MINES PARIS | PSL

Proposition de thèse de doctorat en informatique

Une approche hybride par règles et apprentissage automatique profond pour la consolidation automatique du droit

Mars 2024

Résumé. Cette thèse se place dans le cadre du projet *Droit Quotidien* qui 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. Ce graphe devra être obtenu par la dérivation du droit, qui est décrit dans un langage naturel "semi-formel", en un langage langage formel décrivant la construction de ce graphe. L'objectif de cette thèse est de résoudre l'un des problèmes centraux de cette dérivation : 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. Parmi les pistes de résolution de ce problème, nous proposons une approche hybride associant des règles symboliques à une ou plusieurs architectures d'apprentissage automatique spécialisées dans le droit.

Mots-clefs. Légistique, droit, machine learning, NLP, LLM, 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 [2]. 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 [4]. 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.

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`

Problème. Le projet Droit Quotidien (DQ) de Mines Paris a comme but la compréhension de la structure des textes de loi et de leurs évolutions, avec une approche *légistique* [2], c'est-à-dire en ne cherchant pas à *comprendre* le sens du droit mais uniquement la manière dont ce droit est créé, publié, cité 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.

Parmi les problématiques traitées par le projet DQ, la *consolidation automatique fiable* des textes de loi français et européens est centrale. Des travaux préliminaires [10], 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 des documents juridiques.

Pour chaque texte modificateur, DQ devra 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 [3], 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 [10], en utilisant uniquement des règles formelles se fondant sur des expressions régulières associées à des grammaires, 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.

Les architectures d'apprentissage automatique développées dans le cadre de cette thèse devront prendre en compte différents aspects. Tout d'abord, la quantité de données disponibles pour le droit français et européen est limitée par rapport aux gigantesques corpus utilisés pour les modèles de langues (LLM) à l'état de l'art. Partant de modèles performants pour le français comme CamenBERT [8, 7], il sera nécessaire d'évaluer si seul un affinage (*fine tuning*) est nécessaire pour la tâche de classification principale ou si l'entraînement d'un vrai modèle le langue spécifique à partir de données juridiques sélectionnées est une meilleure approche [5, 9]. Par ailleurs, s'inscrivant dans l'axe transverse "calcul efficace et sûr" de l'institut des transformations numériques des mines de Paris, il conviendra d'évaluer précisément l'efficacité en temps et en énergie du modèle développé, tout en prenant en compte l'état de l'art de l'entraînement des modèles et de la curation de données [6, 9].

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 aux adresses email ci-dessus. Un entretien sera effectué pour les candidats sélectionnés.

Références

[1] POC Consolidation : un exemple d'innovation au service du droit, February 2022.

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

[3] Samuel Fabrizi, Maria Iacono, Tesei Andrea, and Lorenzo De Mattei. 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.

[4] Thierry-Xavier Girardot. Accéder au droit : importance et défis de la consolidation. *Documentaliste – Sciences de l'Information*, 51(4) :30–32, 2014.

[5] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks Are All You Need, October 2023.

[6] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, March 2022.

[7] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoan Dupont, Laurent Romary, Eric Villemonte de La Clergerie, Benoît Sagot, and Djamé Seddah. Les modèles de langue contextuels Camembert pour le français : impact de la taille et de l'hétérogénéité des données d'entrainement. In Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, and Stéphane Schneider, editors, *JEP-TALN-RECITAL 2020 - 33ème Journées d'Études sur la Parole, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, 22ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 54–65, Nancy / Virtuel, France, June 2020. ATALA.

[8] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT : a tasty french language model. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7203–7219. Association for Computational Linguistics, July 2020.

[9] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb Dataset for Falcon LLM : Outperforming Curated Corpora with Web Data, and Web Data Only, June 2023.

[10] Georges-André Silber. Towards an automatic consolidation of french law. In *POPL 2023 - programming languages and the law workshop*, January 2023.