ai

Abstract

A system that extracts obligations automatically from heterogeneous regulations could be of great help for a variety of stakeholders including financial institutions. In order to reach this goal, we propose a methodology to build a training set of regulations written in Italian coming from a set of different legal sources and a system based on a Transformer language model to solve this task. More importantly, we deep dive into the process of human and machine-learned annotations by carrying out both quantitative and manual evaluations of both of them.

1 Introduction

Compliance practitioners in financial intuitions are overburdened by the high volume of upcoming regulations coming from different legal sources, such as the European Union, National legislation, central banks and independent administrative authorities sources, to name a few. Part of the compliance offices work consists of extracting obligations from this vast amount of regulations to trigger compliance processes. It is worth noting that extracting obligations from such a big amount of regulations is tedious and repetitive work. In this scenario having systems to automate this process might be very useful to cut down the costs. Machine Learning (ML) and Natural Language Processing (NLP) may come in help. However, given the variety of legal sources, training this kind of system is a complex activity because it requires a sufficient amount of annotated data, which are expensive especially if the annotations require legal domain experts.

The obligations extraction topic has been already studied with different approaches. Bartolini et al. (2004) used a shallow syntactic parser and hand-crafted rules to automatically classify laws paragraphs according to their regulatory content and extract relevant text fragments corresponding to specific semantic roles. Similarly Sleimi et al. (2018) represent automatically legal texts semantics using an RDF schema with a system based on a dependency parser and hand-crafted rules. Sleimi et al. (2019) used the same representation to build a question-answering system with a focus on obligations. Biagioli et al. (2005) represent law paragraphs using Bag of words either with TF or TF-IDF weighting (Salton and Buckley, 1988) and used Support Vector Machines (SVM) to classify each paragraph as a type of provisioning including obligations. A similar approach is adopted by Francesconi and Passerini (2007): they classify legislative texts paragraphs according to the proposed provision model. They represent them in a similar way as (Biagioli et al., 2005) and use two learning algorithms: Naive Bayes and SVM. Sleimi et al. (2020), propose to address the problem of the complexity of regulatory texts by writing them following a set of standard templates which could be easily parsed.

Contributions In this work we offer four main contributions. (i) We propose a methodology for building training corpora relying on non-expert annotators and we apply this methodology on a set of heterogeneous regulations written in Italian, coming from a set of different legal sources. (ii) We assess the quality of the introduced methodology relying on an inter-annotator agreement score and we carry out an error analysis to highlight if and when expert annotators are required. (iii) We use the dataset produced to train and test an obligations classification system based on neural networks as this approach has been proven to provides state of the art results for several Italian classification tasks (De Mattei et al., 2018; Cimino et al., 2018; Occhipinti et al., 2020). (v) We conduct a manual error analysis to investigate the pros and the limitations of the mentioned system.

2 Task Description

The task we tackle consists of classifying regulations clauses either as obligations or not. By obligation, we mean, from a juridical point of view, a legal constraint imposed by law and addressed to a juridical person.

Being interested in developing a system that supports financial institutions, we distinguish two categories of obligations, classifying them as relevant or irrelevant for financial institutions. Then each clause can be classified in one out of the following three categories: (i) `not obligation`, (ii) `relevant obligation` and (iii) `not relevant obligation`. This classification schema allows practitioners to retrieve in one click all the obligations or the relevant only so that they can decide whether to have a complete overview of the laws they are consulting or to focus only on the obligations that actually affect their institutions.

To distinguish the two categories, we look at the subject to whom the obligation is addressed: if it is a public institution, we classify it as an irrelevant obligation, in all other cases as a relevant obligation. This simplification applied to the classification criterion may seem extreme since it implies that any type of obligation not addressed to a public institution must be considered relevant for a financial institution. However, we believe that applying this distinction is a good strategy because the documents we analyze are already filtered, i.e., they belong to a category of laws that impact financial institutions. Consequently, within them, if an obligation is not directed at a public institution it will almost certainly be directed somehow to financial institutions.

2.1 Special Cases

Legal jargon is not merely a tool used for argumentation or narrative, but a constitutive element of the law. Consequently, the structure of legal texts has particular characteristics that must respond to precise and predictable patterns. Despite this, there are cases in which the language can be ambiguous. Since our goal is to build a dataset in line with compliance practitioners expectations we analyzed some special cases with a group of experts in order to provide clear guidelines to annotators.

One such case is when an obligation is expressed indirectly, for example through the formulation of a right. If an article talks about rights of any kind, it assumes that those rights must be respected. So, for example, the right of a client in terms of obtaining a loan (client's point of view) corresponds to a duty of the bank, which is obliged to grant it if the client has what it takes (bank's point of view). Similarly, an employee's right to go on vacation means that the employer must guarantee vacation days. For this reason, in deciding how to classify a part of a law, in addition to the interpretation by the annotator, the concept of "priority" comes into play. Since our application is designed to support financial institutions, our priority is to highlight the obligations that they must take into account in order not to risk penalties. Consequently, if a sentence represents both a right for one subject and duty for another, we prioritize the obligation in classifying it.

Another case where the priority factor comes into play is that of clauses that contain both relevant and irrelevant obligations. In these cases, since we cannot break the clause down into several parts, we give priority to the relevant obligation. In terms of risk, it is better to classify an irrelevant obligation as relevant, rather than the other way around.

In addition, we have to consider that obligations may be reported implicitly. For example, if a person can perform an action only under certain conditions, it is implied that those conditions can be interpreted as obligations. According to this principle, we do not classify a sentence such as "Spectators may enter the theatre" as an obligation. On the contrary, we do so when a condition is added, as in the case of the sentence "Spectators may enter the theatre only if they have the ticket."

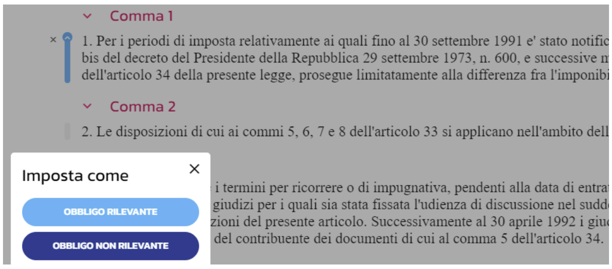
Even if we, as readers, do not pay attention to it, normative texts often contain implicit information that readers are naturally able to trace through reading, such as an implied subject, or a reference to another part of the document or to an external document. Unlike a reader, an automatic classifier, not having provided with enough context, may encounter difficulties in handling this kind of case.

3 Data Annotation

We extracted the dataset from Daitomic[1], a product that automatically collects legal documents from a wide variety of legal sources, represents automatically them accordingly to the Akoma Ntoso standard (Palmirani and Vitali, 2011) and makes them available through a dedicated User Interface. The adoption of Akoma Ntoso lets us represent the structure of heterogeneous legal texts in a unified format that makes us able to apply the same operations on very different kind of poorly encoded documents such as PDF, HTML and DOCX files.

The corpus has been manually labelled by three trained annotators with no previous background in legal domain and contains 71 regulations for a total of 10.628 clauses. We selected regulations that touch heterogeneous topics such as data privacy, financial risk, tax compliance and many more but all of them are known to be relevant for financial institutions. In order to deal with the problem of heterogeneity of normative sources, we found it appropriate to take texts from different sources, so that we could train the model in a balanced way. In particular, we extracted the texts from thirty of the most important regulatory sources for Italian financial institutions, including Gazzetta Ufficiale Italiana, EUR-Lex, Consob, Banca d'Italia and many more. From these sources, we selected texts of different types: acts, regulations, decisions, directives, communications, statutes, and more. In this way, we created a very heterogeneous dataset that can be considered representative of the wide variety of existing regulations.

The annotations were carried out directly from the graphical user interface of the Daitomic application, which allows, within the consultation section, to mark the requirements present in the law and to classify them as relevant or not relevant. The application texts are already structured, so they present a tree structure divided into chapters, articles, paragraphs, clauses, etc, where we annotated the smallest parts, i.e. clauses. Each clause is flanked by a sidebar, clicking on which automatically opens the pop-up shown in Figure 1, which allows the annotators to choose the label that they consider most appropriate. As a result of this choice, the sidebar will turn light blue if the obligation is classified as relevant to financial institutions, and dark blue if it is not relevant.



1. Per i periodi di imposta relativamente ai quali fino al 30 settembre 1991 e' stato notific
bis del decreto del Presidente della Repubblica 29 settembre 1973, n. 600, e successive m
dell'articolo 34 della presente legge, prosegue limitatamente alla differenza fra l'imponibi

2. Le disposizioni di cui ai commi 5, 6, 7 e 8 dell'articolo 33 si applicano nell'ambito dell

i termini per ricorrere o di impugnativa, pendenti alla data di entra
giudizi per i quali sia stata fissata l'udienza di discussione nel sudd
zioni del presente articolo. Successivamente al 30 aprile 1992 i giu
del contribuente dei documenti di cui al comma 5 dell'articolo 34.

Figure 1: Pop-up for setting the label of the obligation.

We picked four of the annotated laws containing as many as 2189 clauses to be annotated by all three annotators.

4 Annotations Evaluation

We used the part of the dataset annotated by all three annotators in order to calculate the inter-annotator agreement (IAA). Using Krippendorff's Alpha reliability, we computed IAA in two different ways, at first checking only whether they had classified the sentences as obligations or non-obligations, then taking into account their choices in distinguishing obligations between relevant and non-relevant. The resulting IAA is 0.58 considering the distinction between relevant and not relevant but increases to 0.70 if no such distinction is applied.

In order to better understand these results we carried out a manual analysis from which turned out that most cases of disagreement are of two kinds (two examples are reported in Table 1). The lack of agreement between annotators can be primarily attributed to the fact that there is often no explicitly expressed subject in a clause, either because it is expressed in the preceding clauses or because it is intuitable from the context, as we can see in the first example. Another frequent reason for disagreement is surely the fact that our annotators, not being experts in the legal field, not always are able to understand the kind of subject to which the obligation is referred, as in the second example. In such cases, expert annotators might be more reliable.

5 Automatic Classifier

We also used the dataset we built to train an automatic classifier. We split the dataset into training (90%) and test (10%) sets. As a learning

[1] https://www.daitomic.com/

Annotator 1	**Annotator 2**	**Annotator 3**	**text**
not relevant	relevant	relevant	I contratti di assicurazione di cui al comma 1, lettera b), sono corredati da un regolamento, redatto in base alle direttive impartite dalla COVIP [...] *en:[The insurance contracts referred to in paragraph 1, letter b), are accompanied by a regulation, drawn up on the basis of the directives issued by COVIP [...]]*
relevant	relevant	not relevant	Il soggetto incaricato del collocamento nel territorio dello Stato provvede altresi' agli adempimenti stabiliti [...] *en:[The person in charge of placement in the territory of the The State also provides for the established obligations [...]]*

Table 1: Example of disagreement among annotators. Correct classifications are shown in blue while incorrect classifications are shown in red.

	Precision	Recall	F-Score
Not Obligations	0.96	0.98	0.97
Relevant Obligations	0.67	0.63	0.65
Not Relevant Obligations	0.84	0.76	0.80

Table 2: System performances evaluation on the test set

	Precision	Recall	F-Score
Not Obligations	0.96	0.98	0.97
Obligations	0.95	0.87	0.91

Table 3: System performances evaluation on the test set with no distinguish between relevant and not relevant obligations

model, we used UmBERTo[2], an Italian pretrained Language Model trained by Musixmatch based on Roberta architecture (Liu et al., 2019), which has been recently proved to provide state of the art performances for other Italian tasks (Occhipinti et al., 2020; Sarti, 2020; Giorgioni et al., 2020). This language model has 12-layer, 768-hidden, 12-heads, 110M parameters. On top of the language model, we added a ReLU classifier (Nair and Hinton, 2010). All the model's weights has been updated during fine-tuning. We applied dropout (Srivastava et al., 2014) with probability 0.1 to both the attention and the hidden layers. We used Cross-Entropy as a loss function and we trained the system until early-stop at epoch 6. The performances obtained on the test set are reported in Table 2. The system performances are fairly good if compared to IAA but not enough reliable to be used in real-world scenarios. However if we evaluate the system without considering the difference between not relevant and relevant obligations (Table 3) we observe much more accurate results suggesting that the systems, similarly to the annotators, performs well in identifying obligations, but struggles in distinguishing between relevant and not relevant obligations.

6 Human vs Automatic Classification

In order to better understand the model capabilities, we ran a manual error analysis, comparing human annotations against automatic classifications on the test set. We identified some categories of typical errors and reported some examples in Table 4. In some cases, the errors of the model are attributable to the non-explicit subject, which the human annotator can derive from the context, as can be seen in the first example, where it is not explicitly specified who should enter the data in the communication. Looking at the second example, we can see a sentence whose main message is the expression of a right, in this case, the right to access a certain file. However, access to the file is allowed only under certain temporal conditions (*at the conclusion of the appeal procedure*), so behind that right is hidden a relevant obligation. Unfortu-

[2] https://github.com/musixmatchresearch/umberto

Human	Machine	text
not relevant	relevant	Nella comunicazione di avvio di cui al comma 2 sono indicati l'oggetto del procedimento, gli elementi acquisiti d'ufficio [...] *en:[In the communication of initiation referred to in paragraph 2 are indicated the subject of the procedure, the elements acquired ex officio [...]]*
relevant	none	L'accesso al fascicolo è consentito a conclusione della procedura di interpello ai fini della tutela in sede giurisdizionale. *en:[Access to the file is granted at the conclusion of the appeal procedure for judicial protection purposes.]*
relevant	none	E' considerata ingannevole la pubblicità', che, in quanto suscettibile di raggiungere bambini ed adolescenti, può', anche indirettamente, minacciare la loro sicurezza. *en:[Advertising that is likely to reach children and adolescents and that may even indirectly threaten their safety is considered misleading.]*
relevant	not relevant	Le amministrazioni interessate provvedono agli adempimenti previsti dal presente decreto con le risorse umane, finanziarie e strumentali disponibili [...]. *en:[The administrations involved shall carry out the obligations provided for in this decree with the human, financial and instrumental resources available.[...]]*
relevant	none	Il presente decreto reca le disposizioni di attuazione dell'articolo 1 del decreto legge 6 dicembre 2011, n. 201, convertito, con modificazioni, dalla legge 22 dicembre 2011, n. 214 [...]. *en:[This decree contains the provisions for the implementation of article 1 of Law Decree no. 201 of December 6, 2011, converted, with amendments, by Law no. 214 of December 22, 2011 [...]]*

Table 4: Example of disagreement between manual (*Human*) and automatic (*Machine*) annotations. Correct classifications are shown in blue while incorrect classifications are shown in red.

nately in these cases, the model is often wrong. Another difficult case to handle is the one shown in the third example in Table 4. This is a sentence that apparently contains simple information: advertising is considered deceptive if it can threaten the safety of children. But behind this message lies an obligation on advertisers to avoid such a situation. Again, the obligation is not explicit, so it is quite understandable that the model could be wrong. Finally, the last two examples show human errors, and it was noted with some interest that where annotators make errors due to distraction or misunderstanding, the model often classifies correctly.

7 Conclusions

In this work we propose a methodology for building training corpora for obligations classification, based on annotations performed by non-experts. We apply this methodology to a set of heterogeneous regulations from a collection of different legal sources. IAA and a manual error analysis highlight that human annotation is in general prone to errors and that non-expert annotators struggle to distinguish between relevant and not relevant obligations. The dataset produced has been used to train and test an obligations classification system based on state-of-the-art pretrained language models. We conduct both an automatic evaluation and a manual error analysis from which turned out that the system, similarly to human annotators, has good performances in recognizing obligations but struggles in distinguish between relevant and not. As future works, we plan to involve domain-expert annotators to evaluate if their contribution can improve the quality of the data and of the model. Also, we will explore techniques to provide more context to the classifier in order to improve the performances on clauses in which the subject is implied.

References

Roberto Bartolini, Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Claudia Soria. 2004. Automatic classification and analysis of provisions in italian legal texts: a case study. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 593–604. Springer.

Carlo Biagioli, Enrico Francesconi, Andrea Passerini, Simonetta Montemagni, and Claudia Soria. 2005. Automatic semantics extraction in law documents. In *Proceedings of the 10th international conference on Artificial intelligence and law*, pages 133–140.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. 2018. Multi-task learning in deep neural networks at evalita 2018. *Proceedings of the Wvaluation Campaign of Natural Language Processing and Speech tools for Italian*, pages 86–95.

Lorenzo De Mattei, Andrea Cimino, and Felice Dell'Orletta. 2018. Multi-task learning in deep neural network for sentiment polarity and irony classification. In *NL4AI@ AI* IA*, pages 76–82.

Enrico Francesconi and Andrea Passerini. 2007. Automatic classification of provisions in legislative texts. *Artificial Intelligence and Law*, 15(1):1–17.

Simone Giorgioni, Marcello Politi, Samir Salman, Roberto Basili, and Danilo Croce. 2020. Unitor@ sardistance2020: Combining transformer-based architectures and transfer learning for robust stance detection. In *EVALITA*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

Daniela Occhipinti, Andrea Tesei, Maria Iacono, Carlo Aliprandi, Lorenzo De Mattei, and Aptus AI. 2020. Italianlp@ tag-it: Umberto for author profiling at tag-it 2020. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online. CEUR. org*.

Monica Palmirani and Fabio Vitali, 2011. *Akoma-Ntoso for Legal Documents*, pages 75–100. Springer Netherlands, Dordrecht.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Gabriele Sarti. 2020. Umberto-mtsa@ accompl-it: Improving complexity and acceptability prediction with multi-task learning on self-supervised annotations. *arXiv preprint arXiv:2011.05197*.

Amin Sleimi, Nicolas Sannier, Mehrdad Sabetzadeh, Lionel Briand, and John Dann. 2018. Automated extraction of semantic legal metadata using natural language processing. In *2018 IEEE 26th International Requirements Engineering Conference (RE)*, pages 124–135. IEEE.

Amin Sleimi, Marcello Ceci, Nicolas Sannier, Mehrdad Sabetzadeh, Lionel Briand, and John Dann. 2019. A query system for extracting requirements-related information from legal texts. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 319–329. IEEE.

Amin Sleimi, Marcello Ceci, Mehrdad Sabetzadeh, Lionel C Briand, and John Dann. 2020. Automated recommendation of templates for legal requirements. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*, pages 158–168. IEEE.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.