MINES PARIS | PSL

Proposition de thèse de doctorat en informatique

Une approche hybride règles/apprentissage pour la compréhension automatique de la structure du droit et de son évolution

Résumé. Cette thèse a comme objectif de contribuer à rendre le droit plus intelligible pour les citoyens et les juristes, au moyen d'un graphe représentant la structure du droit et son évolution, obtenu grâce à la compréhension automatique des textes juridiques. Le droit, décrit dans un langage naturel "semi-formel", devra être traduit dans un langage formel décrivant la construction de ce graphe. Le problème central de cette traduction est d'automatiser la consolidation des textes de loi au fil du temps en transformant les instructions de modification contenues dans les textes modificateurs en programmes de modification.

Mots-clefs. Légistique, droit, machine learning, NLP, DSL, langages naturels, langages formels, expressions régulières, grammaires.

Contexte. Les textes composant les droits français et européens sont mis à jour par des textes modificateurs votés et publiés dans le Journal officiel de la République française (JORF) ou dans le Journal officiel de l'Union européenne. Le cycle de vie d'un texte juridique législatif ou réglementaire démarre par la publication de sa version intégrale dans un JO et se poursuit par les publications éventuelles de textes le modifiant. Le texte intégral modifié, appelé sa version *consolidée*, n'est jamais publié dans le JO et n'a pas de valeur juridique : seule la version initiale et la suite des modifications ordonnées du texte font foi [4]. Le site français Légifrance [1] indique :

> *la consolidation consiste à intégrer dans un acte unique, sans valeur officielle, les modifications et les corrections successives apportées à un texte ; son objectif étant de faciliter la connaissance de leurs droits et obligations par les citoyens.*

Légifrance présente depuis 2008 la majeure partie des textes juridiques français dans leurs versions d'origines ainsi que dans leurs versions consolidées successives, conséquences des modifications apportées à ces textes dans le temps. L'opérateur français du site Légifrance, la Direction de l'Information Légale et Administrative (DILA), reporte manuellement les modifications décrites en langage naturel dans les textes afin de d'obtenir, à chaque date de modification, la version consolidée complète du texte mis à disposition sur le site. Le même processus est à l'œuvre au niveau européen, effectué par l'Office des publications de l'Union européenne (OPOCE) [2].

Cette commodité d'accès aux textes dans une version plus simple à lire et à utiliser a *de facto* changé le statut de ces versions consolidées : elles sont vues par la plupart des utilisateurs, y compris les professionnels du droit, comme le reflet du droit applicable [3]. De plus, les rédacteurs de nouveaux textes, dans les parlements ou dans les ministères, partent de cette version consolidée pour concevoir les textes modificateurs. Il est donc extrêmement important que ce travail de consolidation soit exempt d'erreurs et disponible le plus rapidement possible.

Problème. Le projet Legistix de Mines Paris a comme but la compréhension de la structure des textes de loi et de leurs évolutions, avec une approche *légistique* [4], c'est-à-dire que nous ne cherchons pas à *comprendre* le sens du droit mais uniquement la manière dont ce droit est créé, publié et modifié dans le temps. Pour atteindre ce but, les méthodes, langages et outils développés dans le cadre de ce projet visent à construire automatiquement un graphe orienté représentant les éléments composants les textes législatifs et réglementaires, leurs relations et leurs évolutions dans le temps. Cette construction automatique de graphe passe notamment par la transformation des textes de loi, considérés comme écrits dans un langage naturel "semi-formel", en un langage informatique formel spécifique (DSL) décrivant la construction du graphe.

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

Parmi les problématiques traitées par le projet Legistix, la *consolidation automatique fiable* des textes de loi français et européens est centrale. Des travaux préliminaires [5], se fondant à la fois sur des expressions régulières utilisées dans plusieurs grammaires composées, similaires aux passes successives d'un compilateur, et sur un nouveau langage spécialisé de type fonctionnel, permettent de décrire les changements appliqués aux textes sous la forme de programmes modifiant le graphe Legistix.

Pour chaque texte modificateur, Legistix cherche à générer de manière complètement automatique un programme informatique dans ce nouveau langage qui, lorsqu'il est exécuté, permet d'effectuer les changements induits par le texte modificateur sur les textes cibles. Dans les travaux antérieurs sur ce sujet, présentés par exemple dans [2], seul le problème de classification des types de modification est abordé. À notre connaissance, nos travaux sont les premiers à présenter une approche complète permettant d'identifier les textes cibles et de transformer les instructions en langage naturel du texte modificateur en un programme informatique dans un nouveau langage spécialisé formalisant les règles effectives de transformation.

Objectif. Les premiers résultats présentés [5], en utilisant uniquement des règles formelles se fondant sur des expressions régulières, ont permis de montrer un taux de réussite de l'outil dépassant largement celui du prototype développé par la DILA indiquant un taux de réussite de 50 % [1].

L'objectif de cette thèse est d'étendre ces résultats en ajoutant notamment une phase de classification par apprentissage automatique (*machine learning*) des changements induits par les textes modificateurs, afin d'améliorer les règles formelles pour atteindre un taux de 100 % d'automatisation, avec une précision et un rappel du système de détection tous deux égaux à 1. La fiabilité du système de règles peut être vérifiée grâce à l'historique de tous les textes consolidés manuellement par la DILA depuis une vingtaine d'année.

Avec ce modèle hybride règles/apprentissage, une boucle de rétroaction devra ensuite être étudiée et mise en place : elle permettra de comparer la classification effectuée par les règles et celle faite par apprentissage. Cette comparaison permettra d'améliorer ensuite manuellement le système de règles, en détectant de nouveaux cas non encore intégrés ou des erreurs de classification. Après cet examen manuel des différences et modifications des règles, un réapprentissage partiel sera nécessaire : il conviendra donc de sélectionner une méthode d'apprentissage ne rendant pas prohibitif le coût de ce réapprentissage. Cette hybridation devrait permettre de maintenir dans le temps l'automatisation à 100 %.

La généricité de l'approche choisie, notamment sur le langage spécialisé décrivant les transformations de texte ou sur le modèle de classification des types de transformation pourra être évaluée en étendant ces travaux au droit de l'union européenne.

Une extension de ces travaux pourra consister à comprendre non par uniquement la forme mais la *nature* des changements juridiques induits par une transformation. En effet, certaines transformations sont cosmétiques, comme le changement d'un numéro d'article ou un changement de nom, d'autres sont législatives, d'autre réglementaires. Sans aller jusqu'à comprendre le sens du droit, l'ajout de la nature des changements serait un outil très utile aux juristes pour comprendre les changements apportés au droit. Par ailleurs, relier exactement les changements opérés sur un texte aux discussions parlementaires ayant abouti à ce texte serait une autre évolution de Legistix très utile aux praticiens du droit, notamment les juges et les avocats pour ajouter au texte litéral l'*intention* du législateur lors de la conception de la loi.

Profil du candidat. NLP, machine learning, expression régulières et grammaires. Langage Python. Master 2 ou diplôme d'ingénieur en informatique. Bon niveau en anglais à l'oral et à l'écrit.

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

Encadrement. Georges-André Silber `georges-andre.silber@minesparis.psl.eu`, maître de conférences et Olivier Herman `olivier.hermant@minesparis.psl.eu`, professeur.

Candidature. CV, notes, lettre de motivation et lettres de recommendations à envoyer à l'adresse email ci-dessus. Un entretien sera effectué pour les candidats sélectionnés. Date de début de la thèse le 01/10/2023, date limite de candidature le 15/5/2023.

Références

[1] Direction de l'information légale et administrative (DILA), éd. *POC Consolidation : un exemple d'innovation au service du droit*. 7 fév. 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*. Juill. 2022. URL : http://ceur-ws.org/Vol-3033/paper26.pdf.

[3] Thierry-Xavier Girardot. « Accéder au droit : importance et défis de la consolidation ». In : *Documentaliste – Sciences de l'Information* 51.4 (2014), p. 30-32. URL : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2014-4-page-30.htm.

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

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