Résumé

Titre : ANALYSE SÉMANTIQUE D'UN CORPUS EXHAUSTIF DE DÉCISIONS JURISPRUDENTIELLES

Une jurisprudence est un corpus de décisions judiciaires représentant la manière dont sont interprétées les lois pour résoudre un contentieux. Elle est indispensable pour les juristes qui l'analysent pour comprendre et anticiper la prise de décision des juges. Son analyse exhaustive est difficile manuellement du fait de son immense volume et de la non-structuration des documents. L'estimation du risque judiciaire par des particuliers est ainsi impossible car ils sont en outre confrontés à la complexité du système et du langage judiciaire. L'automatisation permettrait de retrouver exhaustivement des connaissances pertinentes pour structurer la jurisprudence à des fins d'analyses descriptives et prédictives. Afin de rendre la compréhension de la jurisprudence exhaustive et plus accessible, cette thèse aborde l'automatisation de tâches d'importante d'analyse métier de la jurisprudence. En premier, est étudiée l'application de modèles graphiques probabilistes d'étiquetage de séquences pour la détection des sections, d'entités nommées juridiques, et de citations de lois dans la décision. Ensuite, les catégories prédéfinies de demandes sont identifiées par classification de documents. Pour chaque catégorie reconnue, sont extraites les demandes des parties. L'approche proposée pour la reconnaissance des quanta demandés et accordés exploite la proximité entre les sommes d'argent et des termes-clés extraits automatiquement. Nous montrons par ailleurs que le sens du résultat des juges est identifiable soit à partir de termes-clés prédéfinis soit par classification de documents. Enfin, les situations ou circonstances factuelles où est formulée une catégorie de demandes sont découvertes par regroupement des décisions. A cet effet, une méthode d'apprentissage d'une distance de similarité est proposée et comparée à des distances établies. Cette thèse discute des résultats empiriques obtenus sur des données réelles annotées manuellement par un expert. Le mémoire est clôturé par une démonstration d'applications à l'analyse descriptive d'un grand corpus de décisions judiciaires françaises.

Mots clés : analyse de données textuelles, décisions jurisprudentielles, extraction d'information, classification de textes, regroupement non supervisé.

Abstract

Title : A SEMANTIC ANALYSIS OF A COMPREHENSIVE CORPUS OF JUDICIAL DECISIONS

A case law is a corpus of judicial decisions representing the way in which laws are interpreted to resolve a dispute. It is essential for lawyers who analyze it to understand and anticipate the decision-making of judges. Its exhaustive analysis is difficult manually because of its huge size and the non-structuring state of the documents. The estimation of the judicial risk by individuals is thus impossible because they are also confronted with the complexity of the judicial system and language. Automation can enable an exhaustive extraction of relevant knowledge for structuring case law for descriptive and predictive analysis. In order to make the comprehension of the case-law exhaustive and more accessible, this thesis deals with the automation of important tasks of business analysis of jurisprudence. First, the application of probabilistic graphical sequence labeling models for the detection of sections, legal named entities, and legal rules citations in decisions is investigated. Then, predefined categories of requests are identified by document classification. For each recognized category, the requests of the parties are extracted. The proposed approach to the recognition of claimed and granted quanta exploits the proximity between money mentions and automatically extracted key-phrases. We also show that the meaning of the judges' result is identifiable either from predefined key terms or by classification of documents. Lastly, situations or factual circumstances in which a category of claims is formulated are discovered by clustering decisions. For this purpose, a method of learning a similarity distance is proposed and compared with established distances. This thesis discusses the empirical results obtained on real data annotated manually by an expert. The thesis is closed by a demonstration of some applications to the descriptive analysis of a large corpus of French judicial decisions.

Keywords : textual data analysis, case law decisions, information extraction, text classification, document clustering

Table des matières

Liste des figures

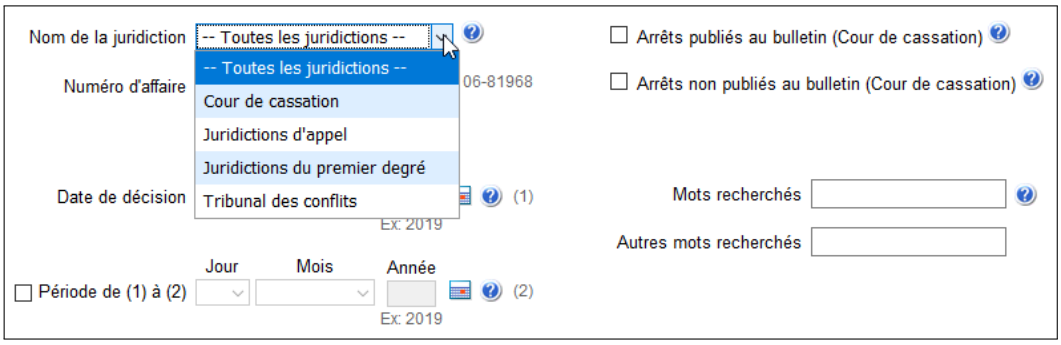
Liste des tableaux

Introduction générale

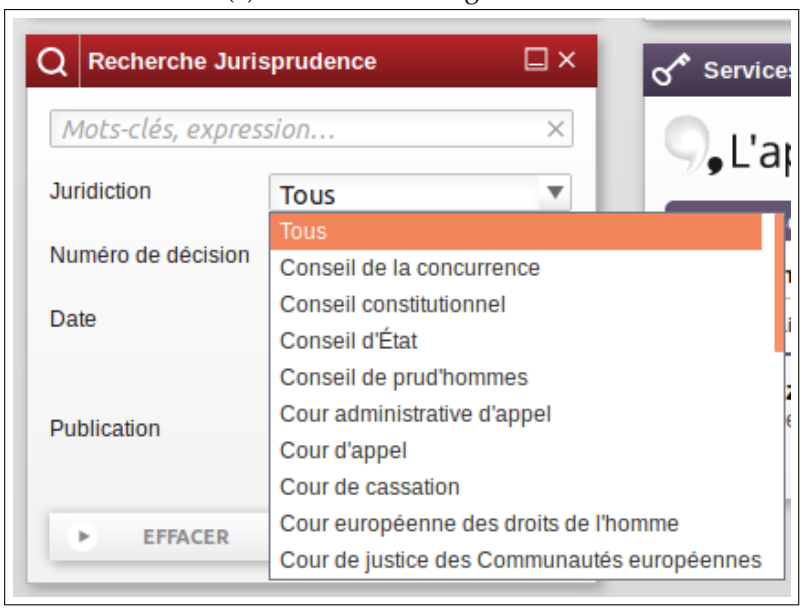
i Contexte et motivations

Une décision judiciaire peut être définie soit comme le résultat rendu par les juges à l'issue d'un procès, soit comme un document décrivant une affaire judiciaire. Un tel document rapporte, notamment, les faits, les procédures judiciaires antérieures, le verdict des juges, et les explications associées. Dans cette thèse, nous désignons par « décision » le document, et par « résultat » une conclusion ou réponse des juges. Une jurisprudence est un ensemble de décisions rendues par les tribunaux. Elle représente la manière dont ces derniers interprètent les lois pour résoudre un problème juridique donné (type de contentieux). Les juristes doivent alors collecter des décisions traitant de situations similaires, les sélectionner, et les analyser afin de mener, par exemple, des recherches empiriques en droit [Ancel, 2003; Jeandidier & Ray, 2006]. Les avocats exploitent aussi les décisions passées pour anticiper les résultats des juges. Ils peuvent ainsi mieux conseiller leurs clients sur le risque judiciaire que ces derniers encourent, et sur la stratégie à adopter pour faire accepter leurs demandes et faire rejeter celles de leurs adversaires. Cette activité de collecte et d'analyse, centrale pour de nombreux métiers du droit, est généralement effectuée manuellement. Elle est par conséquent sujette à plusieurs difficultés liées à l'accès et à l'exhaustivité des documents traités même lors de l'étude d'une question spécifique. Il faut notamment souligner ici que les documents sont dispersés dans les nombreux tribunaux, et que les procédures administratives ne facilitent pas toujours leur accès du fait de la nécessité de préserver la confidentialité des parties. En effet, les décisions n'étant pas « anonymisées » la plupart du temps, elles restent alors inaccessibles aux juristes qui en font la demande. Un certain nombre de documents sont néanmoins accessibles sur internet grâce à des sites de publication de don-

nées ouvertes gouvernementales[1]. Ces sites publient régulièrement des décisions récemment prononcées.



(a) Formulaire de Légifrance



(b) Formulaire de Dalloz

Figure 1 – Exemples de critères des moteurs de recherche juridique

Il existe aussi des moteurs de recherche juridiques qui permettent de retrouver des décisions intéressantes. Cependant, qu'ils soient payants (Lexis-

1. Données ouvertes gouvernementales : data.gouv.fr en France, judiciary.uk en Grande-Bretagne, scotusblog.com aux Etats-Unis, et scc-csc.ca au Canada.

Nexis [2], Dalloz [3], Lamyline [4],...) ou gratuits (CanLII [5], Légifrance [6], ...), les critères de recherche offerts par leurs moteurs de recherche limitent grandement la pertinence des résultats pouvant être obtenus. En effet, il ne s'agit en général que de combinaisons de mots-clés et autres méta-données (date, type de juridiction, ...), ou d'expressions régulières, comme l'illustre la Figure 1 Page 2. La manipulation de tels critères est difficile pour constituer des échantillons pertinents suivant une sémantique souhaitée tels que l'ensemble des décisions traitant d'une catégorie de demande ou d'une circonstance factuelle donnée.

Justice	**2013**	**2014**	**2015**	**2016**	**2017**
civile	2 761 554	2 618 374	2 674 878	2 630 085	2 609 394
pénale	1 303 469	1 203 339	1 206 477	1 200 575	1 180 949
administrative	221 882	230 477	228 876	231 909	242 882

Source : http://www.justice.gouv.fr/statistiques-10054/chiffres-cles-de-la-justice-10303/

Tableau 1 – Nombre de décisions prononcées en France par an de 2013 à 2017

Plus de 4 millions de décisions sont prononcées en France chaque année d'après les chiffres du ministère français de la justice (Tableau 1 Page 3). Dans ce contexte, l'analyse manuelle ne peut être limitée qu'à une infime proportion de documents disponibles. En effet, au regard de la croissance rapide du nombre de décisions, même une étude sur une question très précise nécessite la constitution d'un large corpus de décisions pertinentes. Par ailleurs, il peut s'avérer très pénible de les lire pour en identifier les données d'intérêt. Les documents sont très souvent longs et complexes dans leur style de rédaction. Par exemple, Certaines phrases comprennent plusieurs clauses discutant d'aspects différents (Figure 2 Page 4). On y retrouve aussi des références à des jugements antérieurs (Figure 3 Page 4).

Il est évident qu'une automatisation du traitement des corpus de décisions s'impose pour répondre aux diverses difficultés d'accès, de volumétrie, et de complexité liées à la compréhension des décisions. Une telle

2. https://www.lexisnexis.fr/
3. http://www.dalloz.fr
4. http://lamyline.lamy.fr
5. https://www.canlii.org
6. https://www.legifrance.gouv.fr

Exposant subir un trouble anormal de voisinage pour être privée d'une vue sur la mer dont elle disposait auparavant, ce en raison de l'absence de taille de haies implantées à proximité de son jardin privatif, elle a attrait devant le juge des référés du tribunal de grande instance de Marseille, le syndicat des copropriétaires LES CATALANS (ci-après désigné : le syndicat des copropriétaires) , et son syndic recherché personnellement, le Cabinet L., à l'effet, au visa de l'article 809 du code de procédure civile d'obtenir leur condamnation sous astreinte de 200 euros par jour de retard à tailler les haies qui bouchent sa vue et la condamnation personnelle du Cabinet L. à lui régler une provision de 2.000 euros à valoir sur dommages et intérêts, outre 1.500 euros sur le fondement des dispositions de l'article 700 du code de procédure civile.

***Source** : extrait de la décision R.G. 15/10226 de la Cour d'Appel d'Aix-en-Provence du 2 Juin 2016*

Figure 2 – Exemple de phrases composée de plusieurs clauses dont une demande de condamnation sous astreinte, une autre de dommages et intérêts pour trouble anormal de voisinage, et une dernière de dommages et intérêts sur l'article 700 du code de procédure civile.

Vu le jugement du tribunal de grande instance de Versailles du 5 décembre 2013 qui a :
- rejeté la demande de démolition de la construction litigieuse,

...
SUR CE LA COUR,

...
PAR CES MOTIFS,
...

Confirme le jugement en toutes ses dispositions à l'exception de celle relative au montant des dommages et intérêts ...

***Source** : extrait de la décision R.G. 14/01640 de la Cour d'Appel de Versaille du 7 Avril 2016*

Figure 3 – Exemple de référence à un jugement antérieur dans une décision d'appel.

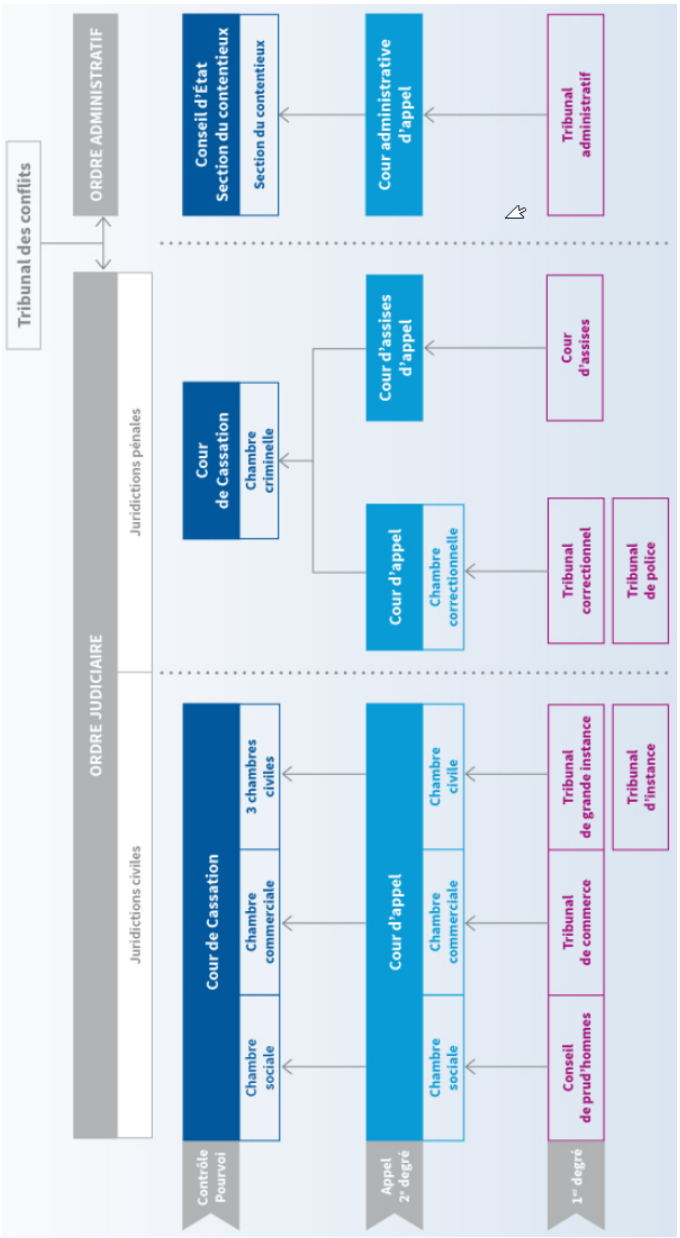
automatisation ferait gagner du temps aux juristes lors de tâches d'analyse métier préalables à leur raisonnement d'experts, tout en leur fournissant une vue exhaustive de la jurisprudence. D'autre part, Cretin [2014] fait remarquer que la justice est complexe dans son organisation (Figure 4 Page 6) et son fonctionnement, et que son langage est peu compréhensible. Il est donc presque impossible pour les profanes en droit d'estimer leurs droits et le risque judiciaire qu'ils encourent dans leur quotidien sans consul-

ter un initié du droit. L'exigence pour le profane étant l'exacte pertinence des ressources, leur accessibilité, et l'intuitivité du processus de leur exploitation [Nazarenko & Wyner, 2017], l'automatisation de l'analyse de la jurisprudence pourrait ainsi améliorer l'accessibilité du droit dans d'indénombrables situations. Par exemple, en comparant le montant qu'on peut espérer d'une juridiction et le coût d'un procès, on peut plus aisément se décider entre un arrangement à l'amiable et la poursuite du litige en justice [Langlais & Chappe, 2009]. Le traitement automatique de la jurisprudence constituerait alors une aide précieuse non seulement pour les professionnels du droit, mais aussi pour les particuliers et entreprises tous soucieux de voir l'issue de leur affaire leur être favorable.

ii Objectifs

Ce mémoire propose des approches pour automatiser l'extraction de connaissances judiciaires à partir des décisions françaises. Le but est de faciliter la structuration et l'analyse descriptive et prédictive de corpus de décisions de justice en adressant les difficultés de l'approche traditionnelle d'analyse de contentieux. L'étude de la jurisprudence pour un contentieux donnée consiste à [Ancel, 2003] :

1. **Choisir un échantillon représentatif** : Des décisions sont collectionnées suivant des contraintes définies : période précise, couverture géographique, types d'affaires, etc.
2. **Sélectionner les décisions** : élimination des décisions qui ne correspondent pas au type de demande d'intérêt.
3. **Élaborer la grille d'analyse** : Un modèle de grille est créé et permet d'enregistrer les informations potentiellement importantes. Chaque ligne de la grille correspond à une demande, et les colonnes font référence aux différents types d'informations qu'il est possible d'extraire sur une demande. Ces variables vont de la procédure suivie, aux solutions proposées, en passant par la nature de l'affaire. Les champs à remplir ne sont pas connus à l'avance ; ce n'est généralement qu'au cours de la lecture des décisions que l'on distingue les informations pertinentes pour l'étude.
4. **L'analyse des décisions et l'interprétation des informations** : Les informations retrouvées dans les décisions sont saisies dans la grille,



Source : http://www.justice.gouv.fr/organisation-de-la-justice-10031/

Figure 4 – Organisation des institutions judiciaires françaises

et des calculs statistiques sont effectués par la suite.

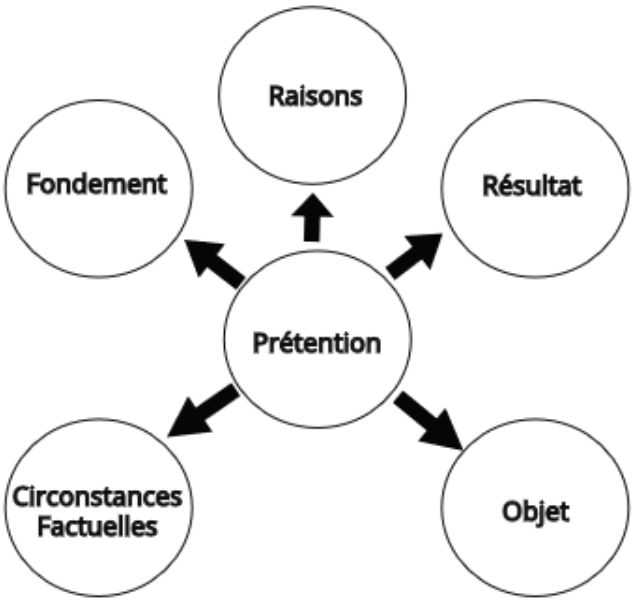
Ancel [2003] évoque principalement le problème de la différence entre l'état capté de la jurisprudence et son état présent. En effet, les longs délais de travail sont caractéristiques de ces études. L'étude de son équipe portait sur les décisions d'expulsion d'occupants sans droit ni titre. La saisie des informations à elle seule a duré 9 mois. De plus, il est difficile d'observer l'évolution des pratiques judiciaires dans le temps et leur différence entre les villes du fait de la faible taille de l'échantillon choisi. Par exemple, Jeandidier & Ray [2006] n'ont analysé que 399 dossiers d'affaires de pension alimentaire correspondant aux audiences s'étalant de fin 1999 à fin 2000 d'un seul tribunal de grande instance. L'équipe de Ancel [2003] n'a quant à elle analysé que 3865 décisions sélectionnées parmi 5656 décisions rendues du 1er juillet au 31 décembre 2001.

La problématique de notre étude est « **comment donner accès à l'analyse automatique de la sémantique d'un corpus jurisprudentiel pour comprendre la prise de décision des juges ?** ». La complexité de cette analyse s'explique notamment par l'interprétation subjective des règles juridiques, l'application non déterministe de la loi, et la technicité du langage judiciaire. Cette problématique intéresse des entreprises telles que LexisNexis [7] et Lexbase SA [8], et plusieurs startups telles que Predictice [9] et CASE LAW ANALYTICS [10]. Afin d'y répondre, nous nous intéressons aux concepts d'informations mentionnées dans les décisions, au centre desquels se trouvent les demandes des parties (prétentions) sur lesquelles portent les conclusions rendues. Ainsi, l'analyse sémantique d'un corpus jurisprudentiel vise l'identification de connaissances sur les demandes (Figure 5 Page 8).

Une demande peut être caractérisée par :

- l'objet qui a été demandé (par ex. dommages et intérêts) quantifié par un quantum ;
- le résultat associé qui est décrit par une polarité (« accepte » ou « rejette »), souvent lié à un quantum accordé, par exemple 5000 euros de dommages et intérêts ou 2 mois d'emprisonnement ;
- le fondement ou la norme juridique qui est la règle qui légitime la prétention ou le résultat ;

7. https://lexmachina.com
8. https://www.legalmetrics.fr
9. http://predictice.com
10. http://caselawanalytics.com



Sens des flèches : l'extrémité est l'information sur l'origine.

Figure 5 – La demande au centre de la compréhension des décisions

- les circonstances factuelles définissent des types d'affaires et qui caractérisent les différentes situations dans lesquelles sont formulées les demandes d'une catégorie donnée ;
- les divers arguments apportés par les parties pour justifier leurs requêtes (raisons des demandes).
- les motivations des solutions des juges (raisons des résultats)

Comme illustré par la Figure 6 Page 9, l'analyse sémantique identifie ou découvre différentes informations descriptives d'un corpus constitué par des décisions collectées à partir de divers moteurs de recherche juridique et des juridictions. Cette thèse s'inscrit dans un projet qui vise, entre autres, à automatiser la constitution d'une base de connaissances sur la jurisprudence française. Une telle base permettrait notamment de mener une grande variété de recherches et d'études expertes. Elle aurait aussi naturellement une importance certaine pour la définition de modèles prédictifs par exemple pour la prédiction des types de demandes à formuler et la prédiction de la solution des juges.

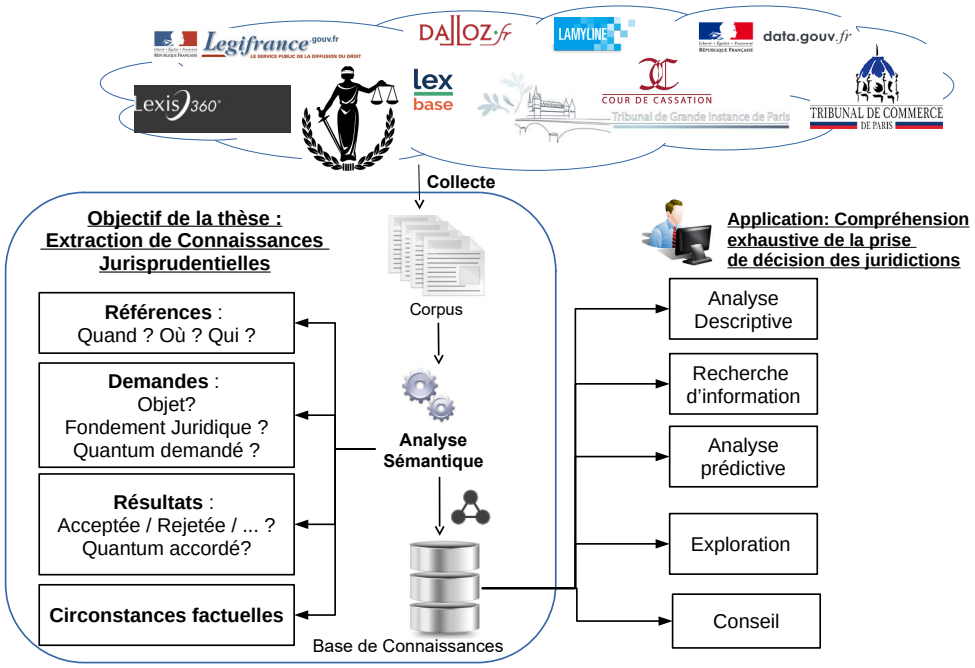


Figure 6 – Objectifs et applications de la thèse

ii.a Collecte, gestion et pré-traitement des documents

Il est nécessaire de trouver des moyens pour collecter le maximum de documents bruts non-structurés, les pré-traiter, et organiser leur gestion afin de les indexer pour faciliter leur traitement. Les décisions de cours d'appel de justice civile sont les plus accessibles à partir des moteurs de recherche juridique (Lexis360, Dalloz, LamyLine, Lexbase, Legifrance, etc.) et de la grande base de données JuriCa alimentée d'environ 180k décisions civiles par an [Lamanda, 2010]. Cependant, l'accès à ces décisions est généralement payant, et le nombre de documents simultanément téléchargeables est très faible sur les sites payants (généralement 10 à 20 décisions au maximum à la fois). La base JuriCa est la plus grosse base de décisions de cours d'appel en France. Elle est gérée par la Cour de cassation. L'accès à cette base est offert par le Service de Documentation, des Etudes et du Rapport [11] (SDER). L'accès est payant pour les professionnels et gratuit

11. https://www.courdecassation.fr/institution_1/composition_56/etudes_rapport_28.html

pour les universités et centres de recherche en partenariat avec le SDER. Lexbase dispose depuis une dizaine d'année d'une licence pour vendre les décisions de JuriCa [12]. Légifrance, le moteur de recherche du ministère de la justice, fournit quant à lui un accès public et gratuit à un nombre considérable de documents. Les décisions y sont identifiées à l'aide de numéros consécutifs et accessibles à partir d'un service Web. Ce dernier a l'avantage de proposer des décisions de tous les ordres et de tous les degrés. Cependant, les décisions des juridictions du premier degré (appelées jugements) restent plus rares sur internet et principalement disponibles auprès des tribunaux. Il faut préciser que nos expérimentations se sont concentrées sur les décisions d'appel en justice civile, et ce choix a été motivé par le fait qu'elles sont les plus disponibles.

ii.b Extraction de connaissances

La difficulté d'extraire de telles informations découlent de l'état non-structuré des documents, et de la complexité et la spécificité du langage employé. L'extraction des connaissances nécessite de mettre en œuvre des techniques de fouille de textes adaptées à la nature des éléments à identifier. Nous avons ainsi abordé l'annotation des références de l'affaire (juridiction, ville, participants, juges, date, numéro R.G., normes citées, ...), l'extraction des demandes et résultats correspondants, et l'identification des circonstances factuelles.

Les méta-données de référence sont des segments de texte qu'on peut directement localiser dans le document. Elles sont donc semblables aux entités nommées dont la reconnaissance est une problématique intensivement étudiée en traitement automatique du langage naturel [Yadav & Bethard, 2018] dans plusieurs travaux et compétitions, aussi bien pour des entités communes [Tjong Kim Sang & De Meulder, 2003; Grishman & Sundheim, 1996], que pour des entités spécifiques à un domaine [Kim *et al.* , 2004; Persson, 2012; Hanisch *et al.* , 2005], et dans diverses langues [Li *et al.* , 2018; Alfred *et al.* , 2014; Amarappa & Sathyanarayana, 2015].

Le problème d'identification des demandes consiste à reconnaître, dans la décision analysée, l'objet, le fondement, le quantum demandé, le sens du résultat correspondant, et le quantum accordé de chaque prétention.

12. Arrêts des cours d'appel : la base JURICA enfin en service chez Lexbase par Emmanuel Barthe `https://www.precisement.org/blog/Arrets-des-cours-d-appel-la-base.html`

La demande s'apparente donc aux entités structurées telles que les évènements ACE [2005] qui sont décrits par un type, un terme-clé, des participants, un temps, une polarité.

Le problème d'identification des circonstances factuelles consiste à constituer des regroupements de décisions mentionnant une certaine catégorie de demande (objet+fondement). Le but est, comme indiqué précédemment, de repérer les différentes situations dans lesquelles cette catégorie de demande est formulée. Chacun des groupes représente donc une situation particulière partagée par les membres du groupe mais bien distinctes de celles reflétées par les autres groupes. Ce problème évoque des problématiques de similarité entre textes, de catégorisation non-supervisée (*clustering*), et de « modélisation thématique » (*topic modeling*).

A l'issue du processus d'extraction, les données extraites sont destinées à enrichir progressivement une base de connaissances. La structuration des données au sein d'une base facilite les diverses analyses automatiques applicables aux décisions et demandes judiciaires.

ii.c Analyse descriptive

L'analyse descriptive exploite l'ensemble des connaissances extraites et organisées pour répondre aux diverses questions que l'on pourrait se poser sur l'application de la loi. Il est intéressant par exemple de comparer les fréquences de résultats positifs et négatifs pour une catégorie de prétention donnée dans une situation précise. Les quanta extraits servent à visualiser les différences entre les montants accordés et réclamés. D'autres analyses plus complexes permettraient d'étudier l'évolution dans le temps et les différences dans l'espace de l'opinion des juges.

iii Méthodologie

Les tâches sont définies par le métier. Comme illustrées précédemment (§ ii.b), les problématiques propres aux textes juridiques trouvent généralement des analogies avec les problèmes d'analyse de données textuelles (*text mining*). Ainsi, les méthodes issues de ce domaine sont applicables aux textes juridiques. Cependant, quelques adaptations sont généralement nécessaires pour obtenir des résultats de bonne qualité hors des domaines pour lesquels ces approches ont été développées [Waltl *et al.* , 2016]. De

plus, la recherche en fouille de textes est souvent réalisée sur des échantillons qui ne reflètent pas toujours la complexité des données réelles. Effectuant l'une des premières études d'analyse sémantique des décisions françaises, nous avons axé notre travail sur le rapprochement des problèmes liés à l'analyse des décisions jurisprudentielles de ceux généralement traités en analyse de textes. Il s'agit ensuite d'établir des protocoles d'évaluation et d'annotation manuelle de données. Selon les problématiques identifiées et les protocoles d'évaluations définis, des méthodes adaptées ont été proposées et expérimentées sur les données réelles annotées manuellement par un expert juriste.

iv Résultats

Une chaîne de traitement pour le sectionnement et l'annotation des méta-données est proposée. Premièrement, le sectionnement a pour but d'organiser l'extraction des informations qui sont réparties dans des sections selon leur nature. L'applicabilité de deux modèles probabilistes, les champs aléatoires conditionnels ou CRF (*conditional random fields*) et les modèles cachés de Markov ou HMM (*hidden Markov Model*), est étudiée en considérant plusieurs aspects de la conception des systèmes d'extraction d'entités nommées.

Par la suite, nous proposons une méthode d'extraction des demandes et des résultats en fonction des catégories présentes dans la décision. L'approche consiste en effet à identifier dans un premier temps les catégories (objet+fondement) présentes par classification supervisée. Un vocabulaire d'expression des demandes et résultats est exploité pour identifier les passages. Puis à l'aide de termes propres à chacune des catégories identifiées, les trois attributs (quantum demandé, sens du résultat, quantum accordé) des paires demande-résultat sont reconnus.

Par ailleurs, nous analysons l'extraction du sens du résultat par classification binaire des documents. L'objectif est de s'affranchir de l'identification préalable de l'expression des demandes et résultats. En effet, les décisions comprenant des demandes d'une catégorie donnée semblent ne contenir, dans une forte proportion, qu'une seule demande. A partir d'une représentation adéquate du contenu de la décision, il est possible de classer cette dernière à l'aide d'un modèle de classification supervisée de documents.

L'identification des circonstances factuelles, quant à elle, est modélisée comme une tâche de regroupement non supervisé de documents. Nous proposons dans ce cas une méthode d'apprentissage d'une distance entre textes, à l'aide d'un algorithme de régression. La métrique apprise est utilisée dans l'algorithme des « K-moyennes » (*k-means*) [Forgey, 1965] et celui des « K-medoïdes » (*k-medoids*) [Kaufman & Rousseeuw, 1987], et comparée à d'autres distances établies en recherche d'information.

v Structure de la thèse

La thèse est organisée en 6 chapitres. Le chapitre **??** positionne nos travaux par rapport à ceux qui ont été réalisés précédemment sur des problématiques d'analyse automatique de décisions de justice. Le chapitre **??** présente les architectures et modèles proposés pour la structuration des décisions et la reconnaissance des entités juridiques ; il discute notamment des différents résultats empiriques obtenus par application des modèles CRF et HMM. Ensuite, le chapitre **??** détaille le problème d'extraction des demandes, puis présente notre méthode et les résultats obtenus. Le chapitre **??** traite de l'identification du sens du résultat par classification directe des décisions, cela en comparant différents algorithmes de classification et de représentations des textes. Le chapitre **??** discute de l'usage de l'apprentissage proposé d'une distance qui est comparée à d'autres distances pour la découverte des circonstances factuelles. Enfin, le chapitre **??** présente des résultats de scénarios d'analyses descriptives pour illustrer l'exploitation potentielle de nos propositions sur un corpus de grande taille.

Conclusion générale

Pourquoi n'avoir pas utilisé des méthodes de deep learning la thèse ? disponibilité des approches de l'état de l'art, peu de données labellisées.

i Évaluation des contributions

Cette thèse porte essentiellement sur la proposition et l'exploration d'approches adressant des problèmes d'analyses de données textuelles rencontrés lors de l'étude de corpus jurisprudentiels par des experts juristes. Trois problèmes principaux y sont abordés. Premièrement, l'annotation, dans les documents, des sections de textes et des entités nommées propres au domaine judiciaire qui peuvent aider à se repérer dans le document et à améliorer la recherche d'information. Le chapitre **??** démontre empiriquement, sur des documents annotés pour la circonstance, l'efficacité de l'application de modèles probabilistes d'étiquetage de séquences, HMM et CRF, sur les deux tâches. Par la suite, l'extraction de données relatives aux demandes, suivant leur catégorie juridique, est discutée dans les chapitres **??** et **??**. Le problème impose d'effectuer les extractions pour une catégorie de demande à la fois car il est impossible d'annoter suffisamment de données pour toutes les catégories prédéfinies. Pour cela nous proposons de filtrer à l'entrée les documents de la catégorie à traiter par une classification binaire. Ensuite, il est proposé une approche ad-hoc identifiant d'une part les quanta demandés et accordés à l'aide de la position de termes clés appris sans exemple grâce à des métriques de pondération des termes, et d'autre part le sens du résultat à l'aide d'un ensemble prédéfini de mots-clés particulier à la rédaction des résultats. Cette méthode, bien que dépendante d'heuristiques, parvient à reconnaître un grand nombre de demandes avec plus ou moins de difficultés selon les catégories traitées. Ensuite, la classification de documents est expérimentée comme approche plus généraliste. Sur l'ensemble des algorithmes explo-

rés, les extensions de l'analyse PLS, appliquées ici pour la première fois sur du texte, démontrent une efficacité proche de celle du meilleur algorithme testé, l'arbre de décision. L'utilité de la restriction des documents à des passages relatifs à la catégorie est aussi démontrée empiriquement. Enfin, le chapitre **??** aborde la problématique de similarité entre deux textes dans un contexte de catégorisation non supervisée des documents. Le but est ici de révéler les circonstances factuelles faisant appel à une catégorie de demande particulière. Une approche d'apprentissage de distance est proposée : elle repose sur le coût d'une transformation d'un des deux textes en l'autre. Cette distance est comparée à d'autres métriques avec l'algorithme des K-moyennes dans des expérimentations qui explorent différents aspects des problèmes de regroupement comme la détermination du nombre de clusters ou la représentation de documents. En somme, les problématiques abordées sont certes variées mais indispensables aux différents maillons de la chaîne complète de traitement automatique de corpus de décisions dont le chapitre **??** montre l'utilité pour visualiser l'état de la jurisprudence, une des nombreuses applications possibles des données extraites.

ii Critique du travail

Au delà des nombreuses problématiques abordées et expérimentations discutées, cette thèse reste limitée par son niveau de contribution théorique d'une part. La proposition globale est une chaîne de traitement employant à chaque niveau des approches soit existantes soit plus techniques. Aussi, un très grand nombre de méthodes de la littérature sont absentes, surtout les plus récentes ; ceci est dû fait à l'ampleur du travail et à la multitudes d'approches existantes. D'autre part, les études menées ont rencontrées comme obstacles la disponibilité d'exemples de référence annotées manuellement. La lenteur et la pénibilité de l'identification des informations à la main se traduit par la faible quantité des données employées pour les expérimentations. De plus, ne disposant que d'un expert, le degré d'accord entre annotateurs n'a été analysé que pour la première problématique de reconnaissance d'entités et de sections. Par conséquent, certaines subtilités propres à l'expert ou des données manquées lors de l'annotation manuelle, peuvent biaiser les résultats observés. Néanmoins, les nombreux résultats obtenus servent de base pour la continuité des études.

iii Travaux futurs de recherche

Les propositions données dans la conclusion des chapitres **??** à **??** pour continuer les travaux peuvent être résumées en 4 catégories principale. En premier, l'exploration de méthodes récentes que celles étudiées permettra d'étendre les résultats expérimentaux. Ensuite, la formalisation des problèmes abordées permettra de définir des approches plus théoriques. Par exemple, la formalisation des demandes comme des relations, entre quantum demandé et quantum accordé, permettra d'explorer le cadre probabiliste et neuronale de la littérature en matière d'extraction des relations. Puis, l'exploration d'autres formulation des problèmes permettra probablement de découvrir des méthodes plus efficaces. Par exemple, on peut percevoir la détermination des circonstances factuelles comme une tâche de modélisation de thématiques (*topic modeling*). Enfin, l'exploration plus approfondie d'autres aspects des problèmes. Par exemple, la reconnaissance d'entités nommées comprend l'identification que nous avons étudiée, et la résolution qui unifie les mentions variantes d'une entité sous un identifiant prédéfinir ou à définir automatiquement.

Il faut aussi remarquer qu'il reste encore des types d'information dont le problème d'extraction n'est pas abordé par cette thèse. Par exemple, les raisons, qui font penchés les juges en faveur d'une décision sur une demande, sont indispensables pour être capable d'anticiper la prise de décision des juges. L'extraction des raisons concernera l'identification et l'analyse des arguments des parties et les motivations des juges.

Par ailleurs, il faudra aussi mieux évaluer la qualité des annotations manuelles expertes ce qui révélera le niveau d'accord non seulement sur les données annotées mais aussi sur leur perception des informations ciblées comme les circonstance factuelles qui semblent subjectives.

iv Perspectives du domaine

D'une part, le conflit entre la qualité des données annotées manuellement et la rapidité de l'automatisation encore imprécise est important. Galgani *et al.* [2015] supportent, par exemple, qu'il est possible en un temps raisonnable d'annoter manuellement un nombre considérable de texte. Il se pose alors la question de savoir à quel point l'exhaustivité est-elle nécessaire pour contraindre les experts à supporter la marge d'erreurs

infligée par les outils d'extraction automatique.

D'autre part, cette thèse est l'un des premiers travaux de recherche de cet largeur sur les décisions françaises. Ainsi, elle ouvre la voie à bien des problématiques comme l'analyse des réseaux de normes, l'anonymisation des décisions, ou l'analyse des arguments, déjà largement étudiés dans d'autres pays, notamment aux États-Unis. En cela, cette thèse encourage la recherche en analyse de donnée textuelle à s'intéresser à l'analyse automatique de la jurisprudence française dont les défis, la disponibilité d'un grand volume de données et la lucrativité du domaine judiciaire ne rendent ce champ d'application que plus attractif. Les cas d'utilisation des données extraites sont très nombreuses dans la recherche en droit, l'aide à la décision des juristes, dans l'enseignement du droit, et surtout dans l'accessibilité des profanes au droit par une estimation automatique de leurs risques judiciaires.

Bibliographie

ACE. 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*. 5.4.3 edn. Linguistic Data Consortium.

Alfred, Rayner, Leong, Leow Chin, On, Chin Kim, & Anthony, Patricia. 2014. Malay named entity recognition based on rule-based approach. *International Journal of Machine Learning and Computing*, **4**(3), 300.

Amarappa, S., & Sathyanarayana, S. V. 2015. Kannada named entity recognition and classification (NERC) based on multinomial naïve bayes (MNB) classifier. *International Journal on Natural Language Computing (IJNLC)*, **4**(4).

Ancel, Pascal. 2003. *Les décisions d'expulsion d'occupants sans droit ni titre - connaissance empirique d'un contentieux hétérogène*. Tech. rept. [Rapport de recherche] Ministère de la Justice.

Cretin, Laurette. 2014. L'opinion des français sur la justice. *Infostat justice*, **125**(Janvier).

Forgey, Edward. 1965. Cluster analysis of multivariate data : Efficiency vs. interpretability of classification. *Biometrics*, **21**(3), 768–769.

Galgani, Filippo, Compton, Paul, & Hoffmann, Achim. 2015. Lexa : Building knowledge bases for automatic legal citation classification. *Expert systems with applications*, **42**(17-18), 6391–6407.

Grishman, Ralph, & Sundheim, Beth. 1996. Message understanding conference-6 : A brief history. *In : Coling 1996 volume 1 : The 16th international conference on computational linguistics*, vol. 1.

Hanisch, Daniel, Fundel, Katrin, *et al.* . 2005. ProMiner : rule-based protein and gene entity recognition. *BMC bioinformatics*, **6**(1), 14.

Jeandidier, Bruno, & Ray, Jean-Claude. 2006. Pensions alimentaires pour enfants lors du divorce - [Les juges appliquent-ils implicitement un calcul fondé sur le coût de l'enfant ?]. *Revue des politiques sociales et familiales*, **84**(1), 5–18.

Kaufman, Leonard, & Rousseeuw, Peter J. 1987. Clustering by means of medoids. *Page 405–416 of :* Dodge, Yadolah (ed), *Statistical data analysis based on the l1-norm*. North Holland/Elsevier. Amsterdam.

Kim, Jin-Dong, Ohta, Tomoko, Tsuruoka, Yoshimasa, Tateisi, Yuka, & Collier, Nigel. 2004. Introduction to the bio-entity recognition task at JNLPBA. *Pages 70–75 of : Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*. Association for Computational Linguistics.

Lamanda, Vincent. 2010. *Discours du premier président de la cour de cassation vincent lamanda lors de l'audience solennelle de début d'année 2010*. `https://www.courdecassation.fr/institution_1/occasion_audiences_59/debut_annee_60/discours_m._lamanda_14858.html`.

Langlais, Eric, & Chappe, Nathalie. 2009. *Analyses économiques du droit : principes, méthodes, résultats*. Editions de Boeck Université. Chap. 4. Analyse économique de la résolution des litiges.

Li, Jianqiang, Zhao, Shenhe, Yang, Jijiang, Huang, Zhisheng, Liu, Bo, Chen, Shi, Pan, Hui, & Wang, Qing. 2018. Wcp-rnn : a novel rnn-based approach for bio-ner in chinese emrs. *The journal of supercomputing*, 1–18.

Nazarenko, Adeline, & Wyner, Adam. 2017. Legal NLP Introduction. *Traitement automatique de la langue juridique / Legal Natural Language Processing - Revue TAL*, **58**(2), 7–19.

Persson, Caroline. 2012. *Machine learning for tagging of biomedical literature*. Closing project report, Technical University of Denmark,DTU Informatics.

Tjong Kim Sang, Erik F., & De Meulder, Fien. 2003. Introduction to the conll-2003 shared task : Language-independent named entity recognition. *Pages 142–147 of : Proceedings of the seventh conference on natural language learning at hlt-naacl 2003 - volume 4*. CONLL '03. Stroudsburg, PA, USA : Association for Computational Linguistics.

Waltl, Bernhard, Matthes, Florian, Waltl, Tobias, & Grass, Thomas. 2016. LEXIA - A Data Science Environment for Semantic Analysis of German Legal Texts. *In : Iris : Internationales rechtsinformatik symposium*. Salzburg, Austria.

Yadav, Vikas, & Bethard, Steven. 2018. A survey on recent advances in named entity recognition from deep learning models. *Pages 2145–2158 of : Proceedings of the 27th international conference on computational linguistics*.