BRINGAY Sandra
Professeur
LIRMM UM CNRS
Université Paul Valéry Montpellier 3
+33 6 83 24 79 33
bringay@lirmm.fr

Rapport sur le mémoire de doctorat présenté par
Gildas Tagny Ngompé

Méthodes d'analyse sémantique de corpus de décisions jurisprudentielles

Pour l'obtention du grade de **Docteur de l'École Nationale Supérieure des Mines d'Alès**
Spécialité : Informatique
École doctorale : Risque et Société

À qui de droit,

La thèse intitulée *« Méthodes d'analyse sémantique de corpus de décisions jurisprudentielles »* est présentée par Gildas Tagny Ngompé pour l'obtention du doctorat de l'École Nationale Supérieure des Mines d'Alès. Ce travail a été réalisé au sein de l'équipe d'accueil CHROME du LGI2P, sous la direction de Stéphane Mussard (Professeur, Université de Nîmes) et de Jacky Montmain (Professeur, IMT Mines d'Alès). Notons que ce travail a été mené dans un cadre d'une collaboration pluridisciplinaire, dont l'objectif est d'analyser automatiquement la sémantique d'un corpus jurisprudentiel pour mieux comprendre le processus de prise de décision des juges. Plus précisément, ce travail porte sur l'étiquetage des séquences afin de repérer des sections dans les textes, sur l'identification des demandes des parties, sur leur orientation et sur les circonstances factuelles associées. Les compétences principales mises en relief dans ce travail sont liées au traitement automatique des données textuelles, à l'extraction d'informations, à la classification de textes et au regroupement non-supervisé.

Le manuscrit de 169 pages est organisé en 6 chapitres. Le premier chapitre décrit très justement les enjeux liés aux thèmes de recherche de Gildas Tagny Ngompé, à savoir l'analyse sémantique de corpus de décisions jurisprudentielles. Le deuxième chapitre décrit une première contribution portant sur l'annotation de sections et d'entités juridiques via des modèles probabilistes d'étiquetage. Le chapitre 3 porte sur l'identification des demandes via des approches d'extraction d'éléments structurés. Le chapitre 4 vise l'identification du sens du résultat des décisions judiciaires via des algorithmes de classification traditionnels et sur deux adaptations de la méthode Gini-PLS. Le chapitre 5 porte sur le regroupement non supervisé des textes selon les circonstances factuelles évoquées. Chaque chapitre contient un état de l'art des méthodes utilisées ainsi qu'une validation expérimentale. Le dernier chapitre 6 donne un exemple d'utilisation des informations extraites via les méthodes présentées dans les chapitres précédents pour analyser un grand corpus. La conclusion générale du manuscrit est consacrée aux résumés des contributions et aux perspectives. Le manuscrit se termine par la bibliographie.

L'introduction de ce manuscrit dresse un panorama du contexte de l'étude des corpus des décisions judiciaires. Gildas Tagny Ngompé définit précisément les problématiques qui motivent ses travaux à savoir la nature des textes qui suivent des normes peu contraintes (e.g. taille variable, syntaxe, vocabulaire, modalités temporelles et aspectuelles, thématiques, etc.). Par exemple, si les éléments à rechercher dans les textes sont généralement bien connus comme une demande de condamnation, leur expression et leur positionnement dans les textes sont extrêmement variables. Pour finir, le volume de textes est très important et à prendre en compte pour un passage à l'échelle des méthodes automatiques.

Gildas Tagny Ngompé décrit ensuite précisément ses objectifs et en particulier la nature des informations recherchées. Il introduit ses contributions et l'architecture de son manuscrit. Ainsi, cette première partie permet de cerner rapidement la problématique et les enjeux et de naviguer par la suite facilement dans l'ensemble du manuscrit.

Le premier chapitre dresse un état de l'art des méthodes actuelles d'analyse automatique des corpus judiciaires qui s'articulent entre annotation et extraction d'information, classification des jugements et similarité entre décisions judiciaires. Ce panorama montre qu'il s'agit d'un domaine très actif depuis déjà plusieurs décennies, que des enjeux forts sont liés à l'adaptabilité des méthodes à différents sous domaines de la justice et que l'effort d'évaluation quantitative est important. Ce chapitre détaillant le contexte des travaux de Gildas Tagny Ngompé est riche, bien structuré et justifie pleinement les contributions principales de la thèse.

Le chapitre 2 porte sur une première contribution, l'annotation des sections (entête, corps, dispositif) et des entités juridiques (e.g. ville, juridiction, juge, avocat, etc.). Le candidat applique deux approches, les modèles de Markov Cachés (HMM) et les champs aléatoires conditionnels (CRF), adaptées pour l'étiquetage de séquences. Différents choix de caractéristiques pour les segments et différents schémas d'étiquetage sont envisagés. Ce chapitre se termine par des expérimentations poussées incluant une analyse des erreurs. Plus de détails sur le processus d'annotation aurait été appréciables pour mieux comprendre la complexité de la tâche. L'interprétation de l'utilisation de la méthode LDA au cours du processus pourrait être précisée. De plus, la comparaison avec une approche par réseau de neurones ne semble pas tout à fait impartiale selon le protocole mis en place. Toutefois, les analyses présentées permettent de justifier le choix de l'approche CRF mise en avant par Gildas Tagny Ngompé.

Le chapitre 3 décrit la méthodologie développée par Gildas Tagny Ngompé pour identifier les demandes, c'est-à-dire une réclamation par un ou plusieurs parties aux juges. Après avoir formalisé les caractéristiques d'une demande (e.g. catégorie, montant demandé et reçu, etc.), du fait du nombre important de catégories de demandes, le candidat a opté pour une extraction par catégorie. Comme état de l'art, le candidat décrit des travaux connexes sur l'extraction d'éléments structurés comme les évènements et leurs relations. La spécificité des données judiciaires aurait pu être développée. L'approche proposée se décline en deux étapes, extraction de la catégorie puis extraction des caractéristiques proches de sommes d'argent. De nouveau, les expérimentations minutieuses et leurs analyses sont importantes et reposent sur la préparation d'un corpus difficile à constituer. Les résultats sont très intéressants et des pistes sur l'application de méthodes d'apprentissage sont évoquées en perspective. Les exemples pédagogiques donnés dans ce chapitre permettent d'appréhender les principes de chacune des briques de la chaîne de traitements proposée, leurs limites et justifient la nouvelle approche proposée par Gildas Tagny Ngompé.

Le chapitre 4 décrit la méthodologie employée par Gildas Tagny Ngompé pour identifier le sens du résultat (demande acceptée ou rejetée) comme une tâche de classification binaire de documents. Dans un premier temps, le candidat décrit les 12 algorithmes expérimentés combinés à différentes pondérations des termes. Deux adaptations de la méthode Gini-PLS sont également proposées. Deux cas ont été explorés pour classer les documents ou les parties de documents, le deuxième donnant évidement de meilleurs résultats. Le choix de la combinaison algorithme et métrique aurait pu être complété. La question des classes non équilibrées se pose également. Si les résultats de cette méthode sont de nouveau très intéressants, une limite porte sur le paramétrage des différentes configurations car seuls les méta-paramètres par défaut ont été utilisés.

Le chapitre 5 décrit la dernière contribution de Gildas Tagny Ngompé dédiée à la découverte des circonstances factuelles de manière non supervisée. L'originalité de l'approche proposée repose sur l'apprentissage d'une distance basée sur la transformation de documents qui s'avère efficace pour mesurer la dis-similarité sémantique définie par les circonstances factuelles. Gildas Tagny Ngompé a

démontré l'efficacité de l'approche en exploitant une faible quantité de catégorisations manuelles. La question des chevauchements reste posée.

Le dernier chapitre 6 présente une analyse préliminaire descriptive d'un grand corpus de décisions jurisprudentielles qui montre des exemples de statistiques pouvant être construites à partir des informations extraites avec les méthodes présentées dans cette thèse.

Le manuscrit se termine par un chapitre de conclusions et une liste de limitations dans laquelle Gildas Tagny Ngompé présente humblement les limites de ses approches, en particulier la volumétrie de ses corpus. Le candidat décrit ensuite des perspectives qui auraient méritées d'être étoffées mais qui mettent en avant des pistes de recherche très intéressantes comme l'exploration d'autres tâches dont la reconnaissance d'entité nommées ou d'autres approches comme des approches neuronales ou encore d'autres applications comme l'anonymisation et l'analyse des arguments.

Le manuscrit est très agréable à lire et bien documenté. Chaque chapitre contient un état de l'art des méthodes étudiées ainsi que le détail des métriques d'évaluation. De manière générale, il contient de nombreux exemples et illustrations permettant de mettre en valeur le travail accompli avec pédagogie. La thématique de ce travail de recherche est très intéressante et d'actualité car l'analyse automatique des grands corpus judiciaires constitue un élément essentiel pour mieux comprendre les décisions des juges. Le manuscrit met en relief le travail conséquent qui a été intégré au sein d'un projet pluridisciplinaire majeur. Ce travail est novateur par sa vision originale de la spécificité des textes juridiques. Ce travail est également conséquent en termes de tâches investiguées, d'implémentations et d'expérimentations. Les résultats présentés sont globalement très prometteurs.

En conclusion, Gildas Tagny Ngompé a mené un travail riche en proposant des méthodes tout à fait pertinentes qui s'incluent dans une vision large de la fouille de textes pour un domaine spécifique, celui de la justice. Ceci demande une connaissance et une maîtrise large du domaine de la science des données que le manuscrit de Gildas Tagny Ngompé met parfaitement en avant.

Pour toutes ces raisons, j'émets un avis très favorable pour la présentation de ces travaux en vue de l'obtention du titre de Docteur de l'École Nationale Supérieure des Mines d'Alès.

Fait à Montpellier, le 7 janvier 2020

Sandra Bringay
Professeur
LIRMM UM CNRS
Université Paul Valéry Montpellier 3