PhD Proposal in Computer Science

A hybrid approach using formal rules and machine learning for the automatic understanding of the structure of the law and its evolution

Abstract. The objective of this thesis is to contribute to improve the intelligibility of the law for citizens and lawyers, by means of a graph representing the structure of the law and its evolution, obtained through the automatic understanding of legal texts. The law, described in a "semi-formal" natural language, will have to be translated into a formal language describing the construction of this graph. The central problem of this translation is to automate the consolidation of legal texts over time by transforming the modification instructions contained in the modifying texts into modification programs.

Keywords. Law, machine learning, NLP, DSL, natural languages, formal languages, regular expressions, grammars.

Context. The texts that compose French and European law are updated by amending texts voted and published in the *Journal officiel de la République française* (JORF) or in the *Journal officiel de l'Union européenne*. The life cycle of a legislative or regulatory text begins with the publication of its full version in a Journal Officiel and continues with the publication of any amending texts. The full amended text, called the consolidated version, is never published in the OJ and has no legal value: only the initial version and the rest of the ordered amendments to the text are authentic [3]. The French Légifrance website[1] indicates:

> *consolidation consists in integrating in a single act, without official value, the successive modifications and corrections made to a text; its objective being to facilitate the knowledge of their rights and obligations by the citizens.*

Since 2008, Légifrance has presented most of the French legal texts in their original versions as well as in their successive consolidated versions, which are the consequences of the modifications made to these texts over time. The French operator of the Légifrance site, the *Direction de l'Information Légale et Administrative* (DILA), manually reports the modifications described in natural language in the texts in order to obtain, at each modification date, the complete consolidated version of the text made available on the site. The same process is at work at the European level, carried out by the *Office des publications de l'Union européenne* (OPOCE)[2].

This convenience of access to texts in a version that is easier to read and use has de facto changed the status of these consolidated versions: they are seen by most users, including legal professionals, as reflecting the applicable law. Moreover, the drafters of new texts, in the parliaments or in the ministries, start from this consolidated version to design the amending texts. It is therefore extremely important that this consolidation work be free of errors and available as soon as possible.

Problem. The Legistix project of Mines Paris aims at understanding the structure of legal texts and their evolutions, with a *legistics* [3] approach, i.e. we do not seek to *understand* the meaning of the law but only the way this law is created, published and modified over time. To achieve this goal, the methods, languages and tools developed within the framework of this project aim at automatically constructing a directed graph representing the elements composing the legislative and regulatory texts, their relations and their evolutions in time. This automatic construction of the graph requires the transformation of the legal texts, considered as written in a "semi-formal" natural language, into a specific formal computer language (DSL) describing the construction of the graph.

Among the problems addressed by the Legistix project, the automatic and reliable consolidation of French and European legal texts is central. Preliminary works [4], based both on regular expressions used in several compound grammars, similar to the successive passes of a compiler, and on a new specialized language of

[1] `https://www.legifrance.gouv.fr/contenu/en-tete/informations-de-mises-a-jour`
[2] `https://eur-lex.europa.eu/collection/eu-law/consleg.html?locale=fr`

functional type, make it possible to describe the changes applied to the texts in the form of programs modifying the Legistix graph.

For each modifier text, Legistix seeks to generate a computer program in this new language in a completely automatic way that, when executed, performs the changes induced by the modifier text on the target texts. In previous work on this topic, presented for example in [2], only the problem of classifying types of modification is addressed. To the best of our knowledge, our work is the first to present a comprehensive approach to identify target texts and to transform the natural language instructions of the modifier text into a computer program in a new specialized language formalizing the actual transformation rules.

Objective. The first results presented [4], using only formal rules based on regular expressions, showed a success rate of the tool largely exceeding that of the prototype developed by DILA indicating a success rate of 50 % [1].

The objective of this thesis is to extend these results by adding a classification phase by machine learning (*machine learning*) of the changes induced by the modifying texts, in order to improve the formal rules to reach a rate of 100 % automation, with a precision and a recall of the detection system both equal to 1. The reliability of the system of rules can be verified thanks to the history of all the texts consolidated manually by the DILA over the last twenty years.

With this hybrid rules/machine learning model, a feedback loop will then have to be studied and implemented: it will allow to compare the classification made by the rules and the one made by learning. This comparison will then allow to manually improve the rule system, by detecting new cases not yet integrated or classification errors. After this manual examination of the differences and modifications of the rules, a partial re-learning will be necessary: it will thus be important to select a learning method that does not make the cost of this re-learning prohibitive. This hybridization should make it possible to maintain automation at 100 % over time.

The generality of the chosen approach, in particular on the specialized language describing text transformations or on the classification model of transformation types, could be evaluated by extending this work to European Union law.

An extension of this work could consist in understanding not only the form but also the nature of the legal changes caused by a transformation. Indeed, some transformations are cosmetic, such as the change of an article number or a name change, others are legislative, others regulatory. Without going so far as to understand the meaning of the law, adding the nature of the changes would be a very useful tool for lawyers to understand the modifications made to the law. Furthermore, linking the exact changes made to a text to the parliamentary discussions that led to that text would be another very useful Legistix development for legal practitioners, especially judges and lawyers, to add to the literal text the legislator's intent when designing the law.

Skills expected and prerequisites. NLP, machine learning, regular expressions, grammars, Python. The candidate must hold an engineering degree or a master of science. Proficiency in french and english, scientific writing in particular, is required.

Location. Centre de recherche en informatique (CRI), Mines Paris, PSL University, Campus Pierre Laffite, Sophia-Antipolis, France.

Supervision. Georges-André Silber `georges-andre.silber@minesparis.psl.eu`, senior lecturer, and Olivier Hermant, professor.

Application. CV, grades, statement of purpose, and letters of recommandation to send to the above email address. Start date of the thesis: 01/10/2023. Application limit: 15/5/2023.

References

[1] Direction de l'information légale et administrative (DILA), ed. *POC Consolidation: un exemple d'innovation au service du droit*. Feb. 7, 2022. URL: https://www.dila.premier-ministre.gouv.fr/actualites/toutes-les-actualites/poc-consolidation-un-exemple-d-innovation-au-service-du-droit.

[2] Samuel Fabrizi et al. "A First Step Towards Automatic Consolidation of Legal Acts: Reliable Classification of Textual Modifications". In: *Proceedings of the Eighth Italian Conference on Computational Linguistics*. July 2022. URL: http://ceur-ws.org/Vol-3033/paper26.pdf.

[3] Secrétariat général du gouvernement and Conseil d'État. *Guide de légistique*. La documentation française, 2017.

[4] Georges-André Silber. "Towards an Automatic Consolidation of French Law". In: *POPL 2023 - Programming Languages and the Law Workshop*. Jan. 2023.