Développement d'un système de Résumé automatique de Textes Juridiques

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

Résumé- Abstract

Nous décrivons notre méthode de production automatique du résumé de textes juridiques. C'est une nouvelle application du résumé qui permet aux juristes de consulter rapidement les idées clés d'une décision juridique pour trouver les jurisprudences pertinentes à leurs besoins. Notre approche est basée sur l'exploitation de l'architecture des documents et les structures thématiques, afin de constituer automatiquement des fiches de résumé qui augmentent la cohérence et la lisibilité du résumé. Dans cet article nous détaillons les conceptions des différentes composantes du système, appelé LetSum et le résultat d'évaluation.

We describe our method for dealing with automatic summarization techniques in the legal domain. This new application of summary helps a legal expert determine the key ideas of a judgement. Our approach is based on the exploration of the document's architecture and its thematic structures, in order to build a table style summary, which improves coherency and readability in the summary. We present the components of a system, called LetSum, built with this approach and some preliminary results of the evaluation.

Mots-clefs – Keywords

Résumé automatique, fiches de résumé, segmentation thématique, textes juridiques
Automatic text summarization, summary table, topic segmentation, legals texts

1 Introduction

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, il faut aussi considérer les besoins de l'utilisateur et de la tâche spécifiée (Minel *et al.*, 2001). 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 répond aux besoins des avocats, des juges et des experts du domaine. On s'intéresse au traitement des décisions juridiques passées 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. Pour une requête dans une base de données de précédents, on va souvent recevoir des centaines documents 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.

Dans les sections suivantes, nous décrivons notre méthode basé sur l'identification des structures thématiques des jugements afin de produire un résumé cohérent sous forme d'une fiche de résumé structurée. Nous présentons les différents modules du système développé, basé sur cette approche ainsi que son implantation et les résultats d'évaluations partielles des fonctionnements des modules.

2 Structuration des textes juridiques

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 http://reports.fja.gc.ca/. Nous avons analysé 50 jugements et leur résumés rédigés manuellement par un arrêtiste, un résumeur professionnel. Nous cherchons les relations entre les informations considérées importantes dans les résumés modèles et les informations dans les documents sources. La taille moyenne des jugements comme entrée de notre système est des décisions qui ont entre 500 et 4000 mots. Pour notre analyse de corpus, nous avons identifié des cadres organisationnels pour le jugement. Les paragraphes qui traitent d'un sujet sont considérés comme membres d'un groupe thématique. Les blocs sont annotés avec une étiquette qui montre leurs rôles thématiques. Nous avons aussi manuellement annoté les unités de citation. Les citations sont les unités textuelles (phrase ou paragraphe) citées 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 éliminé lors des filtrages d'information.

Les travaux linguistiques de Charolles (Charolles, 2002) définissent la notion de *cadre de discours* qui identifie les circonstances relatives à un certain état ou à une série d'événements. Les cadres de discours partitionnent l'information dans des rubriques homogènes. Dans ce contexte, nous avons cherché les niveaux de discours qui partitionnent des décisions juridiques en différentes structures discursives malgré la variabilité des catégories des jugements (Mailhot, 1996). L'identification de ces structures sépare les idées clés des détails secondaires d'un jugement et elle améliore la lisibilité de résumé qui peut produire des textes plus cohérents. La table 1 montre la structuration d'une jurisprudence et ses différents niveaux de discours comme: *Données de la décision*, *Introduction*, *Contexte*, *Raisonnement juridique* et *Conclusion*. Ainsi, dans la présentation du résumé final nous proposons de conserver cette organisation des structures du texte afin de constituer une fiche de résumé de décisions.

Structures thématiques	**Explications**
Données de la décision	Numéro de greffe, Référence neutre, Date du jugement, Nom de la cour de décision, Identification des parties, L'intitulé du jugement.
Introduction	Qui? A fait quoi? À qui?
Contexte	Faits recompose l'histoire du litige; Histoire judiciaire
Citation	Prétention des parties; Citation des références juridiques
Raisonnement juridique	Discussion, analyse du juge et détermination des faits; Expression des motifs de la solution retenue.
Conclusion	*Dispositif*, la décision finale de la cour.

Table 1: Fiche de résumé montre des **structures thématiques** dans un document de jurisprudence

3 Méthode de la constitution de fiches de résumé

Nous proposons une approche de résumé automatique basée sur l'identification des structures thématiques et les rôles argumentatifs, 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 rôles argumentatifs 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 thème en regardant le rôle argumentatif 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. La constitution de la fiche de résumé se fait en quatre étapes (figure 1) : segmentation thématique et les rôles argumentatifs, 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.

Pour la tâche de segmentation, nous avons fait quelques expérimentations avec deux segmenteurs décrits par Hearst (Hearst, 1994) le système TexTiling et le segmenteur C99 décrit par Choi (Choi, 2000), les deux sont les segmenteurs statistiques qui impliquent une fonction de *clustering* sur document, pour trouver des classes divisées par thèmes. Mais les résultats de ces segmenteurs numériques n'étaient pas à un niveau satisfaisant pour trouver les structures thématiques des jugements judiciaires donc nous avons décidé de développer un processus de segmentation thématique basé sur les connaissances spécifiques du domaine juridique.

L'implantation de notre approche est un système appelé LetSum, développé en langage Java et Perl. L'entrée du système est un document de jurisprudence qui peut avoir un des formats XML, HTML, SGML, RFT ou un texte sans balise. Pour traiter le document, le texte est divisé en paragraphes, phrases et en unités plus petites comme mots, nombres et ponctuations. L'analyseur syntaxique utilisé, pour déterminer les catégories syntaxiques des mots est le modèle décrit par Hepple (Hepple, 2000). Les règles et les grammaires sémantiques sont écrits en langage JAPE (Java Annotations Pattern Engine) qui peuvent être exécutés avec le transducer de GATE (Cunningham *et al.*, 2002). GATE offre la possibilité d'extraction de certaines entités

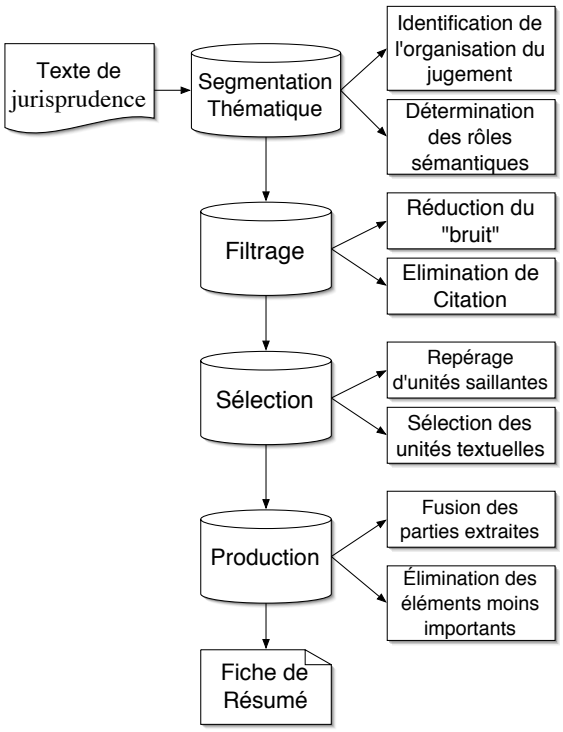

Figure 1: Les étapes de la constitution de la fiche de résumé

nommées (comme noms des personnes, dates, lieux, etc.) et des coréférences.

Segmentation thématique: cette étape détermine l'organisation du document original et encadre les blocs du texte associés avec un rôle argumentatif précis dans la jurisprudence. Notre segmenteur thématique est fondé sur les connaissances sémantiques du jugement. Afin de séparer un segment thématique et de détecter son rôle sémantique nous avons utilisé des conditions telles que la présence des titres des sections, les positions (absolue et relative) d'un segment, l'identification des styles direct et indirect et la présence des marqueurs linguistiques.

Filtrage: cette étape identifie les *exemples négatifs* qui peuvent être supprimés dans les documents, sans perdre les informations pertinentes pour le résumé. Dans un jugement, les citations occupent un volume important dans le texte soit 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 blocs de 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. L'identification de citation est basée sur deux types de marqueurs : directs et indirects (Mourad & Desclés, 2003). La première catégorie de marqueur direct contient les indicateurs linguistiques. Les exemples de marqueurs linguistiques identifiés en trois classes ; les verbes de citation comme *conclude, define, indicate, provide, read, reference, refer, say, state, summarize*, les concepts (nom, adverbe, adjectif) comme *following, section, subsection, page, paragraph, pursuant* et les indices complémentaires comme les nombres, les préposition, les subordonnés relatives et les marqueurs typographique comme deux-points et guillemet. La deuxième catégorie comprend les marqueurs de citation indirects. Les unités textuelles citées indirectement sont les unités voisines des phrases citées directement. Nous avons donc intégré un mécanisme d'identification des in-

tégrations linéaires entre les phrases suivantes de la première phrase citée avec les marqueurs directs.

Sélection des unités saillantes: cette étape construit une liste d'unités saillantes candidates pour chaque niveau structural du résumé. LetSum calcule un poids pour chaque phrase dans le jugement d'après les fonctions heuristiques 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 et les fréquances des mots dans le texte et le corpus ($tf * idf$).

Production de la fiche de 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, nous avons constaté que 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* est 60% et les unités avec le rôle *Conclusion* occupent 6% du résumé.

4 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 similaires dans le domaine juridique afin de fournir des outils linguistiques permettant d'aider les avocats et les juristes. FLEXICON (Smith & Deedman, 1987) a été 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. Le projet SALOMON (Uyttendaele *et al.*, 1996) produit le résumé automatiquement de cas criminels belges (écrits en Hollandais). Le but est d'identifier et d'extraire les informations importantes à partir des jurisprudences. Dans ce projet les connaissances linguistiques sont utilisées. Il extrait les concepts et les unités textuelles saillantes grâce à l'identification des *cue words*, segments indicateurs et patrons de contexte.

5 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. Dans cet article, nous avons présenté notre méthode pour produire un résumé juridique flexible et cohérent. Cette approche est basée sur une analyse de corpus des 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 thèmes. Chaque niveau structural de cette fiche est associé à un rôle argumentatif : données de la décision, introduction, contexte, raisonnement juridique ou conclusion. Le système étant en cours de développement mais les résultats préliminaires sur le segmetuer (F-mesure = 0.90) et le module de filtrage (F-mesure = 0.96) sont très encourageants. Nous complétons actuellement le développement des modules statistiques qui calculent la similarité d'une phrase avec les phrases voisines et la similarité entre d'une phrase avec les autres phrases dans le même segment thématique.

Remerciements

Nous tenons à remercier l'équipe LexUM du laboratoire d'informatique juridique du Centre de recherche en droit public (CRDP) pour leur collaboration.
Merci à l'équipe de Jean-Pierre Desclès, laboratoire LaLICC de l'Université Paris-Sorbonne.
Ce projet est financièrement soutenu par le Centre de recherche en droit public du Canada.

Références

CHAROLLES M. (2002). Organisation des discours et segmentation des écrit. In *Inscription Spatiale du Langage: structures et processus, IRIT*, Toulouse, France.

CHOI F. (2000). Advances in domain independent linear text segmentation. In *Proceding of the 1 st North American Chapter of the Association for Computational Linguistics*, p. 26–33, Seattle, Washington.

CUNNINGHAM H., MAYNARD D., BONTCHEVA K. & TABLAN V. (2002). Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia.

HEARST M. A. (1994). Multi-paragraph segmentation of expository text. In *the 32nd Meeting of the Association for Computational Linguistics*, Los Cruces, NM.

HEPPLE M. (2000). Independence and commitment: Assumptions for rapid training and execution of rule-based part-of-speech taggers. In *the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, p. 278–285.

MAILHOT L. (1996). *Ecrire la décision: guide pratique de rédaction judiciaire*. Québec, Canada: Editions Yvon Blais.

MINEL J.-L., DESCLÉS J.-P., CARTIER E., CRISPINO G., BEN HAZEZ S. & JACKIEWICZ A. (2001). Résumé automatique par filtrage sémantique d'informations dans des textes. *Revue Technique et Science Informatiques*, (3).

MOURAD G. & DESCLÉS J.-P. (2003). Identification et extraction automatique des informations citationnelles dans un texte. In *Ci-Dit, Colloque international et interdisciplinaire*, Bruxelles.

SMITH J. C. & DEEDMAN C. (1987). The application of expert systems technology to case-based law. *ICAIL*, p. 84–93.

UYTTENDAELE C., MOENS M.-F. & DUMORTIER J. (1996). Salomon: Abstracting of legal cases for effective access to court decisions. In *Proceedings of JURIX 96 Ninth International Conference on Legal Knowledge Based Systems*, p. 47–58: Tilburg: University Press.