Introduction Générale

i Contexte et Motivations

Une décision jurisprudentielle peut être définie soit comme le résultat rendu par les juges à l'issue d'un procès, ou bien un document décrivant une affaire judiciaire. Un tel document rapporte, en fait, les faits, les procédures judiciaires antérieures, la solution des juges, et les raisons qui les y ont conduits. Dans ce mémoire, nous désignons par « décision » le document, et par « résultat » la conclusion, ou réponse, ou solution finale des juges. Une jurisprudence [1] est un ensemble de décisions rendues par les tribunaux et qui représente la manière dont ces derniers interprètent les lois pour résoudre un problème juridique donné (type de contentieux). Les juristes doivent collecter ces documents, en sélectionner et analyser pour leurs études. En effet, leur analyse aide à mieux comprendre la prise de décision des juridictions pour mener, par exemple, des recherches empiriques en droit (Ancel, 2003; Jeandidier and 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 dans un contentieux. Cette activité de collecte et d'analyse est manuelle en général, et par conséquent, sujette à plusieurs difficultés liées à l'accès et à l'exhaustivité des documents.

Les documents sont dispersés entre les tribunaux. Les procédures administratives ne facilitent pas toujours leur accès du fait de la nécessité de préserver la confidentialité des parties. Les décisions n'étant pas anonymisées ne peuvent être rendues aux juristes qui en font la demande. Un certain nombre de documents sont néanmoins accessibles sur internet grâce

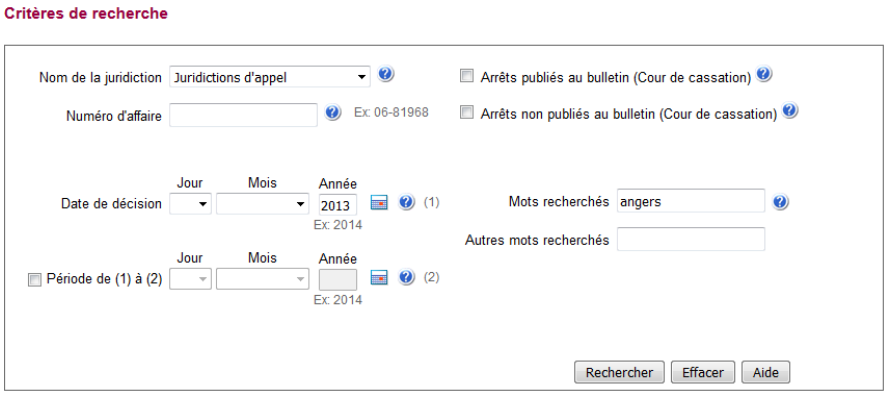
1. `http://www.toupie.org/Dictionnaire/Jurisprudence.htm`

à des sites de publication de données ouvertes gouvernementales, comme `http://data.gouv.fr` en France, `https://www.judiciary.uk` en Grande Bretagne, `http://www.scotusblog.com/` aux Etats-Unis, et `https://www.scc-csc.ca/` au Canada. Ces derniers publient régulièrement des décisions récemment prononcées. Il existe aussi des moteurs de recherche juridique qui permettent d'effectuer de retrouver des décisions intéressantes. Cependant, qu'ils soient payants (LexisNexis [2], Dalloz [3], Lamyline [4],...) ou gratuits (CanLII [5], Légifrance [6], ...), leurs critères de recherche limitent la pertinence des résultats. En effet, il ne s'agit en général que de combinaisons de mots-clés et autres métadonnées (date, type de juridiction, ...), ou d'expressions régulières, comme l'illustre la Figure 1. Ces critères manquent la sémantique juridique qui ramènerait, aux juristes, des échantillons plus pertinents.

L'exhaustivité de l'analyse, ou tout au moins sa représentativité, rencontre un frein face à l'énorme volume existant de documents. En effet, plus de 4 millions de décisions sont prononcées en France par an d'après les chiffres du ministère français de la justice (Tableau 1). Au regard de la croissance rapide de la quantité de décisions, on imagine facilement que même pour une étude sur une question très précise, le corpus utile reste large. Par ailleurs, il peut s'avérer très pénible de lire les décisions pour en identifier les données d'intérêt. Les documents sont très souvent longs. Ils sont aussi complexes dans leur style de rédaction. Par exemple, les phrases sont longues et comprennent plusieurs clauses discutant parfois d'aspects différents. On y retrouve aussi des références à des jugements antérieurs, et des omissions, ...

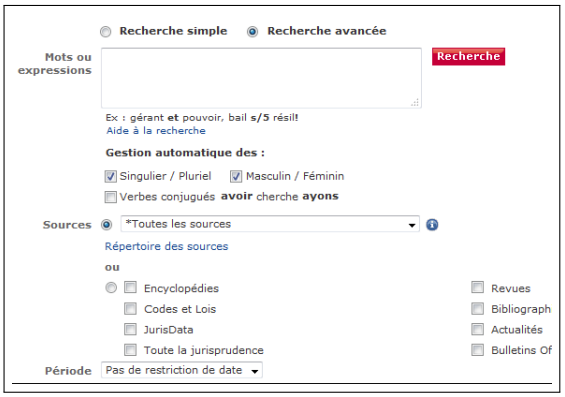
Il est évident qu'une automatisation du traitement des corpus de décisions s'impose pour répondre aux diverses difficultés d'accès, de volumétrie, et de complexité liées à la compréhension des décisions. L'automatisation ferait gagner du temps aux juristes sur des tâches de traitement préalables à leur raisonnement d'experts, tout en leur fournissant une vue pertinente de la jurisprudence. D'autre part, Cretin (2014) fait remarquer que la justice est

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`



(a) Formulaire de Légifrance

(b) Formulaire de Dalloz



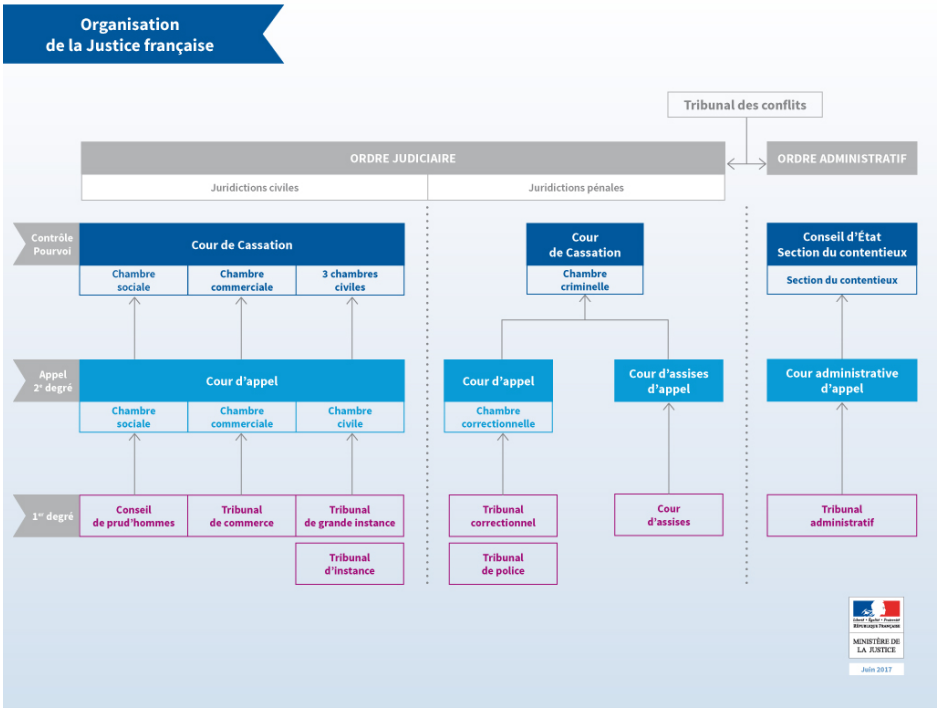
(c) Formulaire de LexisNexis

Figure 1 – Exemples de critères des moteurs de recherche juridique

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



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

Figure 2 – Organisation des institutions judiciaires françaises

complexe dans son organisation (Figure 2) et son fonctionnement, et son langage est pratiquement incompréhensible. Il est donc presque impossible pour les profanes d'estimer leurs droits et le risque judiciaire qu'ils encourent dans leur quotidien sans consulter un initié en droit. L'automatisation pourrait aussi améliorer l'accessibilité du droit dans ce cas. L'exigence pour le profane étant l'exacte pertinence des ressources, leur accessibilité, et l'intuitivité du processus de leur exploitation (Nazarenko and Wyner, 2017). Le traitement automatique constitue, en résumé, une aide précieuse non seulement pour les professionnels du droit, mais aussi pour les particuliers et entreprises soucieux de voir l'issue de leur affaire leur être favorable. 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 and Chappe, 2009).

ii Objectifs

Ce mémoire discute des résultats d'une étude visant à automatiser l'extraction d'information à partir des décisions françaises. Le but est de faciliter la recherche, l'analyse descriptive, et l'analyse prédictive sur une masse de documents. L'approche traditionnelle d'analyse d'un contentieux (Ancel, 2003) consiste à :

1. **Choisir un échantillon représentatif** : collection des décisions suivant des contraintes définies : période précise et d'une couverture géographique, types d'affaires, ...
2. **Sélectionner les décisions** : élimination des décisions qui ne correspondent pas au type de demande d'intérêt.
3. **Elaborer la grille d'analyse** : création d'un modèle de grille (tableau) qui permettra d'enregistrer les informations potentiellement importantes. Chaque ligne correspond à une demande et les colonnes sont les différents types d'informations qu'on peut 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 ; c'est au cours de la lecture des décisions qu'on retrouve les informations qui paraissent intéressantes.
4. **L'analyse des décisions et l'interprétation des informations** : saisie des décisions et calculs statistiques dans un logiciel tableur.

Ancel (2003) évoque principalement le problème de la différence entre l'état capté de la jurisprudence et son état présent. D'une part en effet, Les longs délais de travail sont caractéristiques de ces études. Nous avons pour exemple, l'étude menée par l'équipe de Jeandidier and Ray (2006) pour l'analyse empirique des déterminants de la fixation de pensions alimentaires pour enfant lors de divorce. Cette analyse a duré 9 mois pour l'extraction manuelle des informations et la modélisation par régression de la relation entre les déterminant extraits et les pensions alimentaires accordées. D'autre part, il est impossible d'observer l'évolution des pratiques judiciaires dans le temps et dans l'espace du fait de la faible taille de l'échantillon choisie.

Notre principal objectif est donc de proposer des solutions pour un traitement rapide et efficace d'une grande masse de décisions.

La question à la base de notre étude est celle de savoir « comment capter automatiquement la sémantique d'un corpus jurisprudentiel pour comprendre la prise de décision des juges sachant que l'interprétation subjective des règles juridiques rend l'application de la loi non déterministe ? ». Cette question intéresse des entreprises telles que LexisNexis, et plusieurs startups à l'exemple de Predictice [7] et CASE LAW ANALYTICS [8]. Afin d'y répondre, nous nous intéressons aux concepts manipulés par les experts, au centre desquels on retrouve la demande ou prétention des parties. Tout autour de la demande, gravitent d'autres concepts importants qui enrichissent la compréhension de la décision (Figure 3) :

— le résultat associé qui est décrit par une polarité (« accepte » ou « rejette »), souvent un quantum accordé (par ex. 5000 euros, 2 mois d'emprisonnement) ;
— le fondement ou la norme juridique qui est la règle qui détermine et légitime la prétention ou le résultat ;
— l'objet qui représente ce qui a été demandé (par ex. dommages et intérêts) ;
— les circonstances factuelles dans lesquelles sont formulées les demandes ; ils décrivent les types de faits caractérisant ainsi les types de contentieux ou d'affaires ;
— les raisons c'est-à-dire les divers arguments apportés par les parties (resp. les juges) pour justifier leurs requêtes (resp. leurs solutions) ;

En fait, cette abstraction couvre l'essentiel de l'information pertinente pour les experts. L'analyse sémantique vise donc à identifier les connaissances sur les nombreuses demandes présentent dans les décisions.

Les travaux de cette thèse s'inscrivent dans un projet qui vise, entre autres, l'automatisation de l'analyse empirique des contentieux pour observer de manière exhaustive et synthétique les pratiques judiciaires. L'objectif final est de concevoir un système capable de fournir une estimation des

7. http://predictice.com
8. http://caselawanalytics.com

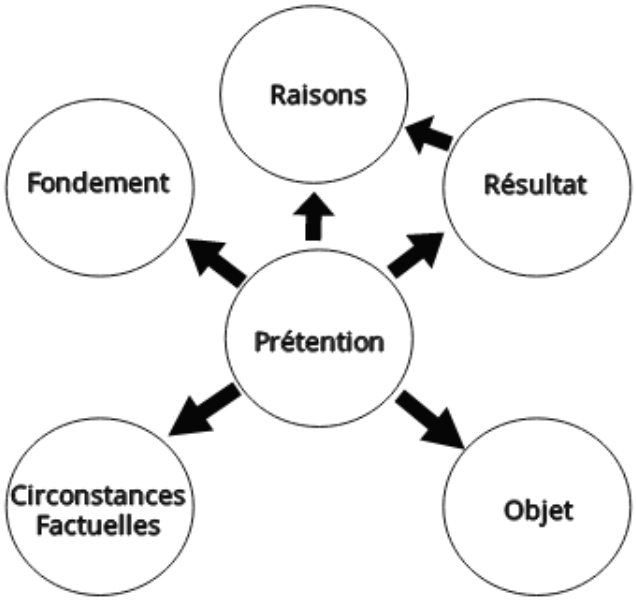


Figure 3 – La demande au centre de la compréhension des décisions

chances d'obtenir un résultat positif suivant des critères comme la juridiction, le type de demande, ou les circonstances du litiges, et d'identifier les facteurs influençant le résultat. Le projet comprend deux phases principales : une phase d'indexation des connaissances de la masse des décisions, suivie d'une phase d'analyse prédictive. La phase d'indexation doit déjà permettre de réaliser automatiquement, de manière exhaustives, des analyses descriptives. Ces dernières consistent, par exemple, à comparer le nombre d'acceptations à la fréquence des rejets. Par conséquent, le système doit apprendre à reconnaître dans les décisions, les informations pertinentes sur les prétention et résultats associés. Pour la phase d'analyse prédictive, le principe consiste à regrouper des paquets de décisions similaires (même résultat sur la même prétention dans les circonstances similaires), pour découvrir les facteurs influençant le sens du résultat (par ex. le fait que « le revenu de l'époux soit le plus élevé du foyer » encourage les juges à accorder la pension alimentaire à l'épouse). En effet, c'est la connaissance de ces facteurs qui permet à l'expert de pouvoir anticiper les décisions judiciaires.

La chaîne de traitement à mettre en œuvre consiste en quatre étapes principales qui s'enchaînent comme le présente la figure 4. Ce document n'aborde

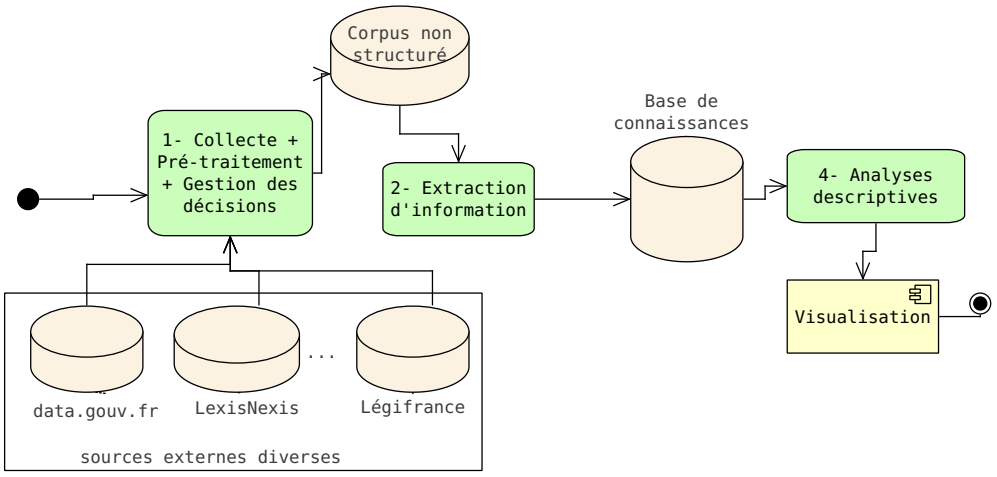


Figure 4 – Chaine d'analyse du corpus jurisprudentiel à mettre en œuvre

pas l'aspect prédictif du projet. Notre étude s'est limitée aux problématiques liées à l'analyse descriptive, et qui sont décrites dans la suite.

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

Le volume de décisions prononcées est énorme et croit très rapidement (Tableau 1). Il est donc nécessaire de trouver des moyens pour collecter le maximum de documents bruts non-structurés, les pré-traiter, et aussi organiser leur gestion afin de les indexer en local pour faciliter leur traitement.

Les décisions de cours d'appel de justice civile sont les plus accessibles à partir des moteurs de recherche juridique (LexisNexis, Dalloz, LamyLine, Legifrance, ...) et de la grande base de données JuriCa de la Cour de cassation. 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). En plus, le nombre de téléchargements par jour est limité. 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 études et du rapport [9] (SDER). Cet accès est payant pour les professionnels et gratuit

9. https://www.courdecassation.fr/institution_1/composition_56/etudes_

pour les universités et centres de recherche en partenariat avec le SDER. 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 à l'aide d'une requête GET du protocole HTTP. Ainsi, il est possible de programmer un client web capable de télécharger l'ensemble des documents de Légifrance. Ce dernier a cet avantage de proposer des décisions de tous les ordres et de tous les degrés. Cependant, les décisions des juridictions du premier (jugements) restent plus rares sur internet et pincipalement disponibles auprès des tribunaux. La disponibilité des décisions du second degré ou d'appel (arrêts) en justice civile est l'une des raisons pour lesquelles notre étude s'est portée sur ceux-ci.

Les décisions existent sous divers formats PDF, DOC, DOCX, RTF, TXT, XML,... Il arrive parfois qu'un fichier téléchargé comprenne plusieurs décisions (sur LexisNexis par ex.). Nous avons par conséquent préférer convertir tous les documents au format plein texte pour homogénéiser les traitements. Par ailleurs, les décisions sont collectées à partir de diverses sources pouvant contenir des documents identiques. Il se pose donc un problème d'identification unique des décisions pour éviter des redondances. Pour cela, nous avons défini un schéma de nomination unique des fichiers. Ce dernier repose sur 3 informations : le type de juridiction (tribunal, cour d'appel, ...), la ville, et le numéro R.G. (registre général) qui est l'identifiant unique de la décision au sein de la juridiction. Par exemple, le numéro « CAREN1606137 » identifie la décision de numéro R.G. « 16/06137 » de la cour d'appel (« CA ») de la ville de Rennes (« REN »). Ces 3 informations sont présentes dans les premières lignes de la décision, et sont facilement identifiables à l'aide d'une routine à base de règles simples. D'autre part, certains moteurs de recherche ne fournissent souvent qu'un résumé au lieu du contenu original des décisions. Il est important de supprimer ces fichiers du corpus.

`rapport_28.html`

ii.b Extraction de connaissances

Les problématiques d'extraction de connaissances constitue l'essentiel de ce mémoire car elles sont les plus importantes et les plus difficiles. La difficulté découle de l'état non-structuré des documents et de la complexité du langage employé. L'extraction des connaissances nécessite de mettre en œuvre des techniques, de fouille de texte, 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 métadonnées de références sont des segments de texte qu'on peut directement localiser dans le document. Leur reconnaissance est donc semblable à celle des entités nommées. C'est une problématique intensivement étudiée en traitement automatique du langage naturel (Yadav and Bethard, 2018) dans plusieurs travaux et compétitions, aussi bien pour des entités communes (Tjong Kim Sang and De Meulder, 2003; Grishman and 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 and Sathyanarayana, 2015).

Le problème d'extraction des demandes et de la réponse correspondante des juges consiste à reconnaître pour chaque prétention : son objet, son fondement, le quantum demandé, le sens du résultat, et le quantum accordé. La paire demande-résultat s'apparente donc à des entités structurées comme 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 des 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 texte, de regroupement non supervisé (*clustering*), et de « modélisation thématique » (*topic modeling*).

Le projet s'intéresse aussi à l'identification des raisons justifiant le résultat des juges sur une demande. Mais ce problème n'a pas encore été abordé.

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

Comme illustrée précédemment (§ ii.b), les problématiques propres aux textes juridiques trouvent généralement des analogies avec les problèmes d'analyse de données textuelles. Ainsi, les méthodes issues de l'énorme progrès réalisé dans ce domaine sont applicables au 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 texte 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 française, nous avons axé notre travail sur le rapprochement des problèmes liés à l'analyse des décisions jurisprudentielles à celles

qui sont généralement traitées en analyse de données textuelles. Il s'agissait ensuite d'établir des protocoles d'évaluation et d'annotation manuelle de données. Selon les problématiques identifiées et les protocoles d'évaluations définies, des méthodes adaptées ont été proposées et expérimentés sur les données réelles annotées par des experts.

iv Résultats

Une chaîne traitement pour le sectionnement et l'annotation des métadonnées a été proposée. 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*), a été étudiée en considérant plusieurs aspects de la conception des systèmes d'extraction d'entités nommées. Le sectionnement a pour but d'organiser l'extraction des informations qui sont réparties dans des sections selon leur nature.

Par la suite, nous avons proposé une méthode d'extraction des demandes et résultats en fonction des catégories présentes dans la décision. L'approche consiste en effet à identifier dans un premier temps les catégories présentes (objet+fondement) par classification supervisée. Ensuite, un vocabulaire d'expression des demandes et résultats est exploiter 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 avons aussi étudié l'extraction particulière du sens du résultat par classification binaire des documents. L'objectif était de s'affranchir de l'identification préalable de l'expression des demandes et résultats.

L'identification des circonstances factuelles, quant à elle, a été modélisée comme une tâche de regroupement non supervisée des décisions. Nous avons proposé dans ce cas une méthode d'apprentissage d'une métrique de dissimilarité sémantique entre texte, à l'aide d'un modèle de régression. La métrique apprise a été comparée à d'autres distances établies en recherche d'information.

v Structure du mémoire

La suite du mémoire est organisée en 6 chapitres. Le chapitre 1 positionne nos travaux par rapport à ceux qui ont été réalisés précédemment sur des problématiques proches. Le chapitre 2 présente la l'architecture de structuration et reconnaissance des entités juridiques, et discute des différents résultats empiriques obtenus par application des modèles CRF et HMM. Ensuite, le chapitre 3 détaille le problème d'extraction des paires demande-résultat, puis présente notre méthode et les résultats obtenus. Le chapitre 4 discute de l'extraction particulière du sens du résultat par classification directe des documents en comparant différents algorithmes et méthodes de représentation des textes. Le chapitre 5 présente notre approche d'apprentissage de la métrique de dissimilarité textuelle, et la compare à des distances établies en recherche d'information sur le problème d'identification des circonstances factuelles. Enfin, le chapitre 6 présente les résultats de scénarios d'analyses descriptives pour illustrer l'exploitation potentielle de nos propositions sur des corpus de grande taille.