Analyse sémantique d'un corpus exhaustif de décisions jurispridentielles pour le développement d'un modèle prédictif du risque judiciaire

Gildas Tagny Ngompé, Sébastien Harispe, Jacky Montmain, Stéphane Mussard, Guillaume Zambrano

▶ **To cite this version:**

Gildas Tagny Ngompé, Sébastien Harispe, Jacky Montmain, Stéphane Mussard, Guillaume Zambrano. Analyse sémantique d'un corpus exhaustif de décisions jurispridentielles pour le développement d'un modèle prédictif du risque judiciaire. L'anticipation de la répression : innovation ou régression ?, Jun 2016, Montpellier, France. hal-01485030

HAL Id: hal-01485030
https://hal.science/hal-01485030v1

Submitted on 23 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ANALYSE SEMANTIQUE D'UN CORPUS EXHAUSTIF DE DECISIONS JURISPRUDENTIELLES POUR LE DEVELOPPEMENT D'UN MODELE PREDICTIF DU RISQUE JUDICIAIRE

Gildas Tagny Ngompe[1;2], Sébastien Harispe[1], Jacky Montmain[1], Stéphane Mussard[2], Guillaume Zambrano[2]

[1] Ecole des Mines d'Alès, Laboratoire LGI2P, 69 Rue Georges Besse, Nîmes, France

[2] Université de Nîmes, Equipe CHROME, Rue du Dr Georges Salan, Nîmes, France

RESUME

Comment pouvons-nous observer, analyser, et anticiper la prise de décision des juges sachant que l'interprétation subjective des règles juridiques rend l'application de la loi incertaine ? Afin d'adresser cette question complexe, nous développons une approche automatisée en vue d'étudier de manière exhaustive la jurisprudence, en proposant notamment de résoudre les défis liés à la recherche et à l'analyse du volume croissant de décisions de justice en France. Le projet porte sur les tâches d'extraction d'information des décisions dans l'objectif de construire une base de connaissances jurisprudentielles. Il propose par la suite d'étudier comment une telle base peut être exploitée afin d'estimer le risque au travers de statistiques descriptives et d'analyses prédictives. Nos premiers résultats montrent que nos modèles d'extraction, appliqués au domaine de la jurisprudence et basés sur le modèle probabiliste CRF (« *Conditional Random Fields* »), offrent de bonnes performances pour la segmentation des textes et l'extraction d'entités : numéro du répertoire général (R.G.), la juridiction, la ville, la date, les juges et leurs fonctions.

Mots-clés: analyse sémantique de décisions jurisprudentielles, extraction d'information

1. INTRODUCTION

Les décisions jurisprudentielles sont essentielles pour les professionnels du droit parce qu'elles sont des sources d'interprétation de la loi. Les juristes qui travaillent sur des problèmes particuliers recherchent ainsi des décisions pertinentes et les analysent afin d'observer des tendances leur permettant de mieux comprendre les décisions des juges. Cette analyse est faite quasiment de façon manuelle et est limitée par plusieurs obstacles. D'abord, l'énorme volume de décisions rend impossible des analyses manuelles exhaustives (plus de 4 millions de décisions en France par an). Par ailleurs, la justice est complexe et son langage

difficilement compréhensible[1] pour permettre à un non-juriste d'estimer les conclusions d'une décision sans l'aide d'un initié en droit. Les technologies actuelles de traitement du langage naturel et de fouille de textes peuvent permettre une analyse automatisée de documents afin d'atténuer ces obstacles. Cette analyse automatisée de décisions jurisprudentielles pourra aider non seulement avocats et chercheurs en droit, mais aussi constituer des aides précieuses pour les particuliers et entreprises soucieux de connaitre les chances que leurs requêtes aboutissent. Comment exploiter un corpus de décisions judiciaires pour analyser, voire prédire, les décisions des juges sachant que l'interprétation subjective des règles juridiques rend l'application de la loi, par nature, non déterministe ?

Afin d'adresser cette question complexe, nous développons une approche automatisée en vue d'étudier de manière exhaustive la jurisprudence, en proposant notamment de résoudre les défis liés à la recherche de décisions similaires et pertinentes, et l'analyse du volume croissant de décisions de justice en France. La machine ne peut pas « comprendre » directement un texte brut rédigé en langage naturel. Il lui faut tout d'abord structurer le corpus de décisions à analyser à partir d'informations les caractérisant : numéro du répertoire général (R.G.), la juridiction, la ville, la date, les juges et leur fonction, les normes utilisées, les demandes et leur quantum, et les résultats des juges et leur quantum... Cette formalisation des informations et de leurs relations (e.g. demande fondée sur une norme) permet une classification des décisions en une base de connaissances. L'objectif premier de notre projet vise ainsi à extraire des informations des contenus textuels du corpus de décisions judiciaires et à les normaliser afin de construire une base de connaissances de la jurisprudence française. Les cas d'application pouvant bénéficier d'une telle base sont nombreux, par ex. : mieux comprendre l'application de règles juridiques, anticiper les résultats de juridictions, rechercher des décisions similaires, ou encore identifier les facteurs qui influencent les résultats des juges.

Dans le cadre de notre projet, nous nous concentrons tout d'abord sur la conception d'un système permettant d'extraire des informations de décisions passées. Ensuite, nous construisons un système d'analyse basé sur les fréquences de résultats des juges étant données les demandes correspondantes et les normes appliquées dans un jeu présélectionné de décisions. Ces fréquences sont des informations relatives à la probabilité que les juges prennent certaines décisions dans une situation donnée.

[1] Cretin, Laurette, « L'opinion des Français sur la justice », *INFOSTAT JUSTICE 125*, 2014.

2. APPROCHES D'ANALYSE DES DECISIONS DE JUSTICE

Les ambiguïtés dans les conditions d'application des règles juridiques sont laissées à l'appréciation des juges sur des affaires réelles. Pour étudier les jugements, les approches automatiques courantes sont principalement prédictives. La complexité de l'application de la loi conduit à la défaillance des systèmes experts juridiques[2]. Voilà pourquoi les approches récentes ont été développées sur la base de l'analyse des décisions de justice passées. Le raisonnement par analogie semble logique parce que nous nous attendons à ce qu'une justice équitable rende des résultats similaires sur des affaires similaires. Certaines de ces approches décrivent un cas avec certaines métadonnées à propos de la juridiction, des juges (par exemple leur parti politique), ou l'affaire (par exemple la question juridique de l'affaire). Par exemple, certaines méthodes statistiques, basées sur des arbres de classification binaire, ont été conçues pour prédire les votes des juges de la cour suprême des Etats-Unis d'Amérique. Après une mauvaise performance obtenue avec un seul arbre de décision par juge (66,7% de succès pour l'arbre de décision contre 67,9% pour des experts humains[3]), une approche à base d'arbres extrêmement randomisés[4] offre une précision meilleure mais toujours faible (70,9%). Par ailleurs, un autre système, "SMILE + IBP"[5], repose sur des "Facteurs" prédéfinis qui catégorisent les faits pour classer les affaires. Ainsi, en combinant un raisonnement basé sur les règles et un raisonnement basé sur les cas, "SMILE +IBP" détermine la partie pour chaque question de droit identifiée. Ce système présente une bonne précision (91,8%), mais il n'a été expérimenté que sur des affaires de détournement de secrets d'affaires. Les expériences ont été réalisées sur un nombre très réduit de décisions (184).

Notre projet propose d'adopter une approche différente de ces travaux, en se concentrant sur la définition d'une autre approche pour analyser sémantiquement et statistiquement un grand volume de décisions jurisprudentielles. Nous ciblons aussi un champ d'applications moins réducteur que la simple analyse prédictive ; en visant notamment la mise en place d'analyses descriptives permettant une observation exhaustive de l'application de la loi à différents moments et à différents lieux.

3. DEFIS ET PREMIERS RESULTATS DE L'EXTRACTION D'INFORMATION

[2]Leith, Philip, « The rise and fall of the legal expert system », *European Journal of Law and Technology*, vol. 1(1), 2010, 179-201 p.

[3] Martin, A. D., Quinn, K. M., Ruger, T. W., & Kim, P. T., « Competing approaches to predicting supreme court decision making », *Perspectives on Politics*, Volume 2(4), 2004,761-767 p.

[4] Katz, D. M., Bommarito, M. J., & Blackman, J., « Predicting the behavior of the supreme court of the united states: A general approach », 2014

[5] Ashley, K. D., & Brüninghaus, S., « Automatically classifying case texts and predicting outcomes, Artificial Intelligence and Law », vol. 17(2), 2009, 125-165p.

Malgré la difficulté d'acquérir des décisions judiciaires, nous avons recueilli plus de 600 000 décisions de diverses juridictions. Construire la base de connaissances jurisprudentielle nécessite une description de ces décisions. Les informations pertinentes peuvent être trouvées en analysant le texte non structuré des décisions. Les différentes natures de ces informations imposent diverses tâches d'analyse de texte. Par exemple, les lieux, les dates, les juges, et les normes sont directement reconnaissables dans le texte et leur extraction est similaire au problème de reconnaissance d'entités nommées – problématique largement étudiée en traitement automatique du langage naturel. Cependant, pour découvrir les demandes dans les décisions, des approches novatrices doivent être définies ; nous proposons, pour détecter ces informations, d'appliquer une classification des décisions. Après avoir identifié les demandes dans une décision, la machine doit identifier et interpréter la conclusion des juges pour savoir si la demande a été rejetée ou accepté.

Nous remarquons que les informations sont distribuées dans trois sections principales sur lesquelles nous avons choisi d'organiser l'extraction séparément : l'entête (références de l'affaire), le texte principal (normes invoquées, prétentions et quantums demandés), et la conclusion (normes appliquées, résultats, et quantums accordés). Afin de faciliter la tâche d'extraction, nous proposons dans un premier temps de sectionner le contenu des décisions. Il semblait logique au premier abord qu'un système basé sur des règles puisse facilement détecter ces sections. Un tel système a été réalisé mais il a montré ses limites en suggérant plusieurs schémas de segmentation pour un même document et un grand nombre de mauvaises segmentations dues à la détection de motifs de segmentation à des emplacements erronés dans les documents.

De nombreux travaux démontrent l'efficacité des modèles graphiques probabilistes dans certaines tâches d'étiquetage de séquences comme la segmentation automatique des foires aux questions (FAQs) en distinguant les questions des réponses [6] ou l'extraction d'entités nommées dans l'entête de publications scientifiques[7].

Ainsi, nous avons conçu et comparé deux approches basées sur les modèles probabilistes : HMM (« *Hidden Markov Model* ») et CRF (« *Conditional Random Field* »). Les deux approches déterminent la section à laquelle appartient chaque ligne d'un document. Pour le

[6] McCallum, A., Freitag, D., & Pereira, F. C., « Maximum Entropy Markov Models for Information Extraction and Segmentation », *Icml*, vol. 17, 2000, 591-598 p.
[7] Peng, F., & McCallum, A., « Accurate information extraction from research papers using conditional random fields », *Information Processing & Management*, vol. 42(4), 2006, 963-979 p.

modèle basé sur le CRF, nous avons annoté manuellement un ensemble de 543 décisions de la Cour d'appel de Nîmes et nous avons utilisé certaines caractéristiques des lignes (numéro, longueur,...). Nous avons utilisé 80% de l'ensemble des documents pour l'apprentissage et 20% pour les tests. L'excellente performance de notre modèle à base de CRF démontre son efficacité dans la segmentation des décisions.

Ensuite, nous avons conçu et comparer les deux modèles d'extraction d'entités basés respectivement sur le HMM et le CRF. L'objectif de ces systèmes est d'étiqueter chaque mot dans les en-têtes avec l'étiquette de l'entité auquel appartient le mot. Après l'étiquetage manuel des entités dans les 543 en-têtes de l'ensemble précédent de décisions, nous avons défini quelques caractéristiques des mots pour le CRF (la position dans la ligne, le nombre de caractères). Avec les deux modèles entrainés avec 80% des données, nous avons observé que le système basé sur le CRF surpasse le système à base de HMM. Cependant, bien que le système à base de CRF semble être très précis pour détecter certaines entités (par exemple les villes, les compétences), il semble moins efficace pour reconnaitre le nom des parties (plaignant et défendeurs). Certaines caractéristiques plus discriminantes sont actuellement étudiées afin d'améliorer la performance du système à base de CRF.

4. CONCLUSION

Au cours de notre travail, nous avons recueilli plus de 600 000 décisions de diverses juridictions. Nous avons étiqueté un ensemble de 543 décisions pour la segmentation et l'extraction d'entités dans les entêtes. L'étiquetage des exemples d'apprentissage a demandé beaucoup de temps et d'effort car nous avons dû le faire manuellement. En outre, certaines caractéristiques ont été définies et calculées pour les systèmes à base de CRF. Nous avons obtenu d'excellents résultats sur notre jeu de données de test, même si certaines entités restent difficiles à reconnaitre.

Les travaux futurs se concentreront tout d'abord sur la reconnaissance d'autres entités avec une priorité mise sur les normes. Ensuite, nous avons l'intention d'utiliser les normes pour découvrir automatiquement les demandes disponibles dans notre corpus - des demandes similaires sont généralement fondées sur des normes similaires ; nous voulons donc regrouper les décisions de telle sorte que chaque catégorie représente une demande.