Proposition de projet doctoral

RÉSUMÉ AUTOMATIQUE DE TEXTES JURIDIQUES

par
Atefeh Farzindar
farzinda@iro.umontreal.ca

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse en cotutelle
Université de Montréal, Équipe RALI
Université ParisIV-Sorbonne, Équipe LaLicc

Sous la direction de
Guy Lapalme, Professeur à l'Université de Montréal et
Jean-Pierre Desclés, Professeur à l'Université Paris-Sorbonne

Septembre 2003

Table des matières

Résumé

Nous proposons une application particulière du résumé et un outil permettant aux juristes de consulter rapidement les idées clés d'une décision juridique pour trouver les jurisprudences pertinentes à leurs besoins. Notre approche repose sur des techniques d'extraction des unités saillantes des décisions judiciaires des cours canadiennes, pour former un résumé indicatif afin d'exploiter les contenus essentiels du texte source. Ce travail est effectué dans le cadre d'une thèse en cotutelle entre l'Université de Paris 4-Sorbonne et l'Université de Montréal en collaboration avec le CRDP (Centre de recherche en droit public) de l'Université de Montréal.

Chapitre 1

Introduction

Dans ce document nous proposons la démarche de notre recherche pour aborder le problème de la création automatique de résumé de texte, ou plus précisément des résumés de textes juridiques de type de décision et jugement délivré par les cours juridiques du Canada.

Ce document présentera d'abord les notions de base des documents de type juridique et dégagera les travaux pertinents pour la génération de résumé automatique dans ce contexte.

De nombreux travaux ont proposé des méthodes pour créer automatiquement des résumés et, plus récemment, des compétitions [DUC, 2003] [TREC, 2002] ont été organisées pour les évaluer. Nos recherches viseront à enrichir ces approches en se basant sur nos expériences de participation aux compétitions *Document Understanding Conference* 2002 et 2003 (DUC est un workshop en résumé automatique associé avec la conférence ACL). Nous proposerons de nouvelles méthodes adaptées aux besoins du traitement de documents juridiques.

1.1 Condensation automatique

Dans le monde actuel, des millions de documents électroniques sont disponibles via les réseaux et les supports informatiques mais il devient de plus en plus difficile d'accéder aux informations intéressantes sans l'aide d'outils spécifiques, dont la technologie du résumé automatique qui permet de choisir un ou plusieurs documents utilement.

Le but principal d'un résumé est de fournir au lecteur une idée précise et complète du contenu d'une source d'information et de le présenter à l'utilisateur sous une forme condensée en langage naturel, il faut aussi considérer les besoins de l'utilisateur et de la tâche spécifiée. Les systèmes de production de résumé peuvent aussi être utilisés comme aide à l'organisation des connaissances ou une aide à l'analyse et à la recherche d'informations sur le réseau internet. [Mani, 2001], [Mani et Maybury, 1999], [Marcu, 2000] et [Minel, 2002] présentent de bonnes introductions à la situation actuelle dans ce domaine.

Un exemple de domaine d'application de résumé est celui de textes juridiques. Chaque jour, les avocats accèdent aux bases de données électroniques qui contiennent des dizaines de millions de documents. Le grand volume d'information juridique et l'effort énorme requis manuellement pour résumer et classer les décisions judiciaires, justifie un système qui automatise le processus de production de résumé de textes légaux.

Dans ce projet de thèse, nous proposons une application particulière du résumé qui peut servir dans le domaine juridique. Nous développerons un outil permettant aux juristes de consulter rapidement les idées clés d'une décision juridique pour trouver les jurisprudences pertinentes à leurs besoins. Un tel système cherche les unités textuelles importantes dans le texte afin d'exploiter les contenus essentiels du texte source.

1.2 Proposition de thèse

Dans le cadre de notre thèse en cotutelle entre l'Université de Paris 4-Sorbonne (laboratoire LaLicc ; Langages, Logiques, Informatique, Cognition et Communication) et l'Université de Montréal (laboratoire RALI ; Recherche Appliquée en Linguistique Informatique), nous collaborons avec le CRDP (Centre de recherche en droit public), pour développer un système de constitution automatique de résumés à partir de jugements en anglais de cours judiciaires du Canada. La sortie du système sera un texte court d'environ 150 mots soit un taux de compression d'environ 10% (rapport entre la longueur de résumé et la longueur de texte source). Son objectif est de présenter les informations importantes du texte original avec les renvois importants (i.e. le numéro de l'article de loi appliqué pour le cas, etc...).

Le but de notre thèse est de développer une approche de résumé automatique adaptée aux textes juridiques. Dans ce contexte, nous nous concentrerons sur les aspects suivants :

- Les particularités des documents dans le domaine juridique. En particulier, les textes de décisions des cours judiciaires.
- Étude des notions de bases de droit et le domaine juridique pour une meilleure compréhension du langage de droit et de la structure des textes de cette application particulière.
- Les approches de résumé automatique générique pour les mono-documents et multi-documents ainsi que des travaux récents pour le traitement des textes juridiques.

Nous pourrons ainsi mieux cibler notre recherche dans ce nouveau domaine d'application du résumé automatique de texte.

1.3 Description du rapport

Dans le deuxième chapitre, nous introduisons d'abord la terminologie nécessaire, l'état de l'art sur la condensation automatique de texte et la problématique ainsi que des différentes

méthodes de création automatique de résumé et évaluation des systèmes. Le troisième chapitre est consacré aux caractéristiques du domaine et du corpus ainsi que les travaux pertinents sur le résumé de texte juridique. Nous décrirons différentes méthodes de production de résumé juridique utilisées actuellement par des experts et les approches automatiques d'un point de vue de leur adéquation par rapport aux textes légaux. Le quatrième chapitre décrira notre proposition de thèse en précisant notre démarche pour l'élaboration d'un système de résumé automatique.

Chapitre 2

Résumé automatique

Le but d'un système de résumé automatique est de produire une représentation condensée du contenu de son entrée, où les informations importantes du texte original sont préservées. Le *texte* ici peut désigner différents genres : document de multimédia, e-mail, hypertexte, etc., utilisé dans divers domaines : scientifique, littérature, juridique, etc. Le résumé produit par le système peut se présenter sous différentes formes et types. Dans ce travail, nos études porteront sur une forme particulière des documents de type juridique, les décisions des cours judiciaires du Canada et les différents aspects de ce domaine d'application. L'objectif de ce projet est de développer un système de création automatique de courts résumés, qui répond aux besoins des avocats, des juges et des experts du domaine.

Dans ce chapitre, nous présentons un aperçu des systèmes de résumé de texte automatique, les définitions de base, les approches de résumés et l'évaluation des systèmes de condensation automatique. Nous commencerons par différentes typologies proposées pour le résumé.

2.1 Types de résumés

Des nombreux types de résumé ont été identifiés [Borko et Bernier, 1975] [Cremmins, 1996] [Spark Jones, 1999] [Hovy et Lin, 1999] selon leur longueur, leur style et leur subjectivité :

résumé indicatif qui fournit une idée du texte sans donner le contenu spécifique, il signale les thèmes du document et son style peut être télégraphique.

résumé informatif qui renseigne sur les informations essentielles contenues dans le document. Les informations sont présentées dans l'ordre du document, mais leur importance relative peut différer de celle du document.

résumé sélectif qui néglige les aspects très généraux d'un document et développe les parties spécialisées.

résumé ciblé qui se concentre sur le(s) topique(s) d'intérêt pour l'utilisateur, alors que le résumé générique reflète plutôt le point de vue d'auteur.

sommaire qui énumère des points principaux d'un discours.

digest ou fiche analytique qui condense un livre ou un article de journal.

Il a été proposé par ailleurs [DUC, 2003] une classification des résumés s'appuyant sur les tâches :

- résumé très court (∼10 mots) (i.e. un titre informatif pour les articles des journaux)
- résumé focalisé par des événements,
- résumé du point de vue indiqué,
- résumé en réponse à une question.

La production du résumé comporte habituellement les processus suivants :

extraction : identification des éléments importants dans le texte,

génération : reformulation du matériel extrait dans un nouveau texte cohérent,

fusion : combinaison des parties extraites,

compression : élimination des éléments de moindre importance.

2.2 Méthodes de résumé automatique

L'idée de produire des résumés automatiques n'est pas nouvelle [Lunh, 1959]. Cependant, les techniques utilisées pendant les années 50 et 60 ont été caractérisées par la simplicité de leur traitement. Jusqu'aux années 70s, les méthodes utilisées étaient plutôt statistiques, les premiers travaux [Lunh, 1959] [Edmundson, 1969] ont étudié les techniques suivantes :

Position dans le texte où les phrases pondérées selon leur position (premier paragraphe, ou *Introduction*, *Conclusion* ou la partie après les titres de section, etc...).

Sélections lexicales où la présence de marqueurs linguistiques est identifiée.

Lieu où les premières (et dernières) phrases de chaque paragraphe devraient contenir l'information sur le sujet.

Fréquence où les mots significatifs sont fréquents dans le document alors qu'ils ont une fréquence faible en général.

Bien que chacune de ces approches ait une certaine utilité, elles dépendent du format et du modèle d'écriture. La stratégie qui consiste à prendre le premier paragraphe n'est pas universelle, elle ne fonctionne que pour les articles de journaux et de magazines. Dans d'autres cas comme pour la production d'un résumé biographique ou d'une décision juridique, la position des phrases n'est toujours un élément prépondérant. Pour élaborer des systèmes de résumé automatique, la plupart des chercheurs se sont penchés sur deux types d'approches : par extraction ou par génération. Les extraits sont créés par réutilisation directe des parties (mots, phrases, etc...) du texte d'entrée, alors que des résumés générés sont créés par régénération à partir d'une représentation abstraite du contenu extrait.

2.2.1 Création automatique de résumé par extraction

Pour créer un résumé par extraction, il faut d'abord repérer les unités textuelles (expressions, propositions, phrases, paragraphes) considérées comme saillantes ; ensuite le système va sélectionner les unités textuelles qui portent des idées principales du contenu du document. Le résumé produit par extraction respecte l'ordre d'apparition des phrases dans le document original.

Les approches d'extraction récentes emploient des techniques plus sophistiquées pour extraire les phrases ; ces techniques se fondent souvent sur l'apprentissage machine pour identifier les indicateurs importants, sur l'analyse de langage naturel pour identifier les passages pertinents ou sur des relations entre les mots.

Nous allons présenter quelques algorithmes de sélection d'unités textuelles pertinentes.

Position des phrases

À cause de la structure des textes, certaines positions dans le texte risquent de contenir les informations importantes. Une méthode simple, surtout pour les articles des journaux, est de prendre le premier paragraphe comme résumé. [Lin et Hovy, 1997] montrent que les phrases importantes sont localisées dans des positions qui dépendent du genre du document (article scientifique, article de journal, ...). Pour créer un résumé, on accorde plus de poids aux phrases qui se trouvent dans les premiers et les derniers paragraphes du document, les premières et les dernières phrases des paragraphes et certaines sections comme *Introduction, Conclusion.*

Expressions prototypiques

Le système de [Edmundson, 1969] utilise des expressions courantes (*cet article propose ... , en conclusion, ...*) qui signalent l'importance de phrases qui les contiennent. Pour extraire ces phrases, il pondère ces expressions par leur localisation, leur fréquence relative dans le corpus, etc. Les *cue phrases* sont regroupées en classes positives et négatives. Dans le cas d'article scientifique, [Teufel et Moens, 1997] rapportent un taux de 54% (rappel et précision) avec l'utilisation d'une liste de 1 423 *cue phrases* identifiées manuellement.

Fréquence des mots

Lunh [Lunh, 1959] utilise la loi de Zipf pour la distribution des mots (quelques mots se produisent très souvent, peu de mots se produisent quelques fois, et beaucoup de mots se produisent rarement) pour développer ce critère : si un texte contient quelques mots avec peu de fréquence, alors les phrases contenant ces mots sont probablement importants. Donc il calcule $tf * idf$, pour chaque mot du texte à résumer, en considérant l'exclusion des mots vides. La technique de pondération utilisée pour calculer du poids d'une phrase est :

$score(M) = flocal * \log(100 * N/fglobal)$

$flocal$: fréquence du mot dans le texte ;

$fglobal$: fréquence du mot dans le corpus de référence ;

N : nombre de documents du corpus ;

Par la suite un score est attribué à chaque unité textuelle UT par addition des scores de chacun des mots contenus dans celle-ci :

$score(UT) = \sum score(M)$

Mots des titres

Edmundson [Edmundson, 1969] utilise cette hypothèse : les principaux thèmes sont véhiculés dans les titres alors les phrases importantes sont celles qui contiennent les mots des titres et les scores ajoutés aux phrases sont fonction de l'occurrence de mots dans les titres et de leur regroupement.

Connection par les chaînes lexicales

Cette approche exploite les techniques de TALN, ainsi que les relations sémantiques entre les différents mots en se basant sur la notion de concept [Hahn, 1990]. La méthode de [Barzilay et Elhadad, 1997] identifie des *chaînes lexicales* entre les mots et il sélectionne les phrases par chaînes lexicales. Les mots peuvent être connectés avec différents types de relations sémantiques : synonymes, hyperonymes, coréférence, ou similarité grammaticale ou lexicale dans les thesaurus. Des chaînes lexicales peuvent être établies entre les phrases du document, entre le titre et le document ou la requête et le document. Alors les chaînes peuvent être pondérées et les phrases sont jugées significatives lorsqu'elles sont traversées par un grand nombre de chaînes [Silber et McCoy, 2002].

Filtrage sémantique

La méthode d'*Exploration Contextuelle* [Desclés, 1988][Desclés, 1997][Minel *et al.*, 2001] vise à identifier les connaissances linguistiques dans le texte en les restituant dans leurs contextes et en les organisant en tâches spécialisées. L'approche développée est fondée sur la construction manuelle d'une base de données de marqueurs linguistiques et une expression de règles d'exploration contextuelle. Ces règles appliquées aux phrases du texte source vont filtrer les informations sémantiques indépendantes du domaine avec les étiquettes sémantiques hiérarchisées comme : énoncés structurants, définition, causalité, etc. La stratégie de sélection

des unités saillantes est fonction des besoins des utilisateurs (filtrage d'informations).

Rhétorique du discours

Cette approche exploite les théories de discours pour trouver les relations entre les propositions des phrases (noyau et nucléide) [Marcu, 1997]. Cette méthode identifie des unités textuelles puis les relations rhétoriques qui les lient. Elle construit des arbres rhétoriques et choisit le meilleur arbre. La pondération est basée sur la forme et contenu de l'arbre de discours, à la dernière étape, il sélectionne des unités saillantes.

Approches mixtes : Combiner les méthodes linguistiques et statistiques

Diverses méthodes ont utilisé une combinaison des différentes fonctions. Dans le système de [Teufel, 1998] cinq traits sont retenus : la présence des motifs lexicaux particuliers, la position de la phrase dans le texte, la longueur de la phrase, la présence de mots thématiques, la présence de mots dans un titre. Il construit des classes rhétoriques de phrases en augmentant les scores en fonction des traits présents dans les phrases. Les phrases ayant les meilleurs scores forment le résumé.

Dans le SUMMARIST [Hovy et Lin, 1999] comparent dix huit caractéristiques différentes et leur combinaison en utilisant l'approche apprentissage par machine C4.5 [Quinlan, 1989].

2.2.2 Génération automatique de résumé

Jing et McKeown [Jing et McKeown, 1999] montrent que des résumés humains sont souvent construits à partir du document source par un processus de *coupage et collage* (cutting and pasting) de fragments de document qui sont alors combinés et regénérés en tant que

des phrases de résumé. Un système peut être développé par génération de résumé, par le processus suivant : réduction du texte en supprimant les fragments sans importance, fusion et génération de l'information pour combiner les fragments extraits, c'est-à-dire une transformation syntaxique et réordonnancement des segments du texte. [Jing, 2002] décrit la construction d'un corpus représentant le procédé de *cut-and-paste* utilisé par des humains ; un tel corpus peut être utilisé pour l'entraînement d'un système de résumé automatique.

Certaines approches de génération de résumé utilisent les templates. Le principe de ces approches est de prédéfinir des templates dont les slots spécifient les informations importantes, de remplir le template à partir d'informations extraites à partir du document source et générer le contenu du template comme texte du résumé.

Saggion et Lapalme [Saggion et Lapalme, 2002] ont développé le système SumUM qui produit de courts résumés automatiques de longs documents scientifiques et techniques par la technique de génération. Pour établir ce système, des résumés écrits par des professionnels ont été manuellement alignés avec les documents originaux pour identifier le type d'information à extraire. Étant donné la forme structurée des articles scientifiques, la majeure partie de l'information dans les résumés a été trouvée dans la première section de l'article (introduction) (40%) dans les titres ou des légendes (23%). À partir de ces observations, les auteurs ont développé une approche, appelée Analyse Sélective, qui produit des résumés indicatifs et informatifs. SumUM produit le résumé en deux étapes : l'utilisateur reçoit d'abord un résumé indicatif où ce dernier peut identifier les topiques du document qui l'intéressent ; le système génère ensuite un résumé informatif qui élabore les topiques choisis par l'utilisateur. L'implémentation est basée sur une analyse syntaxique et sémantique, l'identification conceptuelle et la régénération des textes. L'exécution de cette méthode se fonde sur la sélection des types particuliers d'information du texte source ; l'instanciation de différents types de templates ; la sélection des templates afin de produire un résumé indicatif ; la régénération d'un nouveau texte court qui indique les topiques du document et l'expansion du texte indicatif avec l'élaboration des topiques.

2.2.3 Résumé automatique de multi-documents

La création de résumés mono document, pour certains domaines comme les articles des journaux qui traitent de grandes collections d'informations, n'est pas suffisante.

C'est pourquoi certaines approches sont proposées pour la production de résumé de documents multiples. Le but est d'identifier ce qui est commun et ce qui diffère dans une variété des documents reliés et d'enlever les informations répétitives du résumé [Mani, 2001].

Lors de DUC 2003 [Farzindar et Lapalme, 2003], nous avons proposé une approche pour le résumé de multi-documents d'articles des journaux en utilisant les informations de *Background* pour produire un seul résumé pour un ensemble de document.

Cette méthode est le résultat de nos observations sur un corpus de journaux pour présenter une série d'événements pour les lecteurs humains. Nous avons étudié différents types de journaux pour comprendre ce qui est considéré comme une bonne méthode pour présenter une collection d'informations au sujet d'un concept spécifique à un lecteur et comment le lecteur peut combiner la nouvelle information de l'article courant avec l'information précédente pour parvenir à suivre une série d'événements.

Une bonne manière pour suivre une succession d'événements dans une histoire est de maintenir certaines informations de contexte (background) à propos de ce sujet et de rechercher de nouvelles informations sur ces topiques dans un nouvel article. Figure 2.1 montre un exemple pris de la première page du journal de la Gazette de Montréal où peut être trouvé *Focus* de l'événement, le **Background**, **New** et **Next** avec le numéro de référence de page pour trouver plus de détails au sujet de cet événement. Cette organisation facilite la poursuite et la compréhension d'une série des nouvelles. Avec un background court au sujet de chaque événement, un lecteur peut facilement retrouver l'information nécessaire dans les documents reliés. Avec le temps, un lecteur obtient plus d'informations sur le contexte pour un nouveau document sur un concept précis qui vont l'aider pendant la recherche de

nouvelles informations tout en évitant une certaine redondance qui apparaît nécessairement dans une série d'événements.

IN FOCUS

Credit-card numbers stolen

Background: Most credit-card holders are protected against unauthorized transactions on their accounts.
New: An "unauthorized intruder" gains access to 8 million credit-card account numbers – including Visa, MasterCard and American Express – by breaching the security of a company that processes transactions for merchants.
Next: Online merchants and security experts fear the stolen numbers will be used on Web sites and at call centres. **Page B1**

FIG. 2.1 – L'utilisation de l'information de *Background* pour garder la trace dans une série d'événement dans les articles des journaux (Montréal Gazette 19/02/2003).

2.3 Évaluer les systèmes de résumé automatique

L'évaluation de la qualité de résumés est un problème difficile, parce qu'il n'y a pas de résumé idéal pour un texte. Cette complexité du problème vient de difficulté théorique de définir une métrique claire pour les différents aspects de résumé comme la complétude, la thématique et la cohérence. Des métriques générales souvent utilisées, sont le rappel et la

précision. Rappelons que ces deux mesures sont calculées, pour une requête d'un utilisateur qui cherche des documents dans un fonds documentaire, à partir des trois paramètres suivants : P, nombre de documents non pertinents fournis par le système ; Q, nombre de documents pertinents fournis par le système ; R, nombre de documents pertinents présents dans le fonds documentaire et non fournis par le système. Le rappel et la précision sont alors calculés comme suit :

$$\text{Rappel} = Q/(Q+R)$$

$$\text{Précision} = Q/(P+Q)$$

Pour appliquer ce mode d'évaluation aux systèmes de résumé automatique, on considère non plus le nombre de documents pertinents, mais le nombre de phrases pertinentes, c'est-à-dire les phrases qui devraient être placées dans le résumé. Le principal défaut de ces critères est de postuler l'existence d'un résumé type, construit avec des phrases extraites du texte source, qui pourrait être utilisé comme référence.

Plusieurs compétitions d'évaluation (dans le modèle de NIST et TREC) reflètent l'intérêt des chercheurs dans le domaine du résumé automatique. Le Document Understanding Conference est une série d'évaluations soutenue par le *National Institute of Standards and Technology* (NIST) dans le cadre du projet TIDES, pour faire progresser la production de résumés automatique et pour permettre à des chercheurs de participer à des expériences à grande échelle. Dans le cadre de DUC, l'organisateur détermine les niveaux des performances de systèmes minimaux (baseline), pour chaque tâche dans les différents niveaux de compétition, et fournit des données de référence (documents et résumés correspondants) pour l'entraînement des systèmes.

Pour DUC 2002 dans le cadre de l'ACL 2002, il y avait les trois tâches suivantes :

1. Résumé automatique d'un mono document.
2. Résumé automatique des documents multiples : étant donné un ensemble de documents

sur un sujet simple, les participants doivent créer 4 résumés génériques de l'ensemble avec approximativement 50, 100, 200, et 400 mots.

3. Un ou deux projets pilotes avec l'évaluation extrinsèque.

NIST a produit 60 ensembles de référence d'une dizaine de documents. Ces ensembles proviennent des données de TREC utilisées dans les compétitions de questions et réponses lors de TREC-9. Chaque ensemble contient des documents courts définis selon quatre types de critères : un événement simple de catastrophe naturelle, un événement simple dans n'importe quel domaine, des événements distincts multiples du même type, et des informations biographiques sur un même individu. NIST a effectué deux types d'évaluation des résumés : une faite par des humains (pour les résumés générés), et une autre faite automatiquement (pour les extraits). Pour évaluer la qualité des résumés, NIST a utilisé douze questions qualitatives portant sur des aspects tels la grammaticalité, la cohérence et l'organisation des résumés.

Pour cette compétition nous avons participé avec une version modifiée de SumUM [Farzindar *et al.*, 2002b], le système de génération de résumé automatique des textes, qui avait été développé pour le résumé automatique de longs documents scientifiques et techniques. Malgré le fait qu'on avait utilisé SumUM sur un domaine complètement différent, les résumés produits par SumUM ont été jugés excellents par rapport à ceux produits par d'autres développés expressément pour ce genre de données et le résultats sont parmi les meilleurs de tous les systèmes qui ont participé à DUC 2002.

Pour DUC 2003 dans le cadre de HLT-NAACL 2003, les quatre tâches suivantes étaient définies :

1. Résumé très court ($\sim$10 mots) comme titre informatif d'un ensemble d'articles des journaux,
2. Résumé court ($\sim$100 mots) pour un ensemble de multi-documents,
3. Résumé court ($\sim$100 mots) selon un point de vue donné,

4. Résumé court (∼100 mots) de cet ensemble en réponse à une question.

Nous avons participé à l'évaluation de DUC pour une deuxième fois à partir d'une version modifiée du système SumUM, et nous avons concentré nos efforts sur le développement des possibilités de multi-documents et également ajouté quelques nouvelles fonctions additionnelles pour produire des résumés courts en réponse à une question [Farzindar et Lapalme, 2003].

Chapitre 3

Résumé des documents juridiques

Des milliers de professionnels utilisent l'information juridique et leurs ordres professionnels sont de plus en plus intéressés aux réponses apportées à ces besoins. D'autres travaillent à préparer ces produits d'information dans diverses maisons d'édition. D'autres encore, au gouvernement ou dans les universités, consacrent leurs énergies à augmenter l'offre gratuite des documents juridiques officiels.

Au Canada, l'institut canadien d'information juridique (CanLII) a comme objectif de créer une bibliothèque de droit virtuelle qui donnera un accès internet gratuit aux décisions de tous les tribunaux canadiens, ainsi qu'aux textes législatifs des gouvernements fédéral, provinciaux et territoriaux.

Cette abondance de textes juridiques sous forme numérique nécessite la création et la production d'outils informatiques performants en vue d'extraire l'information pertinente sous une forme condensée.

Mais pourquoi s'intéresse-t-on au traitement des décisions juridiques passées, ainsi qu'à leurs résumés ? D'abord parce qu'une décision judiciaire apporte généralement une solution à

un problème juridique entre deux ou plusieurs parties. Elle tient lieu de “loi” entre les parties. La décision comporte aussi des motifs qui justifient la solution. Donc les motifs constituent un *précédent* puisqu'il est possible d'en extraire une règle de droit pouvant servir à disposer d'affaires semblables. Alors ce qui arrive souvent, pour régler un problème juridique d'une affaire, dans une situation qui n'est pas claire, quand la solution ne se trouve pas directement dans la loi, les avocats cherchent les précédents pour les affaires semblables. Souvent pour une requête, dans une base de données de précédents, on va recevoir des centaines documents qui sont très longs à étudier. Lire tous ces documents, pour trouver les décisions pertinentes pour cette affaire pouvant être fastidieux, les experts et les étudiants de droit sont demandeurs de résumés de décisions judiciaires.

Sur le site de la bibliothèque de droit de Université de Montréal dans la rubrique *Ce mois-ci à la bibliothèque de droit - Questions et réponses sur l'index et résumés*, les étudiants demandent souvent la possibilité de savoir, si une décision donnée est repérable, sous forme de résumé dans les différents bases de données. Dans différents journaux, les journalistes demandent aux juges de publier des sommaires accompagnant leurs jugements pour faciliter les reportages par les médias.

On peut avoir intérêt à savoir, étant donné que plusieurs décisions sont disponibles en texte intégral dans les bases de données, comment on peut récupérer leurs résumés sous forme d'extraits condensés automatiques pour éviter la production manuelle de résumé qui peut être très coûteuse.

La Société québécoise d'information juridique (SOQUIJ) décrit l'objectif de production de résumé par l'humain pour les ressources juridiques comme suit :

> Le sommaire de l'arrêtiste poursuit deux objectifs : d'abord livrer l'essence du jugement clairement et avec concision pour permettre une consultation facile et rapide ; et de fournir suffisamment d'informations sur le jugement pour permettre au lecteur de décider, en connaissance de cause, si celui-ci peut être pertinent à

sa recherche.

Dans ce projet de thèse, en considérant ce besoin de production des résumés de textes légaux, nous allons préciser notre cadre de recherche pour enrichir une méthode de constitution automatique de fiches de résumés applicable pour le domaine juridique.

Dans ce chapitre, nous introduisons d'abord les notions de base de droit ainsi que les caractéristiques du domaine et du corpus. Par la suite nous décrivons l'état de l'art et les travaux pertinents sur la condensation automatique de texte et la problématique du domaine.

3.1 Types de résumés juridiques

Voici quelques types de résumés de texte juridique, en fonction du contenu du texte, du but spécifié pour chaque résumé, et de la tâche demandée par l'utilisateur :

Résumé indicatif — Le résumé d'un texte juridique mentionnant brièvement tous les sujets contenus dans ce dernier.

Exemple — Les résumés créés par le Répertoire électronique de jurisprudence du Barreau (REJB) et Société québécoise d'information juridique (SOQUIJ). Ils reprennent les textes intégraux de décisions comme les documents sources, qu'ils sont publiés par les différents cours judiciaires canadiennes comme la Cour d'appel, la Cour suprême du Canada, et etc.

URL – Société québécoise d'information juridique : `http://www.soquij.qc.ca/`
– Barreau du Québec : `http://www.barreau.qc.ca/`

Digest ou Fiche analytique — Extrait des éléments pertinents du texte de jugement.

Exemple — les fiches analytiques construites par le Bureau du Commissaire à la magistrature fédérale. Ces fiches sont des résumés des décisions intéressantes, mais pas assez importantes pour être publiées au recueil en entier.

URL — Recueils de la Cour fédérale : `http://reports.fja.gc.ca/`

Sommaire (Headnote) — Le résumé des points juridiques pertinents déterminés par une cour. Les sommaires sont écrits par les arrêtistes et ils sont utiles pour une consultation rapide du jugement. Le résumé de l'arrêtiste est toujours placé en-dessous de l'intitulé de la décision et transcrit en caractères italiques.

Résumés législatifs — Les résumés sont établis par la Direction de la recherche parlementaire de la Bibliothèque du Parlement. Ce sont les documents explicatifs sur les projets de loi, examinés par le Parlement. Il en existe un sur la plupart des projets de loi. Ils sont rédigés après l'étape de la première lecture, par des analystes qui connaissent bien le domaine des politiques ou du droit visé par le projet de loi.

URL — `http://www.parl.gc.ca/LEGISINFO/index.asp?Lang=F`

3.2 Corpus juridique

Alors que de grandes quantités de documents juridiques deviennent disponibles électroniquement, les manières d'organiser ces documents pour fournir des moyens pratiques d'accès prennent de l'importance. L'idée de résumé automatique de décisions judiciaires peut être une réponse à ce besoin. Il s'agit d'un texte court qui rend compte des éléments essentiels d'une décision afin de pouvoir réutiliser ultérieurement l'information résumée pour la recherche et les travaux.

Une ressource documentaire importante dans le domaine juridique est formée par les recueils de jurisprudence. La jurisprudence (case-law) réunit un ensemble des jugements, arrêts, décisions et avis publiés qui interprètent et précisent le sens des textes, les lois et règlements, par lesquelles les tribunaux supérieurs statuent sur des points de droit. Ces recueils, qui font l'objet d'annotations, ont été publiés dans des champs disciplinaires très divers : en matière civile, commerciale, administrative, etc. Leur présentation est le plus

souvent chronologique.

Pour ce projet de recherche, deux grands corpus de jurisprudence sont à notre disposition : le premier est la base des données *CanLII* (`http://www.canlii.org/`), qui est une ressource permanente en droit canadien. Il diffuse des sources primaires du droit. Il est réalisé par l'équipe LexUM du Centre de recherche en droit public (CRDP) de l'Université de Montréal. Le deuxième est le *Recueil des arrêts de la Cour fédérale* (`http://reports.fja.gc.ca/`), la Cour Fédérale applique du droit fédéral, qui est de type *"common law"*, c'est-à-dire que les *précédents* judiciaires- les décisions antérieures qui ont résolu un problème juridique- y jouent un rôle important comme source de droit. Le Bureau du Commissaire à la magistrature fédérale (BCMF) diffuse ces décisions de Cour fédérale ainsi que les résumés de certaines décisions.

Dans la première étape de ce travail de thèse, à la suggestion du CRDP, nous travaillerons sur les corpus de décisions publiés de la Cour fédérale en raison de son bilinguisme et du fait que ses décisions sont susceptibles d'intéresser des juristes provenant de toutes les provinces. Le fait d'avoir déjà des résumés modèles pour certaines décisions est aussi un avantage. Les décisions et les jugements délivrés par la cour fédérale, en général, sont bien structurées avec une rédaction judiciaire de bonne qualité. Nous pouvons donc nous en servir comme d'un corpus modèle pour notre étude.

3.3 Caractéristiques des textes

La différence entre le langage courant et le langage juridique change certains éléments à considérer pour créer un résumé. Par exemple, les statistiques des mots importants sont différentes par rapport à celles des autres corpus. Les articles des journaux répètent souvent le message le plus important, en droit, les termes pertinents peuvent n'apparaître qu'une seule fois.

À cause de ce langage particulier, la structure du document à traiter devient plus complexe. Nous avons étudié des notions de base en droit dans le domaine juridique, pour une meilleure compréhension du langage du droit. Nous nous intéressons plus particulièrement à l'étude de corpus du Recueil des arrêts de la Cour fédérale afin d'identifier la structure de textes juridiques et d'analyser les unités textuelles qui contiennent les informations importantes dans le texte qui pourraient être des candidates pour former un résumé du document. Ce Recueil est publié dans les deux langues ; anglais et français, et il reçoit 2000 décisions par année de la Cour fédérale. Les textes de jugement en tant que la jurisprudence rendue par le tribunal, est une source importante pour les juristes.

Étudier les méthodes d'analyse d'une décision juridique nous permet de bien connaître les différentes parties du texte de la jurisprudence qui contiennent les divers éléments avec les valeurs sémantiques différentes. À travers ce langage du droit, il faut identifier les informations importantes et éliminer les unités sans rapport direct avec l'idée principale du texte. Par exemple une "citation" sert à mettre en lumière certains aspects pendant le procès mais elle explique les aspects secondaires du document ; donc, dans le résumé, il faut la supprimer. Par contre identifier les motifs du juge est très important car les motifs constituent la partie la plus importante d'une décision.

Après avoir examiné les résumés manuels des rapports des décisions, nous avons remarqué qu'il existe une *macro structure* pour le texte du jugement, donc on peut définir un plan de décision qui contient les différents aspects du texte. La connaissance de cette organisation du texte facilite la sélection des informations importantes.

Le plan d'une décision juridique se divise en trois parties [Mailhot, 1998] [Laprise, 2000] : introduction, développement et conclusion. L'introduction et la conclusion forment chacune approximativement 10% du texte alors que le développement en constitue 80% [Boissonnault *et al.*, 1980]. Figure 3.1 montre le plan d'une jurisprudence, d'après la juge Mailhot de la Cour d'appel du Québec [Mailhot, 1996].

FAIRE UN PLAN

A- ***Données de la décision*** —

Date du jugement ;

Nom de la cour de décision ;

Identification des parties ;

Introduction — décrit brièvement la situation qui se présente au tribunal et répondre à ces questions : « QUI ? A FAIT QUOI ? À QUI ? »

Questions en litige — identifie le problème juridique.

Faits — recompose l'histoire du litige.

Prétentions des parties — le point de vue d'une partie.

B- ***Raisonnement juridique*** —

Discussion —

- exposé des prétentions respectives des parties ;
- commentaires du juge et détermination des faits ;
- expression des motifs(ratio) de la solution retenue.

Conclusion — le dispositif.

FIG. 3.1 – Plan de la jurisprudence se partage en deux parties : la première partie est **Données de la décision** qui contient les éléments sur lesquels s'appuie le juge pour motiver la décision, la deuxième partie est **Raisonnement juridique** qui constitue l'essence de la décision.

L'analyse d'un texte du jugement afin de résumer les décisions, se partage en deux parties : la première partie est **données de la décision** qui contient les éléments sur lesquels s'appuie le juge pour motiver sa décision. La deuxième partie est **raisonnement juridique** qui constitue l'essence de la décision en droit. Cette partie contient une solution au problème juridique qui pourra être réutilisée dans un autre cas semblable.

D'abord les données de la décision : **date du jugement**, **nom de la cour de décision** consiste à donner la référence complète de la décision (i.e. le nom de recueil, tribunal), **identification des parties** identifie la relation entre les parties sur le plan juridique (i.e. demandeur et défendeur), **introduction** décrit brièvement la situation qui se présente au tribunal et répondre à ces questions : « QUI ? A FAIT QUOI ? À QUI ? », **question(s) de droit** identifie le problème juridique dont le tribunal est saisi, **faits** recompose l'histoire du litige à partir des faits relatés lors de la présentation de la preuve et retenus dans le jugement, **prétentions et arguments des parties** concernent les prétentions en droit, le point de vue d'une partie sur le droit applicable au litige.

Deuxième partie d'une décision consiste en un raisonnement juridique qui répond aux questions de droit : **motifs du juge** constituent la partie la plus importante du résumé d'une décision, puisqu'ils sont la justification d'un dispositif qui transmet la solution. Les motifs du tribunal doivent être aussi la réponse aux questions de droit soulevées par les parties. Après un jugement, les motifs (ratio) deviennent règle de droit. **Conclusion** ou le dispositif détermine le sort ou solution trouvée réservé au litige. Par exemple en droit pénal, il faut spécifier si la personne a été condamnée ou acquittée.

3.4 Travaux en résumé de texte juridique

Nous présentons deux types de création de résumé juridique : la production manuelle et les approches automatiques.

Dans notre recherche, le résumé manuel pourrait être utilisé comme référence, c'est pourquoi nous décrivons les résumés produits par deux organismes canadiens qui produisent des résumés juridiques manuellement. Dans le chapitre 4, nous présenterons notre étude du corpus et l'alignement de résumés modèles et de textes sources pour en retirer les informations importantes.

3.4.1 Production de résumé manuel

Jurisprudence Express

Jurisprudence Express publie le résumé de tous les jugements motivés rendus par les cours de justice du Canada. Les résumés des décisions publiées, reprises en texte intégral dans le recueil, lequel contient également un plan de classification, une table de la législation citée, une table de la jurisprudence citée, une table de la doctrine citée, une table d'interprétation, un index, un suivi des appels ou révisions judiciaires des décisions publiées et une table de corrélation entre l'express et le recueil.

L'arrêtiste relève dans son traitement les principales références faites dans le jugement à la loi et à la jurisprudence. Les coordonnées du jugement (nom des parties, nom du juge, date, district, etc.) suivent le sommaire.

AZIMUT est la banque en ligne qui contient les résumés et les textes intégraux des jugements rapportés au Jurisprudence Express. Comme la plupart des résumés juridiques, l'accès aux résumés est payant. La figure 3.2 montre l'exemple d'un résumé manuel créé par Société québécoise d'information juridique (SOQUIJ). Ce plan expose les différents champs importants d'un jugement qui respecte aussi le schéma général d'une jurisprudence présenté à figure 3.1.

Résumés SOQUIJ 2001/02/06

Résumé 379 mots, décision 1514 mots, Texte intégral : 256 pages.

Parties — G. c. Québec (Procureur général) (G. c. P.G. du Québec) *

Juridiction — Cour d'appel (C.A.), Montréal, 500-09-001092-923

Décision de — Juges Mailhot, [...]

Date — 1999-04-23 Références AZ-99011340

Indexation — SOCIAL (DROIT) — sécurité du revenu — aide sociale — bénéficiaire de moins de 30 ans [...]

Interprétation — Charte canadienne des droits et libertés dans Loi de 1982 sur le Canada (L.R.C. 1985, app. II, no 44, annexe B, partie I), art. 7 [...]

Résumé — Appel d'un jugement de la Cour supérieure ayant rejeté une action en nullité, intentée par voie de recours collectif, de l'article 29 a) du Règlement sur l'aide sociale. Rejeté, avec dissidence.

L'appelante, qui représente tous les bénéficiaires de moins de 30 ans qui ont été assujettis à la disposition contestée entre le 17 avril 1987 et le 1er août 1989, soutient que celle-ci est inconstitutionnelle. Cette disposition avait pour effet, avant l'abrogation du règlement par l'adoption de la Loi sur la sécurité du revenu, entrée en vigueur le 1er août 1989, de réduire l'aide sociale versée aux bénéficiaires de moins de 30 ans, aptes au travail et vivant seuls. [...]

Décision — Mme la juge Mailhot : Le juge Iacobucci, dans l'arrêt Law c. Ministre de l'Emploi et de l'Immigration du Canada, énonce la démarche suivant laquelle il peut être décidé d'une atteinte au droit à l'égalité. Le programme d'aide accordé, envisagé dans sa globalité et dans son contexte, ne produit pas d'effets défavorables au sens de l'article 15 de la charte canadienne. Dans le cas contraire, l'article 1 de la charte rachèterait la mesure contestée. [...]

FIG. 3.2 – Exemple d'un résumé manuel écrit par SOQUIJ.

Fiches analytiques

Le Recueil des arrêts de la Cour fédérale rend disponibles toutes les décisions de la Cour fédérale destinées à être publiées dans le recueil officiel, dans leur intégralité ou sous forme de fiche analytique. Ces fiches analytiques sont des résumés des décisions intéressantes, mais pas assez importantes pour être publiées au recueil en entier. Il présente le contenu intégral du recueil officiel, anglais et français à compter de l'année 1993, publié dans les deux langues officielles. Ces décisions sont importantes pour des juristes de toutes les provinces du Canada, parce qu'ils représentent un précédent pouvant servir à disposer des cas semblables. Figure 3.3 montre un exemple du résumé qui est construit par extraction manuelle des segments pertinents du texte de décision de la Cours fédérale. Ce type de résumé est un bon exemple du résumé souhaité pour les juristes qui va nous servir comme référence.

3.4.2 Approches automatiques de production de résumé

De nombreuses approches sont proposées afin de fournir les outils qui permettent d'aider les avocats et les juristes. Les systèmes existants pour cet objectif, comme des compagnies privées QuickLaw au Canada et WESTLAW et LEXIS aux États Unis, il semble que jusqu'à nos jours le résultat ne satisfait pas complètement les exigences spécifiques de ce domaine d'application. Une raison de la difficulté de ce travail est la complexité du domaine. Pour développer un tel système du traitement du texte juridique, il faut une équipe composée d'informaticiens et de juristes pour analyser le problème.

Dans cette section nous présentons quelques modèles proposés pour produire les résumés de textes légaux.

Fiche analytique de la Cour fédérale

Résumé 209 mots, Texte intégral 2275 mots

Entre — Alli c. et Canada (Ministre de la Citoyenneté et de l'Immigration)

Dossier — IMM-1984-01

Référence neutre — 2002 CFPI 479

Juge — O'Keefe

Date — 26-4-02

Résumé —

Demande de contrôle judiciaire d'une décision de la SSR selon laquelle le demandeur n'était pas un réfugié au sens de la Convention–Le demandeur, un citoyen nigérian, avait revendiqué le statut de réfugié au sens de la Convention en se fondant sur sa religion et sur son appartenance à un groupe social, à savoir les personnes contraintes à participer à des meurtres rituels et à des pratiques néfastes–Le demandeur avait refusé d'assumer le poste que son père occupait à titre de chef Oluwo Apeno, soit un culte dans le cadre duquel on se livrait à des sacrifices humains ; il avait été déshérité par sa famille et s'était converti au christianisme-On avait dit au demandeur que son peuple serait béni par les dieux s'il était tué et si son sang était utilisé pour d'autres sacrifices–Le demandeur s'est enfui–La SSR avait conclu qu'il n'existait aucune possibilité sérieuse que le demandeur soit persécuté s'il retournait au Nigéria et que l'État pouvait assurer à ce dernier une protection adéquate–Demande accueillie–La Commission a commis une erreur susceptible de révision au sujet des faits se rapportant à la protection étatique étant donné que la preuve documentaire montrait que la protection étatique ne pouvait pas être obtenue de la police.

Fig. 3.3 – Une fiche analytique d'une décision de la Cour fédérale.

FLEXICON, University of British Columbia

FLEXICON (Fast Legal Expert Information CONsultant) [Smith et Deedman, 1987][Gelbart et Smith, 1991][Smith *et al.*, 1995] est un système développé pour la gestion des informations juridiques et la production du résumé qui combine le traitement du texte avec raisonnement à base de cas. Cette approche utilise des modules d'extraction pour identifier les concepts, les cas, les législations, les faits et leurs relations dans la décision, afin de construire un profil structuré de document et produire automatiquement un sommaire (headnote). Les concepts sont identifiés par unification des mots du texte avec une liste d'expressions significatives, en appliquant des règles heuristiques simples.

FLEXICON génère des représentations de structures de documents juridiques. Le système effectue les opérations suivantes :

- extraction des informations d'en-tête sur le cas (case header) : les parties, la date, la cour, la juridiction, et des juges ;
- classification du sujet de la loi ;
- identification des paragraphes clés contenant l'essence de l'opinion de la cour.

Pour présenter les parties importantes du texte, le système génère une liste de quatre aspects juridiques : des concepts les plus significatifs, des faits, des cas et de l'ensemble des lois appliquées. Il calcule les poids de cette liste par ordre décroissant. Il extrait les paragraphes importants au complet et il élimine les paragraphes très courts et ceux qui contiennent les *citations* moins importantes dans le jugement.

Pour notre travail, malgré la simplicité de cette approche, le profil du document proposé nous intéresse, car ce système a été développé pour les textes juridiques canadiens de type *common law* qui est notre corpus de texte.

SALOMON, Katholieke Universiteit Leuven

Le projet SALOMON [Uyttendaele *et al.*, 1996][Moens *et al.*, 1996] produit le résumé automatiquement de cas criminels belges (écrits en Hollandais). Ce système extrait les unités importantes des documents à partir du texte de jugement pour former un résumé. Le but est d'identifier et d'extraire les informations importantes à partir des jurisprudences.

Deux méthodologies ont été utilisées pour développer SALOMON. D'abord il identifie la catégorie de cas, la structure et les unités non pertinentes des textes. Ce processus est basé sur la représentation des connaissances réunie dans une grammaire de texte. Ensuite, des données générales et fondamentales du sujet de la décision sont extraites. Deuxièmement, le système produit un résumé informatif des unités textuelles de l'opinion de la cour en utilisant les techniques statistiques.

Dans ce projet les connaissances linguistiques plus profondes sont utilisées. Il extrait les concepts et les unités textuelles saillantes grâce à l'identification des *cue words*, segments indicateurs et patrons de contexte développé en hollandais. Cette recherche montre aussi l'intérêt d'utiliser la structure du discours des textes juridiques.

SUM, University of Edinburgh

Ce projet [Grover *et al.*, 2003] utilise l'information rhétorique et la structure du discours pour générer un résumé flexible, avec un taux de compression élevé qui s'applique sur les textes légaux. Dans cette étude, les propriétés structurales du texte sont examinées pour identifier les parties plus importantes de texte source. Un échantillon de corpus des jugements de House of Lords, la cour suprême de l'Angleterre, a été annoté comme ressource afin déterminer les rôles rhétoriques et de produire les résumés. Il place les phrases extraites par rapport à la représentation structurée du texte, ensuite il les combine pour générer le résumé.

Il est basé sur l'approche de [Teufel et Moens, 2002] proposée pour les articles scientifiques. SUM résume les textes légaux en deux étapes :

- Décider quels rôles argumentatifs sont importants dans le texte source et lesquels sont utiles dans le résumé.
- Dans une collection de textes pertinents, décider pour chaque phrase, le rôle argumentatif qui la décrit mieux.

Ce projet est en cours de développement, leur dernier article rapporte que l'identification de temps des groupe de verbes et d'identification de groupe principal de verbe sont terminés.

3.5 Discussion des méthodes de génération de résumé

Alors que nous avons le problème de grandes quantités de textes juridiques et le besoin de les présenter sous forme d'un résumé court, notre recherche montre qu'il n'y pas en beaucoup de travail sur ce domaine et que le problème du traitement du texte légaux reste sans solution.

Les différentes méthodes utilisées pour traiter des textes juridiques montrent qu'une diversité de systèmes sont utilisés pour les tâches d'indexation, de catégorisation, et de production de résumés. La création automatique de résumé, semble la plus compliquée et demande une étude approfondie du sujet. Les différentes approches proposées essayent de régler une partie du problème mais le résultat d'évaluation des systèmes et la qualité des résumés produits n'est pas encore à un niveau satisfaisant.

Différents systèmes ont été développés pour différentes langues, comme l'allemand, le hollandais, l'anglais ou le français, mais une approche qui peut être efficace pour identifier les indicateurs marquants les phrases importants dans une langue ne sera peut-être pas aussi utile pour d'autres langues avec d'autres types de styles. Il en est de même pour l'organisation du jugement qui peut différer selon les lois et la tradition juridique de chaque pays.

Dans le chapitre suivant nous présentons la démarche de notre thèse suivie du plan de recherche qui nous amènera à développer un système de constitution automatique de résumé de textes juridiques.

Chapitre 4

Démarche et thèmes de recherche

Notre principal objectif est de concevoir un système de résumé automatique pour les textes juridiques de type jurisprudence. Le but d'un tel résumé est de présenter les principaux points traités dans les jugements pour permettre au lecteur de consulter rapidement l'extrait de la décision afin de trouver les documents pertinents. La réalisation de ce projet demande une analyse spécifique pour exploiter les connaissances nécessaires sur les différents aspects linguistique, informatique et juridique des textes légaux. Dans le cadre de ce travail de thèse nous profitons d'une collaboration entre trois équipes de linguistique informatique et de droit (LaLICC, RALI et CRDP).

Afin d'illustrer les particularités de notre domaine, nous allons présenter les différentes étapes de notre étude sur les textes juridiques. D'abord nous décrivons l'étude de corpus sur les différentes collections de données juridiques. Cette étude sur les textes juridiques se réalise en deux parties, d'abord nous avons effectué une étude globale des décisions judiciaires délivrées par les différentes cours canadiennes dans les deux langues, anglais et français, ainsi qu'une recherche sur les méthodes actuelles de production des résumés manuels. La deuxième partie de l'étude de corpus va porter sur des décisions de la Cour fédérale du Canada en anglais. Cette partie a pour but d'identifier les structures de texte ainsi que les marqueurs

linguistiques et les indicateurs (features) qui ont les valeurs importantes dans les jugements. Ces indicateurs et les informations qu'ils apportent nous aideront à identifier les unités textuelles pertinentes dans le texte source.

Dans cette section nous décrivons notre méthode de développement d'un système de résumé automatique des textes légaux.

4.1 Étude de corpus sur les textes juridiques

Dans la première étape d'étude de corpus, notre principal objectif est de connaître les outils de rédaction juridique et de structuration du texte juridique en général. À la suggestion du CRDP, nous avons étudié deux collections de documents juridiques, d'abord une échantillon de corpus, 8 jugements et 8 résumés, produits par REJB (Répertoire électronique de jurisprudence du Barreau) en français, ensuite une collection de résumés des fiches analytiques du Recueil des arrêts de la Cour fédérale du Canada et leurs textes intégraux, en anglais et français. Pour identifier les points considérés importants par les arrêtistes et les résumeurs professionnels, nous avons aligné manuellement ces résumés et leur texte source. Nous allons présenter la description de ces corpus et nos observations.

A- L'échantillon de corpus, 8 jugements et 8 résumés produits par REJB

Nous avons analysé en détail cet échantillon de corpus. Les documents sources viennent de la collection de Cour du Québec `http://www.canlii.org/qc/jug/qccq/`

Cette collection présente des décisions rendues par la Cour Supérieure du Québec depuis 1997. Table 4.1 montre le taux de compression de texte original pour produire le résumé

manuel, qui est calculé par le nombre de mots du résumé cible divisé par le nombre de mots de texte source.

	Langue	No Réf	# mots Doc	# mots Résumé	% Cmpr	Extraction
1	Français	REJB 2001-27340	14 319	558	4%	Non
2	Français	REJB 2002-32023	4 967	283	6%	Non
3	Français	REJB 2002-35246	7 516	733	10%	Non
4	Français	REJB 2000-16759	5 134	526	10%	Non
5	Français	REJB 2002-35246	4 985	862	17%	Non
6	Français	REJB 2003-37475	3 955	685	17%	Non
7	Français	REJB 2003-38272	5 910	1 035	18%	Non
8	Français	REJB 2000-16192	1 879	340	18%	Oui

Tab. 4.1 – Les résumés REJB et les textes sources sont alignés. Les différentes unités du résumé sont extraites des différents paragraphes du texte original de la décision.

Observations

Il y a 8 jugements et 8 résumés de mono-document générés manuellement, comme échantillon de corpus. Les résumés de REJB sont de type littéraires : le texte du résumé est reformulé et les éléments du jugement sont réorganisés selon la compréhension de l'humain. Le résumé littéraire est difficile à aligner avec le texte original à cause de certains segments dans le résumé qui sont des extraits de plusieurs unités de texte source.

Seul le résumé numéro 8, REJB 2000-16192, a été construit par le CRDP à partir du résumé de REJB, pour le ramener de 1800 mots (texte source) à 340 mots (résumé). Dans cet exemple le taux de compression est 18% et la technique utilisée est l'extraction des phrases importantes, l'assemblage de ces phrases sélectionnées forme un court texte qui montre le résumé du type souhaité par le CRDP. D'après leur demande, un résumé plus télégraphique d'environ 200 mots serait également intéressant. La moyenne de la taille des résumés étudiés est 12.5% du document original.

La constitution d'un dictionnaire spécialisé des concepts pour les textes juridique, qui met en valeur les marqueurs linguistiques les plus importants dans le domaine, peut faciliter la tâche de résumer. D'après *plan de jurisprudence* il y a des sections séparées comme *Fait, Conclusion, ...* qui semblent plus importantes et qui contiennent plus d'informations sur le sujet. Dans la plupart des cas, la première section de texte définit la situation, les sections au milieu décrivent le procès et les dernières sections sont les conclusions et les résultats de jugement. En théorie, et ce qui est souhaitable, en termes de rédaction d'une décision, le début d'un jugement est de toute première importance. Pour énoncer la conclusion il y deux possibilités : annoncer la conclusion au début, l'énoncer à la fin. Traditionnellement, les juges n'expriment leur conclusion qu'à la fin après avoir exposé les faits, les questions à décider, les arguments présentés par les diverses parties et les diverses prémisses qui sous-tendent leur raisonnement et leur cheminement vers la solution retenue.

Certains termes utilisés dans les textes des résumés modèles étudiés sont plutôt des abréviations des termes de document source. Pour vérifier la signification de l'abréviation utilisé au recueil de jurisprudence, il est possible de consulter la liste des abréviations juridiques, disponible à la bibliothèque de droit, disponible aussi à l'URL :
`http://www.bib.umontreal.ca/DR/abrev.htm`.

B- Une collection de résumés des fiches analytiques de Recueil des arrêts de la Cour fédérale du Canada

Le deuxième corpus étudié (voir Table 4.2) est le recueil des arrêts de la Cour fédérale du Canada tel que décrit à la section 3.2.2. Certaines décisions de la Cour fédérale sont disponibles sous forme des résumés des fiches analytiques bilingues anglais et français à `http://reports.fja.gc.ca/`. Nous avons analysé 4 groupes de dossiers, que nous avions choisis parmi les différents volumes de jugements dans les différentes années. Les textes légaux étudiés sont en anglais et français et nous avons aligné les condensés avec leur textes

intégraux afin d'identifier les relations entre le texte source et son résumé pour analyser la structure du texte de la décision et les segments considérés importants par les arrêtistes.

	Volume		Langue	No Réf	# mots Doc	# mots Rés	% Cmpr	Extr
1	2003 vol. 1	Trad.	Français	2002FCA253-fr	455	143	31%	Oui
	2003 vol. 1		Anglais	2002FCA253-en	397	95	24%	Oui
2	2002 vol. 1	Trad.	Français	2002 FCT479-fr	2 275	209	9%	∼Oui
	2002 vol. 1		Anglais	2002 FCT479-en	1 928	121	6%	∼Oui
3	2001 vol. 1	Trad.	Français	A-63-00-fr	678	163	24%	∼Oui
	2001 vol. 1		Anglais	A-63-00-en	654	126	19%	∼Oui
4	2000 vol. 1		Français	IMM-6717-98-fr	1 555	221	14%	Oui
	2001 vol. 1	Trad.	Anglais	IMM-6717-98-en	1 443	143	10%	Oui

TAB. 4.2 – Les résumés et les textes intégraux des différents volumes de la Cour fédérale, la technique utilisée pour produire de résumé est l'extraction manuelle des unités importantes.

Observations

Les textes intégraux et les résumés sont étudiés en deux langues : anglais et français. Les textes en deux langues sont des corpus parallèles. La décision et son résumé sont rédigés dans une langue, il y a par la suite une traduction officielle de ce jugement et le résumé dans l'autre langue. Les traductions sont certifiées, donc les textes traduits sont très bien construits.

La moyenne de la taille de résumé de cette collection en anglais est 14.75% et en français est 19.5% du texte source. Les textes de jugements sont accompagnés une section supplémentaire (*Order* en anglais et *Ordonnance* en français) qui mentionne les points juridiques pertinents déterminés par la cour. Ils sont écrits par les arrêtistes de la cour fédérale et ils pourraient être utilisés au moment d'une consultation rapide du jugement. Les résumés sont produits plutôt avec la technique d'extraction mais certaines modifications sont ajoutées à ces extraits Figure 4.1 montre un exemple de ces extraits. Pour deux langues, anglais et français, de document A-63-00, il compare le texte intégrale avec son résumé. La différence

entre les phrases du texte source et les phrases extraites du résumé est montrée en caractères italiques et gras. Les mots italiques, dans les textes intégraux montrent les mots à supprimer pour faire le résumé, et les mots gras dans le résumé montrent les mots ajoutés ou modifiés.

On peut extraire certaines informations du numéro d'identification du jurisprudence. Pour chaque décision il y a deux numéros, un *numéro de référence neutre* qui est unique pour chaque jugement et un *numéro du dossier* qui contient tous les jugements de différentes cours sur un problème juridique. Voici quelques informations cachées du numéro du jurisprudence : au niveaux de la Cour fédérale le *numéro de dossier* qui comporte un "T" montre une décision de "*Trial*", i.e. T-787-00, et un "A" pour "*Appeal*", i.e. A-706-96. Pour le *numéro de référence neutre* les quatre premiers chiffres montrent l'année du jugement, les trois caractères font une distinction entre la Cour fédérale et la Cour fédérale première instance et les deux dernières lettres montrent la langue du texte i.e. 2002 FCT479-fr.

Nous avons remarqué qu'il y a déjà une certaine classification pour les textes par rapport aux noms de fichier : par exemple les références « *IMM* » correspondent à «LE MINISTRE DE LA CITOYENNETÉ ET DE L'IMMIGRATION », donc la structure de texte est presque prédéfinie, le Figure 4.2 compare deux dossiers *IMM*. Dans les deux dossiers les premiers paragraphes, qui définissent la situation de cas, utilisent les patrons predéfinis. Les mots écrits en gras sont les mots réservés qui se répètent dans les dossiers du même catégorie.

4.2 Résultat de l'étude de corpus

Nous comptons nous concentrer sur la collection *Recueil des arrêts de la Cour fédérale* en anglais, qui est une référence importante pour tous les avocats et les experts en Amérique du nord. D'après nos observations, les textes de décisions juridiques de la Cour fédérale sont bien structurés avec une bonne qualité de rédaction. Malgré ces aspects positifs des textes de décisions des jugements, jusqu'à nos jours les systèmes de production de résumé

Texte intégral Dossier : A-63-00	Résumé
[1] *The issue in this appeal is whether an* Appeal Board of the Public Service Commission has the jurisdiction to hear and decide appeals under subsection 21(1.1) of the Public Service Employment Act by persons who do not meet the criteria established under subsection 13(1).	**Appeal from decision by Sharlow J. ([2000] 2 F.C. D-15)** Appeal Board of Public Service Commission does not have jurisdiction to hear, decide appeals under Public Service Employment Act, **s.** 21(1.1) by persons who do not meet criteria established under s. 13(1)
[1] *La question posée dans cet* appel *consiste à savoir* si un comité d'appel de la Commission de la fonction publique est habilité, en vertu du paragraphe 21(1.1) de la Loi sur l'emploi dans la fonction publique, à connaître et trancher des appels de la part de personnes qui ne satisfont pas aux critères prévus au paragraphe 13(1).	**A**ppel **de la décision du juge Sharlow ([2000] 2 C.F. F-9) que** le comité d'appel de la Commission de la fonction publique **n'a pas compétence pour connaître et trancher des appels** en vertu de l'art. 21(1.1) de la Loi sur l'emploi dans la fonction publique **interjetés** par **des** personnes qui ne satisfont pas aux critères prévus **à l'art**. 13(1)

FIG. 4.1 – Dans deux langues, les résumés sont produits avec la technique d'extraction mais certaines ajustements sont faits à ces extraits. Les caractères italiques, dans les textes intégraux montrent les mots à supprimer pour faire le résumé, et les mots gras dans le résumé montrent les mots ajoutés ou modifiés.

Dossier : **IMM-6717-98** Date : 1999-10-25	Dossier : **IMM-5163-01** Date : 2003-01-10
[1] **Il s'agit d'une demande de contrôle judiciaire d'une décision rendue par la Commission de l'immigration et du Statut de Réfugié,** (Section du statut), en date du 30 novembre 1998 **qui concluait que les demandeurs n'étaient pas des réfugiés au sens de la Convention.**	**Il s'agit ici d'une demande de contrôle judiciaire d'une décision de la Section du statut de réfugié de la Commission de l'immigration et du statut de réfugié**(la « CISR ») **rendue** le 15 octobre 2001 **statuant que les demandeurs ne sont pas des réfugiés au sens de la Convention,** tels que définis au paragraphe 2(1) de la Loi sur l'immigration, L.R.C. (1985), ch. I-2 (la « Loi »).
[2] **Les demandeurs** sont **citoyens** République Démocratique du Congo. Ils allèguent une crainte de persécution en raison **d'opinions politiques** et de leur appartenance à un groupe social particulier, soit la famille.	**Le demandeur** principal, sa femme et leurs deux filles mineures sont des **citoyens** mexicains. Ils revendiquent le statut de réfugiés du fait **des opinions politiques** imputées au demandeur principal et, pour sa femmes et leurs filles, de l'appartenance à un groupe social, soit « la famille ».

FIG. 4.2 – Comparaison de ces dossiers montrent les patrons prédéfinis pour les textes de mêmes catégories. Les mots écrits en gras sont les mots réservés qui se répètent dans les dossiers du sujet semblable.

automatique ont des difficultés à traiter ce genre de texte. À cause du langage particulier utilisé en rédaction d'une telle décision, la qualité du résumé produit automatiquement par les systèmes de résumé est très basse. Nous avons effectué quelques tests avec le système SumUM actuel, qui produit les résumés des articles scientifiques et qui est jugé un système de bonne qualité au niveau de génération du résumé, mais le résultat a été décevant. Sur la vingtaine de phrases marquées importantes dans les résumés modèles écrits manuellement, une seule phrase a été identifiée par SumUM. La complexité du texte juridique représente un défi pour les systèmes de résumés. Cette expérience montre l'intérêt de développer un système explicitement pour résumer les textes juridiques qui peut répondre à une vaste demande de domaine du droit.

4.3 Méthode

Dans cette recherche, nous allons enrichir les approches proposées pour la production du résumé. Le projet sera de développer un système qui prendra en entrée un texte d'une décision juridique de la Cour fédérale du Canada et la sortie sera un résumé court (un taux de compression de 10% environ) du texte source. D'après nos études sur les méthodes de productions de résumé manuel par les professionnels, la technique adoptée est d'extraire les informations importantes dans le texte et de les représenter sous la forme d'un résumé cohérent. Nous allons automatiser ce processus pour créer le résumé juridique pendant notre recherche.

Dans la deuxième partie de notre étude de corpus, nous identifierons les indicateurs linguistiques dans le texte qui signalent les informations importantes du texte. À l'aide des indicateurs et des informations qui les porte, on peut segmenter le texte en différentes unités textuelles. Ces segments vont être pondérés d'après leur importance afin de trouver les phrases candidates qui semblent pertinentes. En fonction de la taille demandée du résumé, il y aura un critère de sélection. Cette étape consiste à sélectionner des phrases ou des unités

textuelles importantes, d'après certaines règles sémantiques qui contrôlent les conditions pour les candidats. La dernière étape combinera les parties extraites afin de produire le résumé du texte source.

4.4 Évaluation

Pour évaluer la qualité de sortie du système, nous proposons deux méthodes :

1. Quand le résumé modèle existe : nous pouvons comparer le résumé généré par le système et le résumé modèle. Lors du workshop DUC pour évaluation manuelle, NIST a utilisé une version de SEE [Lin, 2001], qui permet d'aligner automatiquement les unités de deux textes pour comparer la similarité entre les résumés modèles et les résumés produits afin de calculer la couverture (coverage) qui calcule la fraction du résumé modèle exprimé dans le contenu du résumé produit par le système.
2. Quand le résumé modèle n'est pas disponible : avec la méthode *Delphi* qui est une réunion d'un groupe d'experts du domaine, où chacun évalue les résumés produits par le système. Suite à une discussion le groupe donne un résultat d'évaluation final pour chaque résumé généré. Dans notre cas, le groupe sera composé des avocats et des informaticiens qui évalueront les résumés du système. Chaque membre du jury répondra à une série de questions concernant la qualité du résumé, à la fin pour chaque résumé produit on réunira les avis du jury.

Pour évaluer les aspects tels la grammaticalité, la cohérence et l'organisation des résumés, NIST a utilisé douze questions qualitatives sur ce sujet. Voici quelques exemples des questions utilisées par NIST lors d'évaluation de DUC 2002 et 2003 : *About how many gross capitalization errors are there ? About how many sentences have incorrect word order ? About many instances of unnecessarily repeated information are there ?*

4.5 Échéancier des travaux

Nous proposons l'échéancier suivant pour effectuer nos travaux dans la Table 4.3.

4.6 Conclusion

Nous avons présenté notre proposition du sujet de thèse ainsi que la démarche de notre recherche pour résoudre le problème de création automatique de résumé de textes juridique. Notre approche repose sur des techniques d'extraction des unités saillantes des décisions judiciaires des cours canadiennes, pour former un résumé indicatif qui fournit les idées clés du texte source.

Notre revue de littérature du résumé automatique et les expériences des participations des différents workshops internationaux sur ce sujet, nous a permis de déterminer que la qualité des résumés produits par les différents systèmes de recherches est loin du résultat attendu. Pour cette nouvelle application du résumé juridique, les approches actuelles ne proposent pas de techniques efficaces.

Notre contribution dans cette recherche portera sur les aspects suivants :

- Identification de la structuration du texte de jurisprudence ;
- Identification des concepts significatifs du domaine juridique et des critères de sélection de portions importantes de textes ; et
- Production d'un résumé court dans la limite de la taille souhaitée.

Le résumé doit couvrir les informations maximales dans un espace minimal, donc le processus d'évaluation du système et les résultats produits, sera dernière étape de ce projet.

Période	Travaux
Sept 01	Début du doctorat à l'Univérsité ParisIV-Sorbonne
Janv 02	Début du doctorat à l'Université de Montréal
Fév 02	Présentation RALI sur «*Résumé automatique des textes*»
Avr 02	Participation à la compétition DUC 2002 pour évaluer SumUM
Juil 02	Présentation et Publication à *Text Summarization Workshop, ACL*, Philadelphia, U.S.A.
Oct 02	Présentation RALI sur «*Summaries with SumUM and its Expansion for DUC 2002*»
Déc 02	Présentation et Publication à l'*ATALA*, *Le résumé de texte automatique : solutions et perspectives*, Paris, France.
Fév 03	Participation à la compétition DUC2003
Avr 03	Fin des cours et examens écrits de connaissances générales
Juin 03	Présentation et Publication à *Text Summarization Workshop, NAACL*, Edmonton, Canada
Sept- Déc 03	Étude sur les corpus juridiques et identification des marqueurs significatifs dans les textes légaux
Janv- Août 04	Elaboration du prototype de résumé automatique
Sept- Nov 04	Évaluation du système
Déc 04	Rédaction de la thèse

TAB. 4.3 – Échéancier des travaux.

Bibliographie

[Barzilay *et al.*, 1999] Regina Barzilay, Kathleen McKeown et M. Elhadad. Information Fusion in the Context of Multi-Document Summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, Maryland, USA, June 20–26 1999.

[Barzilay et Elhadad, 1997] Regina Barzilay et M. Elhadad. Using Lexical Chains for Text Summarization. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain, July 1997.

[Boissonnault *et al.*, 1980] Pierre Boissonnault, Roger Fafard et Vital Gadbois. *La dissertation, outil de communication*. Edition la Lignée, Ste-Julie, 1980.

[Borges *et al.*, 2001] Filipe Borges, Danièle Bourcier, Evelyne Andreewsky et Raoul Borges. Conception of cognitive interfaces for legal knowledge. In *ICAIL 2001*, pages 231–232, St. Louis, U.S.A., 2001.

[Borko et Bernier, 1975] Harold Borko et Charles Bernier. *Abstracting Concepts and Methods*. Academic Press, New York, 1975.

[Charolles, 1991] M. Charolles. Le résumé de texte scolaire. Fonctions et principes d'élaboration. *Pratiques*, 72 :7–27, Décembre 1991.

[Cremmins, 1996] Edward T. Cremmins. *The Art of Abstracting*. Information Resources Press, Arlington, VA, 2nd édition, 1996.

[Desclés, 1988] Jean-Pierre Desclés. Langage et cognition. *RSSI, Association Canadienne de Sémiotique*, 8(2), 1988.

[Desclés, 1997] Jean-Pierre Desclés. Systèmes dexploration contextuelle. *Co-texte et calcul du sens, (Claude Guimier). Presses de luniversitaires de Caen*, pages 215–232, 1997.

[DUC, 2003] DUC. Document Understanding Conference 2003. NAACL, Text Summarization Workshop. `http://duc.nist.gov`, May 31 - June 1 2003.

[Edmundson, 1964] H. P. Edmundson. Problems in Automatic Extracting. *Communications of the Association for Computing Machinery*, 7 :259–263, 1964.

[Edmundson, 1969] H. P. Edmundson. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16(2) :264–285, April 1969.

[Endres-Niggemeyer, 2000] Brigitte Endres-Niggemeyer. SimSum : an empirically founded simulation of summarizing. *Information Processing & Management*, 36 :659–682, 2000.

[Farzindar *et al.*, 2002a] Atefeh Farzindar, Guy Lapalme et Horacio Saggion. Evaluation à DUC2002 d'une adaptation de SumUM, un système de résumé automatique. In *Association pour le Traitement Automatique des LAngues (ATALA) Le résumé de texte automatique : solutions et perspectives*, Paris, France, décembre 14 2002.

[Farzindar *et al.*, 2002b] Atefeh Farzindar, Guy Lapalme et Horacio Saggion. Summaries with SumUM and its Expansion for Document Understanding Conference. In *DUC02 : ACL'2002 Workshop in Automatic Text Summarization*, Philadelphia, U.S.A., July 6–12 2002.

[Farzindar et Lapalme, 2003] Atefeh Farzindar et Guy Lapalme. Using Background Information for Multi-Document Summarization and Summaries in Response to a Question. In *DUC03 : NAACL'2003 Workshop in Automatic Text Summarization*, Edmonton, Alberta, Canada, May 31 - June 1 2003.

[Gelbart et Smith, 1991] Daphne Gelbart et J. C. Smith. Beyond Boolean search, Flexicon, a legal text-based intelligent system. In *the Third International Conference on Artificial Intelligence and Law*, New York, U.S.A., 1991.

[Grover *et al.*, 2003] Claire Grover, Ben Hachey et Chris Korycinski. Summarising Legal Texts : Sentential Tense and Argumentative Roles. In *HLT-NAACL 2003 Workshop : Text Summarization (DUC03)*, Dragomir Radev et Simone Teufel, éditeurs, pages 33–40, Edmonton, Alberta, Canada, May 31 - June 1 2003.

[Hahn, 1990] Udo Hahn. Topic Parsing : Accounting for Text Macro Structures in Full-Text Analysis. *Information Processing and Management*, 26(1) :135–170, 1990.

[Hovy et Lin, 1999] Eduard Hovy et Chin Yew Lin. Automated Text Summarization in SUMMARIST. In *Advances in Automatic Text Summarization*, Inderjeet Mani et Maybury Mark T., éditeurs, pages 81–94. The MIT Press, 1999.

[Jing et McKeown, 1999] Hongyan Jing et Kathleen R. McKeown. The Decomposition of Human-Written Summary Sentences. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*,

M. Hearst, Gey. F. et R. Tong, éditeurs, pages 129–136, University of California, Beekely, August 1999.

[Jing et McKeown, 2000] Hongyan Jing et Kathleen R. McKeown. Cut and Paste-Based Text Summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference and the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 178–185, Seattle, WA, April 2000.

[Jing, 2002] Hongyan Jing. Using Hidden Markov Modelling to Decompose Human-Written Summaries. *Computational Linguistics*, 28(4), 2002.

[Laprise, 2000] Gisèle Laprise. *Les outils du raisonnement et de la rédaction juridiques*. Les Editions Thémis, 2000.

[Lexis,] Lexis. Compagnie. U.S.A. `http://www.lexis.com/`.

[Lin et Hovy, 1997] Chin-Yew Lin et Eduard Hovy. Identifying Topics by Position. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 283–290. Association for Computational Linguistics, March 31 - April 3 1997.

[Lin et Hovy, 2002] Chin-Yew Lin et Eduard Hovy. From Single to Multi-document Summarization : A Prototype System and its Evaluation. In *Proceedings of the 2nd Document Understanding Conference at the 40th Meeting of the Association for Computational Linguistics*, pages 457–464, Philadelphia, PA, July 2002.

[Lin, 1995] Chin-Yew Lin. Knowledge-Based Automatic Topic Identification. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. 26-30 June 1995, MIT, Cambridge, Massachusetts, USA*, pages 308–310. ACL, 1995.

[Lin, 2001] Chin-Yew Lin. Summary Evaluation Environment, 2001. `http://www.isi.edu/$\sim$cyl/SEE`.

[Lluelles, 2000] Didier Lluelles. *Guide des références pour la rédaction juridique*. Editions Thémis :, Montréal, Québec, Canada, 2000.

[Lunh, 1959] H.P. Lunh. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, pages 159–165, 1959.

[Mailhot, 1996] Louise Mailhot. *Ecrire la décision : guide pratique de rédaction judiciaire*. Editions Yvon Blais, Québec, Canada, 1996.

[Mailhot, 1998] Louise Mailhot. *Decisions, Decisions : a handbook for judicial writing*. Editions Yvon Blais, Québec, Canada, 1998.

[Mani *et al.*, 1998] Inderjeet Mani, David House, G. Klein, Lynette Hirshman, Leo Orbst, Thérèse Firmin, Michael Chrzanowski et Beth Sundheim. The TIPSTER SUMMAC Text Summarization Evaluation. Rapport Technique MTR 98W0000138, The Mitre Corporation, McLean, Virginia, 1998.

[Mani et Maybury, 1999] Inderjeet Mani et Mark Maybury. *Advances in automatic text summarization*. Kluwer Academic Publishers, Boston, U.S.A., 1999.

[Mani, 2001] Inderjeet Mani. *Automatic Text Summarization*. John Benjamins Publishing Company, 2001.

[Marcu, 1997] Daniel Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD thesis, University of Toronto, 1997.

[Marcu, 2000] Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge/London, 2000.

[McKeown et Radev, 1995] Kathleen R. McKeown et Dragomir R. Radev. Generating Summaries of Multiple News Articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, Seattle, Washington, July 1995.

[Minel *et al.*, 1997] Jean Luc Minel, Sylvaine Nugier et Gerald Piat. How To Appreciate the Quality of Automatic Text Summarization ? Examples of FAN and MLUCE Protocols and Their Results on SERAPHIN. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization at the 35th Meeting of the Association for Computational Linguistics, and the 8th Conference of the European Chapter of the Assocation for Computational Linguistics*, Inderjeet Mani et Mark T. Maybury, éditeurs, pages 25–30, Madrid, Spain, 1997.

[Minel *et al.*, 2001] Jean-Luc Minel, Jean-Pierre Desclés, Emmanuel Cartier, Gustavo Crispino, Slim Ben Hazez et Agata Jackiewicz. Résumé automatique par filtrage sémantique d'informations dans des textes. *Revue Technique et Science Informatiques*, (3), 2001.

[Minel, 2002] Jean-Luc Minel. *Filtrage sémantique : du résumé automatique à la fouille de textes*. Editions Hermès, Paris, France, 2002.

[Moens *et al.*, 1996] M.-F. Moens, R. Gebruers et C. Uyttendaele. SALOMON : Final Report. Rapport technique, Katholieke Universiteit Leuven, 1996.

[Quicklaw,] Quicklaw. Compagnie. Canada. `http://www.quicklaw.com/`.

[Quinlan, 1989] J.R. Quinlan. Induction of Decision Trees. *Machine Learning*, pages 81–106, 1989.

[Radev *et al.*, 2002] Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Arda Çelebi, Hong Qi, Elliott Drabek et Danyu Liu. Evaluation of Text Summarization in a Cross-lingual Information Retrieval Framework. Rapport technique, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, June 2002.

[Radev *et al.*, 2003] Dragomir Radev, Jahna Otterbacher, Hong Qi et Daniel Tam. MEAD ReDUCs : Michigan at DUC 2003. In *DUC03*, Edmonton, Alberta, Canada, May 31 - June 1 2003. Association for Computational Linguistics.

[Saggion et Lapalme, 1998a] Horacio Saggion et Guy Lapalme. The Generation of Abstracts by Selective Analysis. In *Intelligent Text Summarization. Papers from the 1998 AAAI Spring Symposium. Technical Report SS-98-06*, pages 137–139, Standford (CA), USA, March 23-25 1998. The AAAI Press.

[Saggion et Lapalme, 1998b] Horacio Saggion et Guy Lapalme. Where does Information come from ? Corpus Analysis for Automatic Abstracting. In *Rencontre Internationale sur l'Extraction le Filtrage et le Résumé Automatique. RIFRA'98*, pages 72–83, Sfax, Tunisie, Novembre 11-14 1998.

[Saggion et Lapalme, 2002] Horacio Saggion et Guy Lapalme. Generating Indicative-Informative Summaries with SumUM. *Computational Linguistics*, 28(4), 2002.

[Saggion, 2001] Horacio Saggion. *Génération automatique de résumés par analyse sélective*. PhD thesis, Université de Montréal, 2001.

[Silber et McCoy, 2000] Gregory H. Silber et Kathleen McCoy. Efficient Text Summarization Using Lexical Chains. In *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI'2000)*, January 9–12 2000.

[Silber et McCoy, 2002] Gregory H. Silber et Kathleen McCoy. Efficiently Computed Lexical Chains As An Intermediate Representation in Automatic Text Summarization. *Computational Linguistics*, 28(4), 2002.

[Smith *et al.*, 1995] J. C. Smith, Daphne Gelbart, Keith MacCrimmon, Bruce Atherton, John McClean, Michelle Shinehoft et Lincoln Quintana. Artificial Intelligence and Legal Discourse : The Flexlaw Legal Text Management System. *Artificial Intelligence and Law*, 3(1-2) :55–95, 1995.

[Smith et Deedman, 1987] J. C. Smith et Cal Deedman. The Application of Expert Systems Technology to Case-Based Law. *ICAIL*, pages 84–93, 1987.

[SOQUIJ,] SOQUIJ. Société québécoise d'information juridique. Québec, Canada. `http://www.soquij.qc.ca/`.

[Spark Jones, 1999] K. Spark Jones. Automatic Summarizing : Factors and Directions. In *Advances in Automatic Text Summarization*, I. Mani et M. Maybury, éditeurs. MIT Press, Cambridge MA, 1999.

[Teufel et Moens, 1997] Simone Teufel et Marc Moens. Sentence Extraction as a Classification Task. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization, ACL/EACL*, 1997.

[Teufel et Moens, 2002] Simone Teufel et Marc Moens. Summarising Scientific Articles - Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4) :409–445, 2002.

[Teufel, 1998] Simone Teufel. Meta-Discourse Markers and Problem-Structuring in Scientific Texts. In *Proceedings of the Workshop on Discourse Relations and Discourse Markers at the 17th International Conference on Computational Linguistics*, M. Stede, L. Wanner et Eduard Hovy, éditeurs, pages 43–49, August 15 1998.

[TIDES, 2000] Translingual Information Detection, Extraction and Summarization (TIDES) Program. `http://www.darpa.mil/ito/research/tides/index.html`, August 2000.

[TREC, 2002] TREC. The 11th Text REtrieval Conference (Trec-11). `http://trec.nist.gov/`, novembre 2002.

[Uyttendaele *et al.*, 1996] C. Uyttendaele, M.-F. Moens et J. Dumortier. SALOMON : Abstracting of Legal Cases for Effective Access to Court Decisions. In *Proceedings of JURIX 96 Ninth International Conference on Legal Knowledge Based Systems*, pages 47–58. Tilburg : University Press, 1996.

[Westlaw,] Westlaw. Compagnie. U.S.A. `http://www.westlaw.com/`.