Résumé de textes juridiques par identification de leur structure thématique

Atefeh Farzindar, Guy Lapalme et Jean-Pierre Desclés

RALI, Université de Montréal
Département d'Informatique et recherche opérationnelle
C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7
farzinda@iro.umontreal.ca, lapalme@iro.umontreal.ca

LaLICC, Université de Paris4-Sorbonne
96, boulevard Raspail 75006 Paris, France
jean-pierre.descles@paris4.sorbonne.fr

RÉSUMÉ. *Cet article présente notre méthode de production automatique de résumé de textes juridiques qui permet aux juristes de consulter rapidement les idées clés d'une décision juridique pour trouver les jurisprudences pertinentes à leurs besoins. La méthodologie repose sur l'exploitation de la structure thématique afin de constituer automatiquement une fiche de résumé augmentant la cohérence et la lisibilité du résumé. La constitution de la fiche de résumé se fait en quatre étapes : la segmentation thématique qui repère la structure du document en quatre thèmes* INTRODUCTION, CONTEXTE, RAISONNEMENT JURIDIQUE *et* CONCLUSION, *le filtrage des unités moins importantes comme les citations des articles des lois, la sélection des unités textuelles saillantes dans les segments thématiques et la production du résumé dans la limite de la taille demandée. La conception des différentes composantes du système, appelé LetSum, est décrite en détail ainsi que son implémentation et le résultat d'évaluations.*

ABSTRACT. *This article present our method for dealing with automatic summarization techniques in the legal domain, which helps a legal expert determine the key ideas of a judgement in order to find relevant documents. Our approach is based on the exploration of the document's architecture and its thematic structure in order to build automatically a table style summary, which improves coherency and readability in the summary. The summary is built in four phases: thematic segmentation which identifies the structure of the document, filtering of less important units such as citations of law articles, selection of relevant textual units and production of the summary within the size limit of the abstract. We present the components of a system, called LetSum, built according to this approach, its implementation and some preliminary results of the evaluation.*

MOTS-CLÉS : *Résumé automatique, fiche de résumé, segmentation thématique, textes juridiques*

KEYWORDS: *Automatic text summarization, summary table, thematic segmentation, legals texts*

1. Introduction

Le but d'un système de résumé automatique est de produire une représentation condensée d'un document, où les informations importantes du texte original sont préservées. Dans ce contexte, il faut aussi considérer les besoins de l'utilisateur et de la tâche. Les systèmes de production de résumé peuvent être utilisés comme aide à l'organisation des connaissances ou à l'analyse et à la recherche d'informations sur le réseau Internet. De nombreuses approches pour la création automatique de résumé ont été proposées : la distribution des mots [LUH 59], les expressions prototypiques [EDM 69] [PAI 81], l'identification des chaînes lexicales [BAR 97], la rhétorique du discours [MAR 97], les étiquettes sémantiques [MIN 01] et le modèle conceptuel et linguistique de génération (*abstraction*) [SAG 01]. Plus récemment, la série de compétitions *Document Understanding Conference* [DUC 03] a été organisée pour évaluer les systèmes de résumé automatique. Mani [MAN 01, MAN 99], Marcu [MAR 00] et Minel [MIN 02] présentent des introductions à la situation actuelle dans ce domaine.

La plupart des approches de production de résumé dépendent du format et du modèle d'écriture. Par exemple, la stratégie qui consiste à prendre le premier paragraphe n'est pas universelle, elle ne donne des résultats satisfaisants que pour les articles de journaux et de magazines. Pour la production d'un résumé biographique, la position des phrases n'est pas toujours un élément prépondérant. La plupart des techniques calculent un poids pour chaque phrase individuellement, alors que des approches récentes prennent en compte les relations entre les phrases.

Nos études portent sur une nouvelle application de résumé automatique et une forme particulière de documents de type juridique : les décisions des cours judiciaires du Canada. L'objectif de ce projet est de créer automatiquement des résumés courts, qui répondent aux besoins des avocats et des experts du domaine [FAR 04a].

Notre approche est basée sur l'exploitation de la structure thématique d'un jugement. L'identification de l'organisation structurelle du document nous donne la possibilité de catégoriser les informations et de trouver les bornes de changement thématique dans le document afin d'en extraire les idées clés. Dans cet article, nous montrerons que la jurisprudence possède une structure thématique. Les différents thèmes dans le jugement découpent la suite des paragraphes en segments thématiques. Pour chacun, nous assignons un rôle rhétorique qu'il joue dans le jugement. Le rôle rhétorique indique le thème du segment. Dans cet article, nous utilisons la notion **thème** qui correspond au rôle rhétorique pour chaque segment thématique. Cette étude montre l'intérêt de l'analyse des phrases dans leur contexte, ce qui permet l'extraction des phrases les plus importantes d'après leurs thèmes [FAR 04b].

Dans les sections suivantes, nous décrivons la motivation de la recherche et les résultats de notre étude sur le corpus juridique. Nous présenterons notre méthode pour l'identification de la structure thématique des jugements afin de produire un résumé cohérent sous forme d'une fiche de résumé structurée. Nous présentons les différents modules du système LetSum (Legal text Summarizer) basé sur cette approche ainsi

que son implantation et les résultats d'évaluations partielles des fonctionnements des modules et l'évaluation globale des résumés produits.

2. Contexte du travail

Au Canada, le Centre de recherche en droit public (CRDP) a mis sur pied le projet CanLII (Canadian Legal Information Institute) qui a pour objectif de créer une bibliothèque de droit virtuelle donnant un accès internet gratuit aux décisions de tous les tribunaux judiciaires canadiens. Le nombre de jugements rendus annuellement est de plus de 200 000. 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, avec des méthodes fiables et peu coûteuses.

Mais pourquoi s'intéresse-t-on au traitement des décisions juridiques passées, ainsi qu'à leurs résumés ? D'abord parce que pour régler un problème juridique dont la solution ne se trouve pas directement dans la loi, les avocats cherchent des jurisprudences dans le passé afin de trouver les cas semblables et d'analyser les solutions adoptées par les avocats des parties et les juges. Car chaque décision tient lieu de *loi* entre les parties et en justifie la solution. Donc une jurisprudence constitue un *précédent* juridique puisqu'il est possible d'en extraire une règle de droit pouvant servir à disposer d'affaires semblables. Pour traiter un cas juridique il faut examiner des centaines de précédents 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.

Au Québec, le Répertoire électronique de jurisprudence du Barreau (REJB) et la Société québécoise d'information juridique (SOQUIJ) sont deux organismes qui fournissent des résumés manuels pour les ressources juridiques, mais le temps et l'expertise nécessaires augmentent le coût d'utilisation de leurs services. Par exemple le prix d'un résumé de SOQUIJ et son texte intégral (Mai 2004) est 7,50$ can. , alors que pour traiter un seul cas il faut souvent examiner des centaines de jugements comme précédents juridiques. Une requête sur un sujet, par exemple les langues officielles du Canada, dans la base de données de CanLII sur la collection des décisions de la Cour fédérale retourne 500 jugements comme précédents juridiques, ce qui montre l'intérêt de la production automatique de résumé. Certains systèmes d'informations juridiques ont été développés par des compagnies privées, QuickLaw au Canada, WESTLAW et LEXIS aux États Unis, toutefois aucun ne satisfait complètement les exigences spécifiques de ce domaine d'application.

Une raison de la difficulté de ce travail est la complexité du domaine : des vocabulaires spécifiques au domaine juridique et des interprétations d'expressions judiciaires qui peuvent amener des ambiguïtés ; par exemple le mot ***sentence*** en anglais peut avoir deux sens différents : l'un est la phrase et l'autre est, dans un sens plus particulier du domaine, la condamnation ou la peine. Un autre exemple est le mot **dispositif**, le sens

courant de ce mot est le mécanisme et la manière donc sont disposées les pièces d'un appareil alors que son sens en droit signifie la partie terminale d'un jugement dans laquelle est indiqué ce qui est décidé.

Les expériences des systèmes de résumé automatique actuels ont été essentiellement limités à l'étude des articles des journaux et à quelques systèmes sur les articles scientifiques [SAG 02]. Il y a des différences importantes entre le langage journalistique et le langage juridique : statistique des mots, mots de titre et les relations de chaîne lexicale entre les mots du titre avec les idées clés du texte, relations entre une phrase avec sa précédente et la suivante, il est de même pour les paragraphes et la structuration du texte. Pour les jugements, nous montrons dans cet article qu'il est possible de définir les structures discursives pour les différentes parties de la décision et de leur assigner des thèmes. Les articles des journaux répètent souvent le message le plus important mais, en droit, rarement plus d'une seule fois. Le traitement des textes juridiques demande une attention particulière et il n'est pas facile d'adapter les techniques déjà développées pour d'autres types de textes.

3. Corpus d'étude

3.1. *Composition*

Notre corpus est composé de 3500 documents de jurisprudences en anglais, rendues par la Cour fédérale du Canada du tribunal de première instance des années 2000 à 2003, qui sont disponibles en format HTML sur le site CanLII à l'adresse suivante `http:/ / www.canlii.org`. Nous avons analysé 50 jugements et leur résumés rédigés manuellement par un arrêtiste, un résumeur professionnel. Un nombre limité de résumés des décisions publiées de la Cour fédérale sont disponibles sous le nom des fiches analytiques sur le site de `http://reports.fja.gc.ca`. Nous avons aussi étudié deux collections de documents, en français, regroupant 15 jugements et leurs résumés, produits par le Répertoire électronique de jurisprudence du Barreau (REJB), et le Recueil des arrêts de la Cour fédérale. Contrairement au projet SALOMON [UYT 96] qui portait sur des cas criminels, notre recherche couvre différentes catégories de jurisprudences comme : *Access to information, Administrative law, Air law, Broadcasting, Competition, Constitutional law, Copyright, Customs and Excise - Customs Act, Environment, Evidence, Human rights, Maritime law, Official languages, Penitentiaries, Unemployment insurance*, etc.

Afin d'estimer la taille moyenne des jugements comme entrée de notre système, nous avons calculé la distribution des mots des 3500 documents de notre corpus. Nous observons que 75% des décisions ont entre 500 et 4000 mots (2 à 8 pages). Donc dans notre travail, nous concentrons nos travaux sur les textes dans cet intervalle de longueur ; 10% des documents ont moins de 500 mots, environ une page, et il n'est donc pas nécessaire de résumer alors que seulement 15% des documents comportent plus de 4000 mots.

Dans notre corpus, la taille moyenne des jugements est 3600 mots et la taille moyenne des résumés est 360 mots, c'est à dire un taux de réduction de 10%.

3.2. *Structuration des textes juridiques*

Dans certains domaines spécifiques comme le domaine juridique, le contexte peut changer la valeur sémantique d'une phrase. Souvent pour interpréter une phrase, il faut considérer sa place dans le fragment textuel et la situation dans laquelle cette phrase est employée. Pour la tâche d'extraction des phrases pertinentes de décisions de justice, il est aussi important de présenter les circonstances dans lesquelles les phrases ont été prononcées. Par exemple pour obtenir une interprétation correcte de la phrase *The application is dismissed*, il est nécessaire de prendre en compte son contexte : si cette phrase apparaît dans la partie du texte qui explique les histoires judiciaires sur le cas, elle signifiera une demande de révision sur une décision précédemment rendue par une autre cour de justice, mais si cette phrase apparaît dans la partie terminale du jugement, elle exprimera la décision finale du jugement en cours.

Pour aborder ce problème, nous avons étudié l'organisation générale de textes de jurisprudences. D'après nos analyses, les jugements sont organisés selon une macro-structure qui contient différents niveaux d'informations, indépendamment du domaine de jugement. Les travaux expérimentaux de la juge Mailhot de la Cour d'appel du Québec [MAI 96], utilisés pour guider les juges à écrire un jugement, renforcent cette idée qu'il est possible de définir une structure organisationnelle pour ces décisions. Les jurisprudences sont organisées par le discours lui-même, ce qui permet de les segmenter en s'appuyant sur l'organisation discursive des contenus. L'ensemble des unités textuelles qui traitent du même sujet forme un segment thématique. Pour chaque thème, il existe une borne qui signale un changement dans le discours ou dans le thème du segment. Pour déterminer les bornes de segments thématiques, nous avons étudié plusieurs éléments indicateurs comme les titres significatifs des sections, les positions des fragments textuels et les expressions linguistiques que nous expliquerons en détail dans la section 6.1.2. À l'intérieur d'un segment thématique les phrases portent sur un même sujet et chaque phrase est influencée par ses voisines.

Pour notre analyse de corpus, nous avons comparé les résumés modèles créés par un humain avec les textes des jugements originaux. Nous avons identifié la structure organisationnelle pour le jugement. Les paragraphes qui traitent d'un sujet sont considérés comme membres d'un groupe thématique. Nous avons annoté les segments avec une étiquette indiquant leurs thèmes. Nous avons aussi manuellement annoté les unités de citation, des unités textuelles (phrase ou paragraphe) données par le juge comme référence (par exemple à un article de loi). Les segments de citation occupent une taille considérable dans le jugement, mais ils ne sont pas considérés importants dans le résumé, donc ces segments seront éliminés lors des filtrages d'information.

Les unités textuelles considérées importantes par les résumeurs professionnels ont été alignées manuellement avec un ou plusieurs éléments du texte source. Le Tableau 1

Décisions de la Cour fédérale du Canada entre :
GUIDES LTD., et JASPER NATIONAL PARK PROFESSIONAL RIVER OUTFITTERS ASSOCIATION, demandeurs et
LE PROCUREUR GÉNÉRAL DU CANADA, défendeur

Texte intégral Dossier : T-1557-98	Résumé de l'arrêtiste	Thème de section
[1] This application for judicial review arises out of a decision (the Decision) announced on or about the 30th of June 1998 by the Minister of Canadian Heritage (the Minister) to close the Maligne River (the River) in Jasper National Park to all boating activity, beginning in 1999.	Judicial review of Minister of Canadian Heritage's decision to close Maligne River in Jasper National Park to all boating activity beginning in 1999 to protect habitat of harlequin ducks-	Introduction
[7] The applicants offer commercial rafting trips to Park visitors in this area each year from mid-June to sometime in September.	Applicants offer commercial rafting trips on River.	Faits
[10] Consequently, a further environmental assessment regarding commercial rafting on the Maligne River was prepared in 1991. The assessment indicated that rafting activity had expanded since 1986, with an adverse impact on Harlequin ducks along the Maligne River.	1991 environmental assessment indicating rafting having adverse impact on harlequin ducks along river.	Faits

Tableau 1. *Alignement des unités textuelles du texte original du jugement avec les unités du résumé rédigé manuellement.*

montre un exemple de cet alignement entre un résumé humain et le texte intégral original. Nous y cherchions les relations entre les informations considérées importantes dans les résumés des arrêtistes et les informations dans les documents sources. Nous avons constaté que les arrêtistes produisent un résumé par extraction des unités importantes tout en suivant des lignes directrices du texte. L'assemblage de ces unités sélectionnées constitue le résumé.

L'identification des structures thématiques sépare les idées clés des détails secondaires d'un jugement et améliore la lisibilité de résumé en produisant des textes plus cohérents. Notre hypothèse est que, malgré la variabilité des catégories des jugements, on peut distinguer une structure sur les informations présentées dans une jurispru-

dence. Afin de repérer les thèmes du jugement, nous avons développé un segmenteur thématique basé sur des connaissances linguistiques et juridiques. Nous en dégageons un plan d'organisation thématique dans lequel les unités composant le discours prennent place au fur et à mesure de leur apparition dans le jugement. Nous allons présenter la fonctionnalité des thèmes et leur importance dans le jugement du point de vue des idées clés et principales. le Tableau 2 montre la structuration d'une jurisprudence et ses thèmes : DONNÉES DE LA DÉCISION, INTRODUCTION, CONTEXTE, RAISONNEMENT JURIDIQUE et CONCLUSION. La présentation du résumé final respectera cette organisation afin de constituer une fiche de résumé de décisions en cinq thèmes :

DONNÉES DE LA DÉCISION présentent la référence complète de la décision et la relation entre les parties sur le plan juridique : nom de la cour de décision, lieu de l'audience, date du jugement, numéro de greffe, référence neutre, identification des parties, intitulé du jugement, autorités et doctrines citées.

INTRODUCTION contient les paragraphes explicatifs placés en tête du jugement qui présentent le sujet. Ils décrivent brièvement la situation qui se présente au tribunal et répondent aux questions **qui ? a fait quoi ? à qui ?**. Il peut arriver que les **questions de droit** (issues), qui identifient le problème juridique dont le tribunal est saisi, viennent directement après l'introduction.

CONTEXTE contient les faits et l'histoire judiciaire et l'ensemble des circonstances dans lesquelles s'insèrent des faits. Il recompose l'histoire du litige à partir des faits et des événements relatés lors de la présentation de la preuve et retenus dans le jugement.

RAISONNEMENT JURIDIQUE discute à partir des contextes et des faits du litige en s'appuyant sur des références juridiques et des autorités afin d'en arriver à une conclusion. Il répond aux questions de droit et comporte une explication sur les motifs du juge. Cette partie est la plus importante du résumé d'une décision, puisqu'elle contient la justification de la décision finale de la cour et 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 deviennent règle de droit.

CONCLUSION est la dernière partie du procès qui fait connaître la décision du juge. Elle prononce le **dispositif**, la partie terminale du jugement dans laquelle est indiqué ce qui est décidé et les montants adjugés s'ils existent. Par exemple en droit pénal, il faut spécifier si la personne a été condamnée ou acquittée.

Dans notre corpus, nous avons identifié cette organisation textuelle sous forme de segments thématiques. En suivant les thèmes, le lecteur est guidée de la définition du problème en litige jusqu'à ce que la cour a retenu comme le résultat, qui peut soit arriver à une solution aux problèmes entre les deux parties, soit renvoyer le dossier à une autre cour pour compléter le processus juridique.

Structure thématique	Explications
DONNÉES DE LA DÉCISION	Nom de la cour de décision Lieu de l'audience Date du jugement Numéro de greffe Référence neutre Identification des parties Intitulé du jugement Autorités et doctrines citées
INTRODUCTION	Qui ? A fait quoi ? À qui ?
CONTEXTE	Faits recompose l'histoire du litige Histoire judiciaire
RAISONNEMENT JURIDIQUE	Analyse du juge et détermination des faits Expression des motifs de la solution retenue
CONCLUSION	Décision finale de la cour

Tableau 2. *Structure thématique d'une jurisprudence*

4. D'où viennent les informations

Au cours de notre analyse de corpus de résumés alignés avec les textes de jurisprudences, nous avons mesuré la distribution de l'information (en nombre de mots) pour chaque thème de jugement. Les volumes occupés par ces champs dans le document source et le résumé sont données au Tableau 3. On y voit l'importance et la contribution de chaque thème de jurisprudence dans le résumé, ce qui nous servira à attribuer un score comme valeurs sémantiques des thèmes. Ces valeurs sémantiques augmenteront ou diminueront la chance des phrases candidates d'être sélectionnées dans le résumé final. Par exemple, à l'étape finale de contrôle de la taille de résumé, nous préférerons plutôt une phrase du segment RAISONNEMENT JURIDIQUE qu'une du segment CONTEXTE, car le premier est jugé plus important à cause de sa valeur sémantique plus élevée.

Un jugement fait souvent référencer à d'autres jurisprudences en les citant. Ces citations peuvent être placées en plusieurs segments thématiques selon les cas. Les phrases citées n'expriment pas les idées clés des jugements, mais les résultats retenus de ces citations sont importants. Habituellement, les citations contiennent deux phases, la première comporte les prétentions et arguments des parties, qui présentent le point de vue d'une partie sur le problème. Cette partie contient les soumissions des parties et leurs positions dans le litige que nous avons considéré comme faisant partie du thème CONTEXTE. La seconde phase de citation est la citation des articles de lois, ce qui traite des prétentions en droits applicables sur le cas. Ces segments cités sont les sections ou les paragraphes des lois, les jurisprudences ou les doctrines que le juge utilise comme références dans son raisonnement. Dans plupart des cas, les citations sont considérées comme étant du thème RAISONNEMENT JURIDIQUE. Dans

Structures thématiques	Jugement	Résumé
INTRODUCTION	5%	12%
CONTEXTE	24%	20%
RAISONNEMENT JURIDIQUE	67%	60%
CONCLUSION	4%	8%
Total	100%	100%

Tableau 3. *Les pourcentages de la contribution de chaque thème dans les résumés et les jugements originaux*

notre système, nous identifions les phrases citées, pour en identifier les résultats afin de les garder dans la liste des unités candidates à extraire.

5. Dictionnaire de connaissances linguistiques

Nous avons construit un dictionnaire de connaissances linguistiques et de vocabulaire spécifique au domaine juridique. Le dictionnaire contient 250 marqueurs linguistiques significatifs que nous avons observés dans le corpus. Ce dictionnaire est organisé en trois catégories. La première contient des marqueurs communs à plusieurs types de texte et elle est divisée en trois classes : les verbes, les concepts (noms, adjectifs et adverbes) et des indices complémentaires comme les marqueurs typographiques (ex. deux-points dans citation), prépositions (ex. ***at*** *page*), numéros (ex. *subsection* ***20***), et subordonnés relatives (ex. *that*). La deuxième catégorie de marqueurs contient les expressions prototypiques (*cue-phrases*) qui expriment les parties importantes de discours (ex. *finally*). La troisième catégorie contient les vocabulaires et les expressions spécifiques au domaine juridique qui permettent de désambiguïser les termes avec différents sens dans un dictionnaire général mais qui ont, en droit un sens particulier, par exemple la différence entre un *appel* téléphonique et un jugement porté en *appel*.

6. Méthode de la constitution automatique de fiches de résumé

Notre étude de corpus, basée sur une analyse des phrases des résumés modèles alignées manuellement avec les jugements sources, fait ressortir la structure thématique d'un jugement. Notre approche de résumé automatique est fondée sur cette analyse. Elle regroupe les paragraphes qui traitent le même sujet, ce qui conduit à détecter les frontières entre les différents thèmes du document. Le découpage du document divise le texte en segments thématiques. Pour chacun, nous attribuons un thème comme étant le sujet traité par les paragraphes du segment. D'après le thème du segment, nous extrairons des phrases saillantes qui contiennent les informations pertinentes sur ce thème. Mais avant l'étape de sélection, un filtrage des segments cités (jugés non im-

portants) diminuera la quantité des textes à analyser. L'ensemble des phrases extraites constitue le résumé.

Dans les sections suivantes, nous expliquons en détail notre méthode de production de résumé automatique qui repose sur l'identification de la structure thématique du jugement, en utilisant la technique d'extraction des unités saillantes, avec une présentation du résumé final sous forme d'une fiche contenant des rubriques homogènes d'informations. Cette fiche permet de présenter les informations considérées importantes associées à des thèmes précis, ce qui en facilite la lecture et la navigation entre le résumé et le jugement source. Pour chaque phrase du résumé produit, l'utilisateur peut en déterminer le sujet en regardant le thème associé à son segment thématique. Si une phrase semble plus importante pour l'utilisateur et qu'il désire plus d'information sur ce sujet, on peut lui proposer le segment thématique entier contenant la phrase sélectionnée, pour obtenir les informations complémentaires sur le sujet. Dans notre système, il y a d'abord un prétraitement du texte, ce qui implique la séparation du document en unités textuelles (paragraphes, phrases, mots, nombres et ponctuations) et modules linguistiques (voir section 7). La constitution de la fiche de résumé se fait en quatre étapes (Figure 1) : segmentation thématique, filtrage des unités moins importantes comme les citations des articles des lois, sélection des unités textuelles candidates pour le résumé et production du résumé selon la taille demandée. L'étape segmentation repère le squelette structurel du texte. Chaque niveau de cette structure a un thème particulier dans le jugement. Selon le thème, nous cherchons à identifier les unités pertinentes dans les segments correspondants, tout en respectant la limite de la taille du résumé.

6.1. *Segmentation en thèmes*

Dans la première étape, nous nous intéressons à déterminer l'organisation du document original et la segmentation thématique ayant pour objet le découpage des textes en segments thématiquement homogènes.

6.1.1. *Expérimentation*

Pour la segmentation thématique, nous avons fait quelques expérimentations avec deux segmenteurs décrits par Hearst [HEA 94] le système TexTiling et le segmenteur C99 décrit par Choi [CHO 00]. Ces deux segmenteurs statistiques utilisent une fonction de *clustering* pour diviser le document par thème. Mais les résultats de ces segmenteurs numériques n'étaient pas satisfaisants pour déterminer les structures thématiques des jugements. TexTiling nous donnait un segment thématique par paragraphe et C99 ne nous donnait que quelques découpages entre les paragraphes et les citations. Nous avons donc procédé à une segmentation thématique basée sur des connaissances spécifiques en domaine juridique. Pour y arriver, il faut identifier les bornes de la structure thématique, pour encadrer les segments du texte associés avec un thème précis dans la jurisprudence.

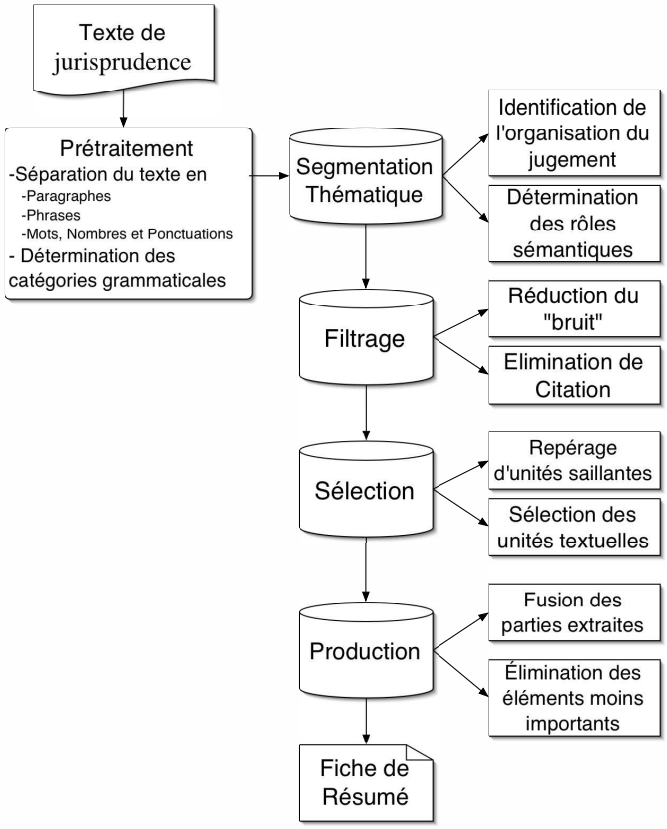

Figure 1. *Les étapes de la constitution de la fiche de résumé*

L'exploitation de la structuration du jugement peut indiquer les citations dans les textes. La détection des citations est importante pour la prochaine étape de filtrage d'information. Afin d'identifier les citations dans les jugements, nous avons développé un module qui identifie les marqueurs directs et indirects de la citation. Nous décrirons ce processus dans la section 6.2.

6.1.2. *Module de segmentation sémantique du thème*

Notre segmenteur thématique est fondé sur les connaissances sémantiques du jugement. Nous avons identifié les indicateurs linguistiques et les vocabulaires réservés du domaine qui signalent les changements thématiques. Pour déterminer les thèmes des segments, les marqueurs les plus importants sont : les indicateurs linguistiques qui se trouvent dans les titres des sections, les positions des paragraphes et les marqueurs linguistiques et les expressions prototypiques du domaine. Pour la segmentation, nous

Catégories des titres	Marqueurs linguistiques	Exemples des titres
INTRODUCTION	*introduction, summary*	*Introduction, Summary*
CONTEXTE	*facts, background*	*The factual background, Agreed statement of facts*
RAISONNEMENT JURIDIQUE	*analysis, decision, discussion*	*Analysis and Decision of the court*
CONCLUSION	*conclusion, disposition, cost*	*Conclusion and Costs*

Tableau 4. *Les marqueurs significatifs qui se trouvent dans les titres des sections*

avons écrit 200 règles sémantiques qui contrôlent la présence des marqueurs linguistiques en considérant les contextes dans lesquels les termes sont utilisés.

Nous avons identifié les titres plus significatifs pour les sections qui séparent les différentes étapes du jugement. Le Tableau 4 montre des exemples de marqueurs identifiés dans les titres et les exemples des titres observés dans les jugements classés selon leur thème dans le document.

Les DONNÉES DE LA DÉCISION peuvent être extraites par les informations présentées dans la première page couverture du jugement. Dans le champ *autorités et doctrines citées*, nous gardons les références et les renvois juridiques cités dans le document. Cette section est séparée au texte de jugement par un titre comme *Reasons for order* ou *Reasons for judgment*. Pour la majorité des jugements des cours canadiennes, le juge commence avec une introduction sur le sujet qui constitue les premiers paragraphes donnant en général des informations importantes sur l'affaire, le thème attribué à ce segment est INTRODUCTION. Après l'INTRODUCTION, nous trouvons les informations sur les faits et l'histoire du litige, ce segment est associé au thème CONTEXTE. Les titres de la section du CONTEXTE peuvent par exemple contenir *The factual background* ou *Agreed statement of facts*. CONTEXTE contient deux niveaux de discours : premier concerne la présentation des parties et leur demande, le deuxième comprend la partie narrative du jugement dans laquelle juge raconte les événements et les faits. Pour la présentation des parties, les marqueurs sont : *appellant*, *applicant*, *defendant*, *plaintiff*, *respondent* suivis de leur identification (ex. *the applicant is company X*). Pour la demande, les verbes suivants sont utilisés *advise*, *indicate*, *request*, etc.

Afin d'augmenter la précision de l'identification des parties, nous avons défini certaines règles qui tiennent compte de la sémantique. Par exemple, si la phrase suivante est placée dans le segment CONTEXTE, le patron *the respondent is a* n'est pas une présentation de partie : *On February 14, 1996, the Minister issued, pursuant to subsection 70(5) of the Immigration Act , an opinion that* ***the respondent is a*** *"danger to the public"*. Les règles sémantiques vérifient donc les points suivants : est-ce que cette phrase reflète une opinion ou un point de vue spécifique ? est-ce que cette phrase reflète une définition dans un cas particulier (*the person constitutes a danger to the public in Canada is ...*) ? est-ce une citation ?

Dans la partie narration du discours, on peut extraire ce qui s'est passé entre les deux parties et l'histoire du problème en litige. Le texte narratif du thème CONTEXTE peut être identifié grâce aux caractéristiques suivantes :

– présence d'une suite d'événements qui font partie d'un fait
– présence des dates complètes des événements
– présence de liens logiques de temps (*then*, *following day*, ...)
– présence de temps (en général au passé)
– des verbes d'action

Dans certains jugements, plusieurs paragraphes, avec un titre comme *Respondent's position* ou *Applicant's submissions*, expliquent la position d'une partie devant le juge. La section ISSUES ou questions en litige comprend des questions de fait ou de droit auxquelles le juge doit répondre afin de déterminer le résultat du litige. Cette section a des titres comme *The issues, The questions of law*. Selon le cas, le juge peut avoir à répondre à une ou plusieurs questions, mais il arrive que ces questions ne soient pas expressément posées par le juge dans ses motifs. Ces questions de droit sont importantes parce qu'elles précisent le but du raisonnement du juge. Ces deux dernières sections précèdent la partie RAISONNEMENT JURIDIQUE et nous les avons considérées comme une borne entre deux thèmes CONTEXTE et RAISONNEMENT JURIDIQUE.

Après avoir exposé les informations historiques sur le sujet, le juge commence le RAISONNEMENT JURIDIQUE, la partie la plus importante dans une décision, qui amène la lecture du CONTEXTE jusqu'à la conclusion finale de la cour. Ce segment peut avoir un titre comme *Analysis* ou *Discussion*. Comme mentionné dans la partie CONTEXTE, dans la plupart des jugements, deux styles différents sont présents dans la décision selon leur position dans le texte. Dans les sections précédant le RAISONNEMENT JURIDIQUE, le juge s'exprime avec un style narratif pour décrire les faits résultant du litige ; c'est le juge qui analyse les lois et les précédents judiciaires pour arriver à une conclusion. Donc dans le discours du thème RAISONNEMENT JURIDIQUE le juge s'exprime par *I*. Les exemples des expressions utilisées dans le RAISONNEMENT JURIDIQUE sont : *In reviewing the sections No. of the Act*, *Pursuant to section No.*, *As I have stated*, *In the present case*, *The case at bar is*, *I am of the view that*, etc.

La dernière section du jugement exprime la CONCLUSION où le juge prononce le **dispositif** dans laquelle est indiqué ce qui est décidé. Habituellement, les phrases contenant le dispositif apparaissent dans les derniers paragraphes du jugement et leur forme est au passif. La présence des marqueurs linguistiques afin de déterminer les résultats du jugement est importante. Quelques exemples de phrases donnant le dispositif sont : *The motion is dismissed, the application must be granted*. Nous avons identifié une liste des marqueurs de classe des verbes qui expriment la décision de la jurisprudence : *allow*, *deny*, *dismiss*, *grant*, *refuse*, *award*, etc.

Dans un jugement, on peut avoir une section contenant un paragraphe comme l'ordonnance (*Order*), avec une ou deux phrases qui décrivent le résultat très bref pro-

Intitulé du jugement : l'accès d'information, Numéro de greffe : T-1819-98
In support of its application the plaintiff maintained that the Department had erred in fact and in law when it decided that the plant inspection reports could be disclosed to Mr. D., since the tests for exceptions to the disclosure of documents contained in s. 20(1)(c) and (d) of the Act had not been met. **Section 20(1)** of the Act **reads as follows :**
20. (1) Subject to this section, the head of a government institution shall refuse to disclose any record requested under this Act that contains : (a) trade secrets of a third party. (b) financial, commercial, scientific or technical information that is confidential information supplied to a government institution by a third party and is treated consistently in a confidential manner by the third party. (c) information the disclosure of which could reasonably be expected to result in material financial loss or gain to, or could reasonably be expected to prejudice the competitive position of, a third party. (d) information the disclosure of which could reasonably be expected to interfere with contractual or other negotiations of a third party.

Tableau 5. *Exemple d'une citation : la ligne entre les deux paragraphes montre un changement thématique qui signale la présence de la citation. Les mots en gras sont les marqueurs de citation. Cette unité pointe vers le bloc énuméré, qu'il va aussi considérer comme une citation parce que les phrases de ce bloc sont liées avec des intégrations linéaires à la première unité contenant des marqueurs de citation.*

noncé par la cour. Cette section peut être placée au début ou à la fin du jugement après la conclusion.

Nous avons considéré la signature du juge comme la fin du jugement avec les marqueurs : le nom ou initiales du juge, son titre (*Judge*), la date et le lieu de l'audience. Par la suite, le jugement peut être complété avec les informations sur les références complètes de tous les articles de lois cités, les doctrines citées, etc., qui ne sont pas pertinents pour le résumé.

6.2. *Filtrage*

L'objectif du filtrage est d'identifier les *exemples négatifs* qui peuvent être supprimés dans les documents sans perdre les informations pertinentes pour le résumé et les *exemples positifs* qui augmentent la chance d'une unité textuelle candidate d'apparaître dans le résumé final. Dans les cas de citation, comme dans l'exemple du Tableau 5, la présence des marqueurs signale un changement thématique de la citation. Cette méthode permet de réduire la quantité de texte à analyser.

Classe des verbes	Classe des concepts	Classe des indices
conclude, define, indicate, provide, read, reference, refer, say, state, summarize, ...	*following, section, subsection, page, paragraph, pursuant, ...*	Ponctuation : deux-points et guillemet ; Prépositions : *at* (*page*), (*pursuant*) *to* ; Numéros : (section, page, article de loi) ; Subordonnées relatives (conjunction) : *that, as, if*.

Tableau 6. *Les marqueurs linguistiques de citation*

Dans un jugement, les citations occupent un volume important dans le texte soit environ 30% du jugement, alors que leur contenu est moins important pour le résumé, donc nous les considérons comme des exemples négatifs. Pour cette raison, à l'intérieur des segments thématiques, nous identifions les citations à supprimer. Les citations comprennent deux catégories : la première, les *prétentions et arguments des parties*, concernent les points de vues des parties sur le litige ; la seconde, les *prétentions en droit* concernent les citations des articles de lois applicables sur l'affaire. Comme exemples positifs, nous identifions les résultats des cas de citations à conserver dans les unités textuelles candidates pour le résumé.

L'identification de citation est basée sur deux types de marqueurs : les marqueurs directs et indirects. La première catégorie de marqueur direct contient les indicateurs linguistiques. Le Tableau 6 montre les marqueurs linguistiques identifiés en trois classes : verbes de citation, concepts (nom, adverbe, adjectif) et indices complémentaires. Les verbes que nous avons identifiés en anglais, concordent avec les travaux linguistiques sur l'identification des verbes de citation par Mourad [MOU 00, MOU 03] effectués sur des textes en français, des types des articles des journaux et des rapports scientifiques.

La deuxième catégorie comprend les marqueurs de citation indirects. Les unités textuelles citées indirectement sont les unités voisines des phrases citées directement. Dans le Tableau 5, les marqueurs de citation directes sont les mots écrits en gras "***Section 20(1)*** *of the Act* ***reads as follows :***", mais les unités textuelles qui suivent cette phrase sont aussi les unités citées. Nous avons donc intégré un mécanisme d'identification des intégrations linéaires entre les phrases suivantes de la première phrase citée avec les marqueurs directs. Afin de trouver les bornes des segments de citation, il est important de préciser le commencement et la fin de citation. Pour chaque unité marquée comme citation, nous avons considéré un voisinage qui contient les unités textuelles (phrase ou paragraphe) situées avant ou après de l'unité marquée citation, avec la probabilité qu'ils soient une citation. Nous avons développé des règles sémantiques contrôlant le voisinage de l'unité citée et qui seront décrites à la section 7 sur l'implémentation du système.

Thèmes	**Marqueurs linguistiques**
INTRODUCTION	Concepts : *decision*, *motion*, *application*, *determination*, *order*, *proceeding*, *against*, *rule*, *reason* Verbes : *raise*, *strike*, *determine*, *describ*, *declare*, *date* Expressions : *application for judicial review, application to review a decision, motion filed by, Statement of Claim*
CONTEXTE	Concepts : Les parties (*appellant*, *applicant*, *defendant*, *plaintiff*, *respondent*) Verbes : *advise*, *indicate*, *request*, *claim*, *allege*, *concern*
RAISONNEMENT JURIDIQUE	Concepts : *opinion, conclusion, summary, because, cost, action, view, reason*, *I*, *my* Verbes : *note, accept, summarise, scrutinize, think, say, satisfy, discus, conclude, find, believe, reach, indicate, persuade, agree, see, review, recommend, answer* Expressions : *in the case at bar, for all the above reasons, in my view, my review of, in view of the evidence, in the result, finally, thus, consequently*
CONCLUSION	Concepts : *motion*, *application*, *action*, les parties Verbes : *allow*, *deny*, *dismiss*, *grant*, *refuse*, *award* Expressions : *for (all) the above reasons*, *for all of the foregoing reasons*

Tableau 7. *Les marqueurs linguistiques utilisés dans le module de sélection des unités textuelles qui signalent les phrases importantes dans le jugement*

6.3. *Sélection des unités textuelles*

La prochaine étape de traitement des segments est la sélection des unités textuelles afin de construire une liste d'unités saillantes candidates pour chaque segment thématique du résumé. Il reste à déterminer quelle information doit absolument se retrouver dans le résumé. Le Tableau 7 montre les marqueurs linguistiques utilisés dans ce module de sélection des unités textuelles qui signalent les phrases importantes dans le jugement. Les marqueurs sont classifiés d'après leurs thèmes dans le jugement. Les marqueurs de l'INTRODUCTION identifient les unités qui signalent le problème et éventuellement les références juridiques utilisées durant le raisonnement. Les marqueurs du CONTEXTE reconnaissent les fragments textuels qui présentent les deux parties et le problème en litige, mais les points de vue de parties ne sont pas pertinents pour le résumé. Les marqueurs du RAISONNEMENT JURIDIQUE déterminent les fragments textuels concernant les analyses du juge et la conclusion retenue par ces analyses. Les marqueurs de la CONCLUSION indiquent les phrases du dispositif qui décrit la décision finale du juge.

Afin de pondérer les unités, nous avons attribué un poids P, une valeur entre 0 et 10 à chaque unité textuelle en nous basant sur des fonctions heuristiques. La somme

des poids permet de calculer un score pour chaque phrase. Pour développer le module de sélection et attribuer les poids, nous avons pris en compte les points suivants :

– La position des phrases dans le paragraphe ; pour la première phrase du paragraphe, la valeur de P est égale à 3, pour la dernière phrase (pour les paragraphes plus d'une phrase) P est égale à 2.

– La position des paragraphes dans le segment thématique : les deux premiers paragraphes dans le segment ont des valeurs de P respectivement égales à 2 et 1. Le poids du dernier paragraphe du segment (pour les segments plus de deux paragraphes) est 1.

– Les valeurs des marqueurs linguistiques : pour les phrases qui contiennent des marqueurs linguistiques (classe des verbe-concept-indice, expressions prototypiques et vocabulaires contrôlés du domaine), nous avons assigné des poids en fonction de l'importance des marqueurs. La valeur de P pour la phrase est entre 7 et 10.

Nous avons utilisé certains modules de calcul statistique pour les termes individuels et les termes dans le contexte du document. Un module calcule les fréquences des mots et un autre calcule $\sum tf * idf$ pour chaque phrase où tf désigne *term frequency* (la fréquence du terme dans le document) et idf *inverted document frequency* qui mesure si ce terme est discriminant ou non-uniformément distribué dans le corpus. Un terme qui a une valeur de $tf * idf$ élevée doit être à la fois important dans ce document et apparaître rarement dans les autres documents. C'est le cas quand un terme correspond à une caractéristique importante et unique d'un document qui peut exprimer un topique de document. Le résultat de cette étape est une liste des phrases ayant les poids les plus élevés et qui occupent environ 30% de la taille de document source.

6.4. *Production du résumé*

Une fois sélectionnées les unités candidates potentielles pour le résumé, cette étape choisit les unités pour le résumé final et les combine afin de produire un résumé d'environ 10% du jugement. Le critère de sélection des unités est basé sur la pondération du segment thématique contenant les unités candidates. Selon nos analyses de corpus présentées dans le Tableau 3, la distribution de l'information dans les résumés des arrêtistes donne la possibilité de mesurer de l'importance des segments thématiques.

Lors de cette étape de sélection de la liste des unités candidates, nous choisissons les unités du segment thématique INTRODUCTION avec les scores plus élevés jusqu'à concurrence de 10% de la taille de résumé. Dans le segment CONTEXTE, les unités sélectionnées occupent 24% de la longueur du résumé. La contribution du segment RAISONNEMENT JURIDIQUE est de 60% et les unités avec le thème CONCLUSION occupent 6% du résumé.

7. Implémentation

L'implantation de cette approche de production automatique de résumé est un système appelé LetSum, développé en Java et Perl. L'entrée du système est un document de jurisprudence qui peut avoir un des formats XML, HTML, SGML, RTF ou un texte sans balise. Pour analyser le document, LetSum commence par le prétraitement du document, le texte est divisé en paragraphes, phrases et en unités plus petites comme les mots, les nombres et les ponctuations. L'analyseur syntaxique utilisé, pour déterminer les catégories grammaticales des mots est le modèle décrit par Hepple [HEP 00]. Les règles et les grammaires sémantiques sont écrites en langage JAPE (Java Annotations Pattern Engine) qui peuvent être exécutés avec le transducer de GATE [CUN 02]. GATE offre la possibilité d'extraction de certaines entités nommées (comme noms des personnes, dates, lieux, etc.) et des coréférences [SAG 03].

L'implémentation de LetSum, selon le modèle de conception de la Figure 1 comprend les étapes suivantes :

1) Segmentation thématique : nous avons utilisé les conditions suivantes afin de découper un segment par thème. La première condition satisfaite arrête le traitement.

- Présence des titres des sections et classement parmi les catégories des titres (Tableau 4) : une section avec un titre significatif va segmenter thématiquement avec le thème identifié par son titre.

- Position absolue d'un segment : le premier paragraphe est une introduction et les deux derniers paragraphes sont la conclusion.

- Position relative d'un segment (selon le résultat de notre étude de corpus, les segments thématiques identifiés dans le jugement sont linéairement ordonnés) : si deux sections sont thématiquement identifiées et qu'une section entre ces deux segments n'a pas encore une étiquette de thème, alors la position de cette section est un marqueur.

- Présence des marqueurs linguistiques : LetSum calcule le nombre de marqueurs de segmentation pour les paragraphes. Si une section a plus de marqueurs de même catégorie, alors cette section va être étiquetée par rapport à la catégorie des marqueurs qu'elle contient.

- Identification des styles narratif et direct (indicateur de la borne des segments CONTEXTE et RAISONNEMENT JURIDIQUE) : détection de deux styles typographiques grâce aux temps et aspect des groupe verbaux.

2) Filtrage de citation qui comporte quatre étapes :

- Identification des marqueurs et des patrons d'unités textuelles ;

- Extraction de la phrase contenant de ces unités ;

- Détection des marqueurs d'intégration linéaire : LetSum identifie d'abord les phrases contenant les unités énumérées, ensuite il regroupe ces phrases dans un bloc. Le système vérifie la condition de relation de citation entre l'unité annotée citation avec le bloc énuméré. Si l'unité de citation pointe vers le bloc énuméré alors ce bloc sera aussi considéré comme une citation.

- Contrôle de la citation pour les voisinages (avant ou après) des unités citées dans le cas d'absence des marqueurs d'énumération.

3) Sélection des unités saillantes : LetSum calcule un poids pour chaque phrase dans le jugement d'après les fonctions heuristiques expliquées à la section 6 basées sur les informations suivantes : la position des paragraphes dans le document, la position des phrases dans le paragraphe, les marqueurs linguistiques, les *cue-phrases*, les vocabulaires contrôlés du domaine juridique. Les modules statistiques du système calculent les fréquences des mots et calculent $\sum tf * idf$ pour chaque phrase. Nous avons calculé *idf* (*inverted document frequency*) en utilisant la collection des jugements de la Cour fédérale du Canada, qui contient les décisions auxquelles le public a accès rendues depuis 1992 jusqu'à 2004. La collection comporte 10317 jugements en anglais disponibles sur la page web de CanLII (`http://www.canlii.org/ca/ cas/fct/`).

4) Production de la fiche de résumé : LetSum produit un résumé court en éliminant les unités moins importantes de la liste des unités candidates au niveau de chaque segment thématique. Le système ne garde que les unités textuelles dans la limite de la taille autorisée du segment, selon les statistiques du Tableau 3. Le système génère une fiche de résumé en assemblant les unités sélectionnées de chaque segment thématique en indiquant le thème des fragments dans la jurisprudence.

8. Un exemple de sortie de LetSum

Le Tableau 8 montre un exemple des fragments textuels pertinents sélectionnés par LetSum appliqué sur un jugement de la Cour fédérale du Canada (2468 mots). Les informations dans le segment de DONNÉES DE LA DÉCISION sont les informations pertinentes extraites de la page de couverture du jugement et les informations dans le champ autorités et doctrines citées sont les traces des références citées dans le jugement. Le module de **Segmentation thématique** sépare le texte en segments thématiques (les thèmes sont écrits à la partie gauche des cases dans le Tableau 8). Le module de **Filtrage** supprime la phrase qui signale la citation ainsi que les paragraphes cités énumérés (paragraphe (15) marqués Citation dans le Tableau 8). Module de **Sélection** choisie les unités textuelles (montré en gras dans le Tableau 8) pour chaque segment thématique. Les unités sont sélectionnées d'après leurs thèmes dans le jugement. À partir toutes ces informations, le module de **Production** contrôle la taille de résumé, si l'ensemble des phrases sélectionnées dépasse 10% du texte source, il élimine les certaines phrases candidates selon l'importance des segments thématiques.

9. Évaluation

Pour évaluer la qualité de sortie du système, on peut considérer deux méthodes : une première compare le résumé généré par le système et un résumé modèle existant. Un outil dans GATE (AnnotationDiff) permet d'évaluer les annotations des unités saillantes. Il est possible d'aligner automatiquement les unités de deux textes pour comparer la similarité entre les résumés modèles et les résumés produits afin de cal-

DONNÉES	Nom de la cour de décision : Cour fédérale du Canada Lieu de l'audience : Ottawa Date : 31/12/97, Numéro de greffe : T-1989-96 Identification des parties : Commissaire aux langues officielles du Canada, Requéran - et - Air Canada, Intimée Intitulé du jugement : Official languages Autorités et doctrines citées : Section 78 of the Official Languages Act
INTRODUCTION	(1) An order was made by this Court on February 4, 1997 authorizing the respondent (Air Canada) to raise preliminary objections to the notice of an originating motion filed by the applicant (the Commissioner). As a result, this **motion filed by Air Canada** on March 18, 1997 raises six alternative preliminary objections **asking the Court to strike out in part the motion made by the Commissioner** on September 6, 1996 **under section 78 of the Official Languages Act.**
CONTEXTE	1. **Facts** (2) **The Commissioner's originating motion**, which was filed with the consent of the complainant Paul Comeau, **concerns Air Canada's failure to provide ground services in the French language at the Halifax airport.** (3) **The Commissioner's motion is filed by the complainant Paul Comeau.** ...
CITATION	(15) The point of departure is paragraph 78(1), which reads as follows : 78. (1) The Commissioner may (a) within the time limits prescribed by paragraph 77(2)(a) or (b), apply to the Court for a remedy under this Part in relation to a complaint investigated by the Commissioner if the Commissioner has the consent of the complainant. (b) appear before the Court on behalf of any person who has applied under section 77 for a remedy under this Part ...
RAISONNEMENT	(18) **In my view, the purpose of section 79 is to enable the Commissioner to prove to the Court that there is a systemic problem and that it has existed for a number of years.** Unless all similar complaints are filed in evidence, the Court cannot assess the scope of the problem and the circumstances of the application. ...
CONCLUSION	7. **Conclusion** (30) This **motion** to strike by Air Canada with respect to the preliminary objections **must accordingly be dismissed.**

Tableau 8. *Un exemple de sortie de LetSum. Le jugement source est divisé en segments thématiques, le bloc de citation va être supprimer par le module de filtrage. Les unités montrées en gras sont jugées pertinentes et les phrases les contenant constitueront le résumé final.*

culer la précision et le rappel qui calcule la fraction du résumé modèle exprimée dans le contenu du résumé produit par le système. La précision mesure la proportion des unités pertinentes parmi tous les unités retrouvées par le système. Le rappel mesure la proportion des unités pertinentes retrouvées parmi tous les unités pertinentes dans le document. F-Mesure est souvent utilisé comme une conjonction entre la précision et le rappel, qui calcule le poids moyen entre les deux.

Un autre méthode d'évaluation de résumé peut être utilisée quand le résumé modèle n'est pas disponible. À l'aide de 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.

Nous avons d'abord évalué les modules de LetSum séparément. Nous avons comparé les sorties de modules de systèmes avec le corpus annoté manuellement. Le corpus de test utilisé est de a 25 000 mots pour 15 jugements qui n'ont pas été utilisés pour entraîner le système, ni servi à la construction du dictionnaire des marqueurs. Le résultat de l'évaluation pour le segmenteur thématique est de 0.90 pour la F-mesure. L'évaluation du module de filtrage pour l'identification de citation est de 0.98 pour la précision et 0.95 pour le rappel qui donne 0.96 pour le F-mesure. Sur 60 cas de citation, 57 unités ont été identifiées correctement. Nous compléterons sous peu les évaluations des modules statistiques ainsi que l'évaluation globale du résumé produit.

10. Travaux futurs

Nous sommes présentement en mesure de produire des résumés en anglais par identification de structures thématiques, l'étape suivante sera d'utiliser les structures thématiques repérés pour les tâches plus spécifiques et d'analyser les phrases extraites en détail. Un exemple d'application de cette analyse est la génération de titres significatifs pour la jurisprudence. La deuxième étape de nos travaux futurs est d'adapter notre méthodologie au français qui est l'autre langue utilisée sur le site CanLII.

10.1. *Génération des titres significatifs*

Dans les articles des journaux, les messages les plus importants apparaissent dans les mots de titre. Mais dans les jugements, il n'y a pas de vrai titre pour le document. Par contre les premiers paragraphes du jugement donnent un aperçu global sur le cas. À partir des premiers segments, que nous avons appelé INTRODUCTION dans la fiche de résumé, nous générons un titre informatif pour le jugement qui est limité à un maximum de 15 mots. Ce titre sera un résumé très court pour le jugement et il sert à identifier rapidement le but du document.

Une application des titres significatifs est la présentation des résultats de la recherche d'informations. Pour une requête dans la base de données de CanLII, des dizaines des documents peuvent être présentés avec leurs titres significatifs. En regardant les titres des jugements, qui montre le but de chaque document, l'utilisateur pourra décider rapidement de la pertinence d'une jurisprudence pour sa recherche.

10.2. *Traitement des jugements en français*

Notre méthodologie de création de résumés automatiques a été développée pour les jugements en anglais, or nous souhaitons l'adapter pour les documents en français. Les principaux composants du système devraient rester les mêmes, mais il faudrait adapter le dictionnaire des marqueurs linguistiques. Nous avons l'avantage d'avoir un corpus bilingue des jugements en anglais et français, ce qui nous donne la possibilité d'aligner les textes parallèles en deux langues duquel nous pouvons extraire les termes français correspondant aux termes anglais que nous utilisons actuellement. Nous développerons aussi les règles sémantiques en français qui vérifient les présences des marqueurs dans le contexte du jugement.

11. Travaux connexes

LetSum est un des premiers systèmes à traiter spécifiquement du problème des résumés de textes de jurisprudences. Toutefois il y a déjà eu des travaux dans le domaine juridique afin de fournir des outils linguistiques permettant d'aider les avocats et les juristes.

FLEXICON de l'Université de Colombie-Britannique [SMI 87][GEL 91][SMI 95] est un système développé pour la gestion des informations juridiques et la production du résumé qui combine le traitement du texte avec le 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. 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.

Le projet SALOMON de Katholieke Universiteit Leuven [MOE 96][MOE 99] produit un résumé automatique de cas criminels belges (écrits en néerlandais). 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 des 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 thématique des textes juridiques.

Le projet SUM de l'Université de Edinburgh [GRO 03] utilise l'information rhétorique et la structure du discours au niveau analyse des termes pour chaque phrase afin de générer le résumé en adaptant au texte juridique l'approche de Teufel [TEU 02] proposée pour les articles scientifiques. Ce projet est en cours de développement.

12. Conclusion

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 eu beaucoup de travail dans ce domaine et que le problème du traitement des textes légaux reste ouvert. Les 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 le néerlandais, l'anglais ou le français, mais une approche qui peut être efficace pour identifier les indicateurs marquants les phrases importantes 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 cet article, nous avons présenté notre méthode pour produire un résumé juridique flexible et cohérent. Cette approche est basée sur une analyse d'alignement manuel entre les phrases des résumés modèles et des jugement, afin d'exploiter la structure thématique du texte. Notre corpus contient les décisions de la Cour fédérale du Canada, en anglais. Nous proposons une nouvelle forme de présentation du résumé à l'utilisateur sous forme d'une fiche de résumé qui divise le résumé en différents segments thématiques. Chaque segment de cette fiche est associé à un thème : DONNÉES DE LA DÉCISION, INTRODUCTION, CONTEXTE, RAISONNEMENT JURIDIQUE ou CONCLUSION. Notre prototype LetSum prend en entrée un jugement en anglais le traite avec les étapes suivantes : prétraitements (découpage des paragraphes, phrases et mots, analyse syntaxique et sémantique), segmentation thématique, filtrage des citations, sélection des unités pertinentes et productions du résumé.

Nous étudions la possibilité d'adaptation de notre méthodologie afin de produire les résumés juridiques en français. Dans la première phase de ce projet, nous avons aligné manuellement les résumés modèles et les documents sources pour d'étudier la faisabilité de production automatique de résumé juridique, mais dans la deuxième phase de projet, nous étudierons les avantages d'utilisation les alignements automatiques [JIN 02] afin de construire un corpus annoté plus volumineux.

Remerciements

Nous tenons à remercier l'équipe LexUM du laboratoire d'informatique juridique du Centre de recherche en droit public de la faculté de droit de l'Université de Montréal pour leur collaboration. Ces travaux ont bénéficié des infrastructures du Text Analysis Portal for Research (TAPoR). À l'Université de Montréal, TAPoR est mis en place grâce aux subventions de la Fondation canadienne de l'innovation (Projet No. 5637) et du ministère de l'Éducation du Québec.
La recherche présentée ici est financièrement soutenu par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG).

13. Bibliographie

[BAR 97] BARZILAY R., ELHADAD M., « Using Lexical Chains for Text Summarization », *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, July 1997, p. 10-17.

[CHO 00] CHOI F., « Advances in domain independent linear text segmentation », *Proceding of the 1 st North American Chapter of the Association for Computational Linguistics*, Seattle, Washington, 2000, p. 26-33.

[CUN 02] CUNNINGHAM H., MAYNARD D., BONTCHEVA K., TABLAN V., « GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications », *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, July 2002.

[DUC 03] DUC, « Document Understanding Conference 2003. NAACL, Text Summarization Workshop », http ://duc.nist.gov, May 31 - June 1 2003.

[EDM 69] EDMUNDSON H. P., « New Methods in Automatic Extracting », *Journal of the Association for Computing Machinery*, vol. 16, n° 2, 1969, p. 264–285.

[FAR 04a] FARZINDAR A., « Développement d'un système de résumé automatique de textes juridiques », *TALN-RECITAL'2004*, Fès, Maroc, 19-22 April 2004, p. 39-44.

[FAR 04b] FARZINDAR A., LAPALME G., « Legal text summarization by exploration of the thematic structures and argumentative roles », *Text Summarization Branches Out Workshop held in conjunction with ACL04*, Barcelona, Spain, 25-26 July 2004.

[GEL 91] GELBART D., SMITH J. C., « Beyond Boolean search, Flexicon, a legal text-based intelligent system », *the Third International Conference on Artificial Intelligence and Law*, New York, U.S.A., 1991.

[GRO 03] GROVER C., HACHEY B., KORYCINSKI C., « Summarising Legal Texts : Sentential Tense and Argumentative Roles », RADEV D., TEUFEL S., Eds., *HLT-NAACL 2003 Workshop : Text Summarization (DUC03)*, Edmonton, Alberta, Canada, May 31 - June 1 2003, p. 33–40.

[HEA 94] HEARST M. A., « Multi-Paragraph Segmentation of Expository Text », *the 32nd Meeting of the Association for Computational Linguistics*, Los Cruces, NM, June 1994.

[HEP 00] HEPPLE M., « Independence and Commitment : Assumptions for Rapid Training and Execution of Rule-based Part-of-Speech Taggers », *the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, October 2000, p. 278-285.

[JIN 02] JING H., « Using Hidden Markov Modelling to Decompose Human-Written Summaries », *Computational Linguistics*, vol. 28, n° 4, 2002.

[LUH 59] LUHN H., « The Automatic Creation of Literature Abstracts », *IBM Journal of Research and Development*, , 1959, p. 159-165.

[MAI 96] MAILHOT L., *Ecrire la décision : guide pratique de rédaction judiciaire*, Editions Yvon Blais, Québec, Canada, 1996.

[MAN 99] MANI I., MAYBURY M., *Advances in automatic text summarization*, Kluwer Academic Publishers, Boston, U.S.A., 1999.

[MAN 01] MANI I., *Automatic Text Summarization*, John Benjamins Publishing Company, 2001.

[MAR 97] MARCU D., « The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts », PhD thesis, University of Toronto, 1997.

[MAR 00] MARCU D., *The Theory and Practice of Discourse Parsing and Summarization*, MIT Press, Cambridge/London, 2000.

[MIN 01] MINEL J.-L., DESCLÉS J.-P., CARTIER E., CRISPINO G., BEN HAZEZ S., JACKIEWICZ A., « Résumé automatique par filtrage sémantique d'informations dans des textes », *Revue Technique et Science Informatiques*, , n° 3, 2001.

[MIN 02] MINEL J.-L., *Filtrage sémantique : du résumé automatique à la fouille de textes*, Editions Hermès, Paris, France, 2002.

[MOE 96] MOENS M.-F., GEBRUERS R., UYTTENDAELE C., « SALOMON : Final Report », rapport, 1996, Katholieke Universiteit Leuven.

[MOE 99] MOENS M.-F., UYTTENDAELE C., DUMORTIER J., « Abstracting of legal cases : the potential of clustering based on the selection of representative objects », *Journal of the American Society for Information Science*, vol. 50, n° 2, 1999, p. 151-161.

[MOU 00] MOURAD G., « Présentation de connaissances linguistiques pour le repérage et l'extraction de citations », *RECITAL'2000*, Lausanne, 16-18 Octobre 2000.

[MOU 03] MOURAD G., DESCLÉS J.-P., « Identification et extraction automatique des informations citationnelles dans un texte », *Ci-Dit, Colloque international et interdisciplinaire*, Bruxelles, 8- 11 Novembre 2003.

[PAI 81] PAICE C. D., « The Automatic Generation of Literary Abstracts : An Approach Based on Identification of Self-Indicating Phrases », NORMAN O. R., ROBERTSON S. E., VAN RIJSBERGEN C. J., WILLIAMS P. W., Eds., *Information Retrieval Research*, London : Butterworth, 1981.

[SAG 01] SAGGION H., « Génération automatique de résumés par analyse sélective », PhD thesis, Université de Montréal, 2001.

[SAG 02] SAGGION H., LAPALME G., « Generating Indicative-Informative Summaries with SumUM », *Computational Linguistics*, vol. 28, n° 4, 2002.

[SAG 03] SAGGION H., BONTCHEVA K., CUNNINGHAM H., « Robust Generic and Query-based Summarisation », *EACL'2003*, Budapest, Hungary, April 12–17 2003.

[SMI 87] SMITH J. C., DEEDMAN C., « The Application of Expert Systems Technology to Case-Based Law », *ICAIL*, , 1987, p. 84-93.

[SMI 95] SMITH J. C., GELBART D., MACCRIMMON K., ATHERTON B., MCCLEAN J., SHINEHOFT M., QUINTANA L., « Artificial Intelligence and Legal Discourse : The Flexlaw Legal Text Management System », *Artificial Intelligence and Law*, vol. 3, n° 1-2, 1995, p. 55-95.

[TEU 02] TEUFEL S., MOENS M., « Summarising Scientific Articles - Experiments with Relevance and Rhetorical Status », *Computational Linguistics*, vol. 28, n° 4, 2002, p. 409-445.

[UYT 96] UYTTENDAELE C., MOENS M.-F., DUMORTIER J., « SALOMON : Abstracting of Legal Cases for Effective Access to Court Decisions », *Proceedings of JURIX 96 Ninth International Conference on Legal Knowledge Based Systems*, Tilburg : University Press, 1996, p. 47-58.