TALN 2004, Fès, 19–21 avril 2004

Anonymisation de décisions de justice

Luc Plamondon (1), Guy Lapalme (1), Frédéric Pelletier (2)
(1) RALI - Université de Montréal
Montréal, Québec, Canada
{plamondl, lapalme}@iro.umontreal.ca
(2) CRDP - Université de Montréal
Montréal, Québec, Canada
pelletif@lexum.umontreal.ca

Résumé - Abstract

La publication de décisions de justice sur le Web permet de rendre la jurisprudence accessible au grand public, mais il existe des domaines du droit pour lesquels la Loi prévoit que l'identité de certaines personnes doit demeurer confidentielle. Nous développons actuellement un système d'anonymisation automatique à l'aide de l'environnement de développement GATE. Le système doit reconnaître certaines entités nommées comme les noms de personne, les lieux et les noms d'entreprise, puis déterminer automatiquement celles qui sont de nature à permettre l'identification des personnes visées par les restrictions légales à la publication.

Publishing court decisions on the Web can make case law available to the general public, but the Law sometimes prohibits the disclosure of the identity of people named in decisions. We are currently developing an automatic anonymization system, using the GATE development environment. The tasks of the system are the recognition of some named entities like person names, locations and company names, then the automatic selection of the ones that may lead to the identification of people whose identities must be legally kept confidential.

Mots-clefs – Keywords

Anonymisation, désidentification, reconnaissance d'entités nommées, textes juridiques.
Anonymization, de-identification, named entity recognition, law texts.

1 Introduction

On estime à 200 000 le nombre de décisions rendues annuellement par les tribunaux judiciaires canadiens, représentant 2 millions de pages de texte, sans compter les décisions rendues par les tribunaux administratifs.

L'accès à la jurisprudence, c'est-à-dire l'ensemble des décisions rendues par les tribunaux, est primordial dans les sociétés démocratiques, non seulement pour assurer la transparence de la justice et l'indépendance de la magistrature, mais aussi pour informer les citoyens et leurs avocats sur l'état du droit. Ceci demeure vrai autant dans les pays dont la tradition juridique relève de la *common law*, comme le Canada et le Royaume-Uni, que dans les pays de tradition juridique continentale européenne, comme la France.

Pour le bénéfice des professionnels du droit, des éditeurs privés publient depuis longtemps des recueils regroupant une sélection des décisions et mettent en ligne des banques de données relativement exhaustives. Cependant, en raison du travail éditorial considérable effectué sur les décisions ainsi publiées, le coût d'accès demeure prohibitif pour le grand public. Depuis l'avènement du Web et avec l'utilisation généralisée des nouvelles technologies dans la préparation matérielle des décisions, il est maintenant possible de rendre le texte brut de l'ensemble des décisions de justice accessible au grand public, et ce à faible coût. C'est dans ce contexte que l'Institut canadien d'information juridique (IIJCan) a entrepris en 2000 un important projet de diffusion libre du droit canadien sur le Web (www.iijcan.org).

Bien que la diffusion intégrale du texte des décisions de justice soit la règle générale, il existe au Canada, comme dans plusieurs autres pays, des restrictions légales quant à la divulgation de l'identité des participants les plus vulnérables aux procédures judiciaires, tels les enfants en matière familiale ou de protection de la jeunesse, les jeunes contrevenants, les adultes sous régime de protection ou les victimes d'agression sexuelle. Il résulte de ces restrictions que plusieurs décisions nécessitent une étape manuelle d'anonymisation avant leur diffusion. L'anonymisation manuelle requiert en moyenne de une à deux minutes de travail par page de texte. Le laboratoire LexUM du Centre de recherche en droit public de l'Université de Montréal estime que le tiers des décisions rendues au Canada ne sont pas publiées gratuitement sur le Web, faute de méthode d'anonymisation fiable et peu coûteuse.

Nous avons entrepris le développement d'un système d'anonymisation automatique de décisions de justice. Le système devra, à terme :

1. reconnaître automatiquement les noms, dates, lieux et autres entités nommées susceptibles de constituer des renseignements permettant d'identifier les personnes visées par les restrictions à la publication ;
2. regrouper les renseignements qui concernent un même individu ;
3. déterminer quels individus requièrent l'anonymat ;
4. procéder à la modification ou la suppression des renseignements relatifs à ces individus ;
5. offrir une interface conviviale pour une révision/correction manuelle.

L'étape 1 relève du domaine de la reconnaissance d'entités nommées. Les textes juridiques présentent cependant des difficultés qui leur sont propres. Nous présentons à la figure 1 un premier prototype qui effectue la reconnaissance des noms propres de personne dans les décisions. Ce prototype a été développé à l'aide de la suite de développement GATE. Nous en parlerons plus abondamment à la section 3.

Les autres étapes en sont encore à l'état de projet. L'étape 2 consistera à regrouper les différentes appellations d'un même individu ("`Jeremy R. Sullivan`", "`Mr. Sullivan`", etc.) et certains renseignements personnels comme sa date de naissance, dans le but de procéder à une anonymisation en bloc si requis. L'étape 3 est nécessaire car tous les individus dont une décision fait mention n'ont pas à être anonymisés : cette étape présentera des difficultés tant au point de vue

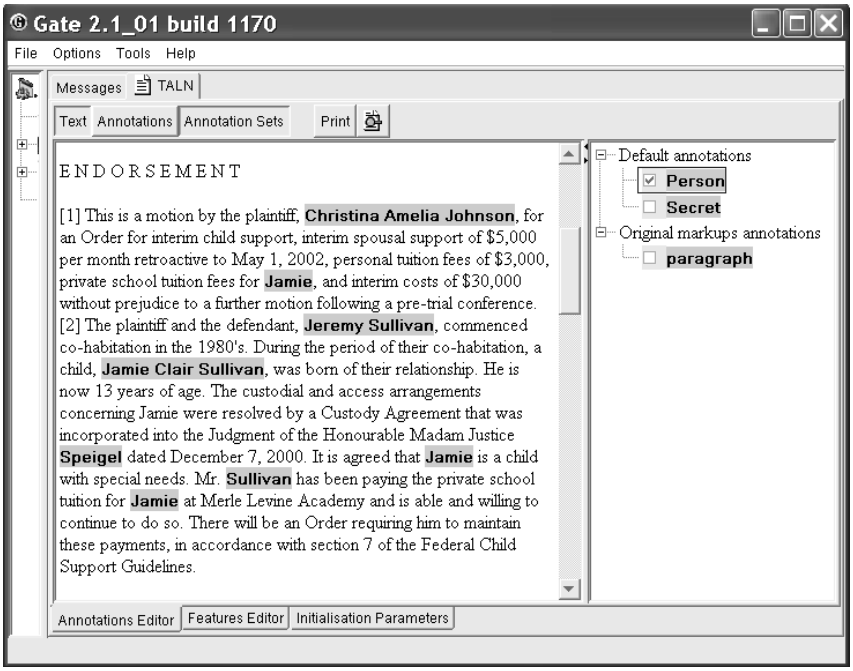

FIG. 1 – Prototype du système d'anonymisation automatique. L'étape illustrée est la reconnaissance des noms propres de personne. L'interface graphique est celle de GATE. La décision montrée en exemple est authentique mais les noms sont fictifs.

informatique linguistique que juridique. Une fois la décision prise, les informations confidentielles devront être modifiées (étape 4). Par exemple, conformément aux pratiques actuelles des éditeurs, les noms de personne devront être remplacés par des initiales ("`Mr. Sullivan` → `Mr. S.`", "`Jeremy R. Sullivan` → `J. R. S.`") et le jour et le mois de naissance seront supprimés ("`born February 3, 1998` → `born [...], 1998`"). La dernière étape (étape 5) est la seule à ne pas être automatisée : il s'agit en fait d'offrir la possibilité à un humain de réviser et de corriger le travail fait par le système.

2 Que faut-il anonymiser ?

Pour anonymiser un document, il ne suffit pas de supprimer tous les noms de personne qu'il contient. En effet, une étude a montré que le code postal, la date de naissance et le sexe d'une personne suffisent à identifier de façon unique 87 % de la population des États-Unis et la seule combinaison du code postal et de la date de naissance suffit à identifier 97 % de la population de la ville de Cambridge au Massachusetts (Sweeney, 2001).

À l'inverse, il faut résister à la tentation de masquer le plus d'informations possible car les textes perdent d'autant plus de leur valeur. Par exemple, l'extrait de décision "`La demande de statut de réfugié de Monsieur A. R., originaire de Y, est refusée`" est inutile aux fins d'une recherche portant sur l'acceptation des demandes de réfugiés en fonction du pays d'origine. La pertinence juridique du document publié doit être maintenue.

Le laboratoire LexUM du Centre de recherche en droit public de l'Université de Montréal s'est penché sur la question du juste niveau d'anonymat (Pelletier, 2003). S'inspirant des pratiques

Renseignements à supprimer systématiquement pour toute personne visée par une restriction à la publication, ainsi que pour chacun de ses proches (parents, enfants, professeurs, voisins, employeurs, collègues, établissement scolaire, etc.) :

1. le nom (et surnom) ;
2. le jour et le mois de naissance ;
3. le lieu de naissance ;
4. les coordonnées (numéro, rue, municipalité, code postal, téléphone, télécopieur, courriel, page Web, adresse IP) ;
5. les identificateurs personnels uniques (numéro de sécurité sociale, d'assurance maladie, de dossier médical, de passeport, de compte bancaire, de carte de crédit, etc.) ;
6. les identificateurs de possessions personnelles (numéro de licence ou de série, désignation cadastrale, nom d'entreprise, etc.).

Renseignements à supprimer selon la situation, s'ils permettent d'identifier une des personnes ci-haut :

7. les petites communautés ou lieux géographiques ;
8. les accusés et co-accusés si leur identité n'est pas déjà protégée par la loi ;
9. les intervenants (experts de la cour, travailleurs sociaux, officiers de police, médecins, etc.) ;
10. les informations inhabituelles (nombre d'enfants s'il est anormalement élevé, revenu s'il est particulièrement élevé, occupation ou fonction exceptionnelle).

FIG. 2 – Liste des renseignements à anonymiser dans une décision de justice.

actuelles de divers éditeurs de décisions de justice, il a conclu qu'il faut supprimer les renseignements énumérés à la figure 2.

S'inspirant des pratiques actuelles de divers éditeurs de décisions de justice, il a conclu que pour les domaines de droit pour lesquels une restriction à la publication s'applique le plus souvent – en particulier les affaires pénales impliquant des victimes d'agression sexuelle ou des personnes mineures, ainsi que les affaires familiales impliquant des enfants – il est possible de supprimer systématiquement les renseignements énumérés à la figure 2 tout en préservant un bon niveau de pertinence juridique au texte. Il est prévisible que cette liste doive être légèrement adaptée suivant le type de tribunal ou le domaine de droit concerné.

La plupart des éléments à masquer sont des entités nommées, c'est-à-dire des noms propres, des dates ou des nombres. Les reconnaître dans un texte constitue déjà un champ à part entière du traitement automatique de la langue. Le second défi est de déterminer automatiquement lesquelles nécessitent d'être anonymisées.

3 Un premier prototype avec GATE

Pour développer un prototype de système, nous avons choisi d'utiliser GATE (*General Architecture for Text Engineering*, http://gate.ac.uk), un environnement de développement conçu spécialement pour l'ingénierie des langues (Cunningham *et al.*, 2003). Cet environnement dé-

finit une architecture standard qui permet de construire des applications à partir de modules indépendants. Par exemple, l'environnement est livré avec des modules autonomes pour la tokenisation de textes, la segmentation en phrases, l'étiquetage grammatical, la reconnaissance d'entités nommées et la résolution de liens de coréférence ; il suffit de mettre les modules bout à bout pour obtenir rapidement une application fonctionnelle. Le développeur peut ajouter ses propres modules à l'endroit approprié du pipeline ou remplacer un module d'origine. Pour faciliter la conception de modules, l'environnement GATE inclut une bibliothèque de classes Java pour effectuer des tâches courantes en manipulation de textes, telles la lecture d'un fichier en différents formats, la constitution du graphe des annotations et l'écriture en format XML. De plus, une interface graphique de développement permet d'annoter manuellement des documents (pour constituer le corpus de référence) ou de visionner les résultats de l'application automatique et de les comparer avec le corpus de référence.

Le système d'anonymisation en cours de développement sera formé, à terme, du pipeline de modules illustré à la figure 3. Il s'agit d'un projet devant s'échelonner sur plusieurs années et au moment d'écrire ces lignes, nous en étions à la phase de reconnaissance des noms de personne.

GATE inclut déjà une série de modules permettant de faire de la reconnaissance d'entités nommées : la suite ANNIE (*A Nearly-New Information Extraction System*) (Cunningham *et al.*, 2003, chapitre 7). Cependant, cette suite a été conçue pour analyser des textes journalistiques alors que nous cherchons à analyser des textes juridiques. Bien que nous ayons pu récupérer certains modules de la suite ANNIE (pour la tokenisation et l'étiquetage grammatical), nous avons dû modifier les lexiques et réécrire entièrement la grammaire de reconnaissance des noms de personne. Cette grammaire est écrite en langage JAPE (*Java Annotation Patterns Engine*), une variante du standard CPSL adapté au langage de programmation Java (Cunningham *et al.*, 2003, annexe B). Nous avons modifié le transducteur lui-même afin de combler des lacunes du langage JAPE et ainsi nous permettre plus de flexibilité dans l'écriture des règles.

Jusqu'à maintenant, nous nous sommes concentrés sur la reconnaissance des noms propres de personne dans les décisions de la Cour supérieure de l'Ontario. Conformément aux travaux de Sweeney (1996) et de Grouin (2002), nous avons privilégié l'utilisation de lexiques et de règles pour couvrir les cas les plus courants. Plus précisément, nous utilisons le lexique de prénoms et le lexique de titres de civilité de GATE que nous avons augmenté avec des titres de fonction propres au monde juridique (*the Honourable*, *Justice*, *Q.C.*, etc.). La présence d'un titre, d'un prénom connu ou d'une initiale (ou de certaines combinaisons de ces éléments) signale le début d'un nom: tous les mots qui suivent et qui débutent par une majuscule sont considérés comme des noms de famille. Certaines constructions permettent d'identifier un nom de famille sans qu'il soit nécessairement précédé d'un titre, d'un prénom du lexique ou d'une initiale, par exemple lorsqu'une décision antérieure est citée avec cette formule: "`Gordon v Young`" (Gordon contre Young). Pour le moment, nous n'utilisons pas de lexique de noms de famille courants.

Afin d'évaluer notre travail de reconnaissance des noms de personne, nous avons constitué un corpus de test fait de décisions totalisant 16 297 mots et contenant 546 noms de personne. Pour l'instant, nous ne nous préoccupons pas de savoir si les noms sont confidentiels ou non. Notre grammaire nous a permis d'obtenir une précision de 98 % et un rappel de 88 % (les noms identifiés partiellement sont ici considérés incorrects). En comparaison, la grammaire de reconnaissance des entités nommées ANNIE (livrée avec GATE) n'atteint que 91 % de précision et 81 % de rappel car elle n'est pas adaptée aux textes juridiques (figure 4).

Les mêmes noms sont répétés plusieurs fois dans une décision. Ainsi, les 546 noms de per-

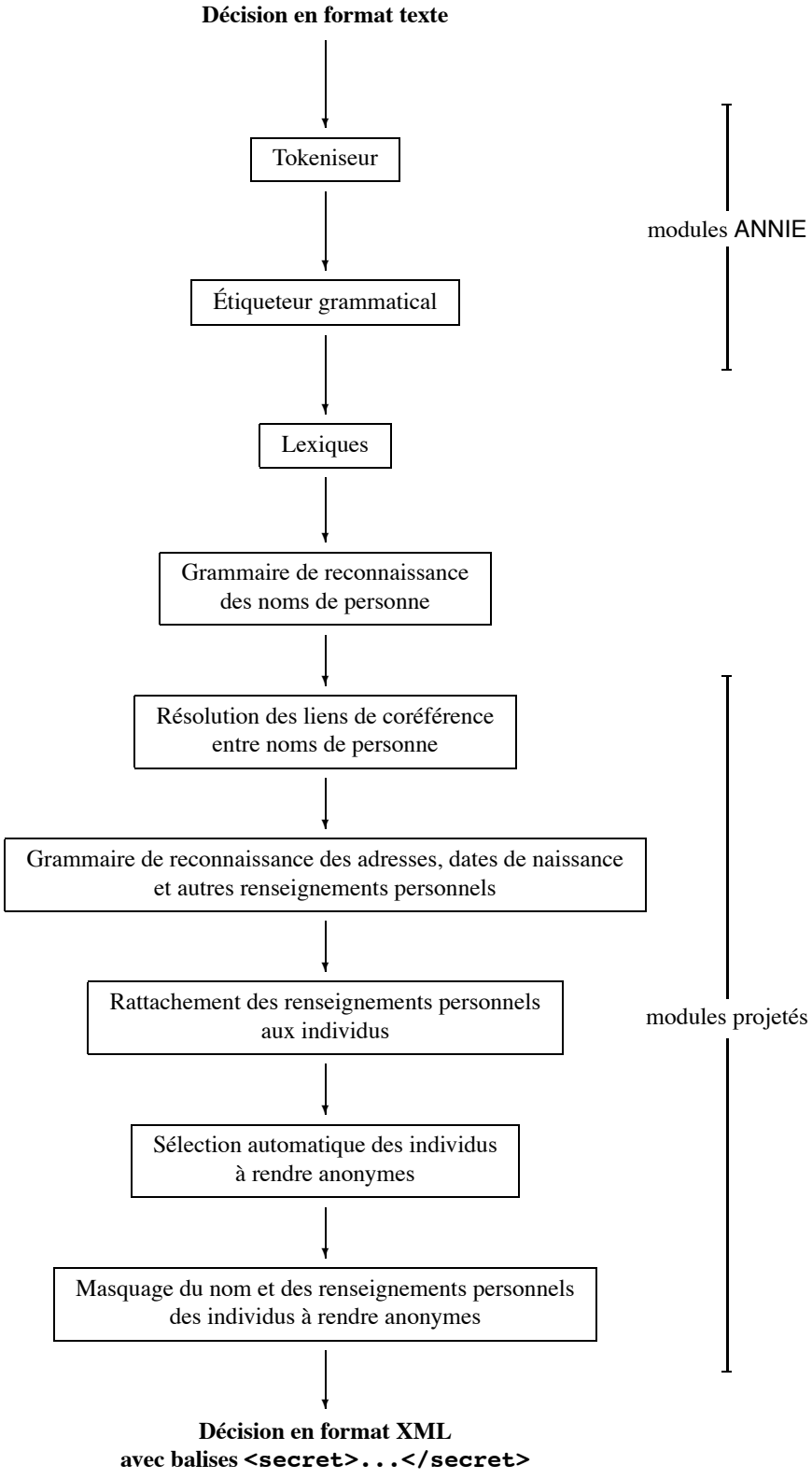

FIG. 3 – Pipeline de modules formant le système d'anonymisation ADJUM.

Grammaire	Occurrences	Précision	Rappel	Formes	Précision	Rappel
Adaptée	546	98 %	88 %	63	88 %	73 %
ANNIE	546	91 %	81 %	63	61 %	48 %

FIG. 4 – Précision et rappel de la grammaire adaptée aux décisions et de la grammaire ANNIE. Les *formes* sont les occurrences dépouillées des doublons.

Omissions	**Occurrences (sur 546)**	**Formes (sur 63)**
Noms de famille sans marqueurs	9	5
Prénoms inconnus	34	2
Noms entièrement en majuscules	22	10
Autres	5	5
Chevauchements	**Occurrences (sur 546)**	**Formes (sur 63)**
Province incluse dans le nom	1	1
1^{er} mot de la phrase inclus dans le nom	1	1
Faux positifs	**Occurrences**	**Formes**
Raison sociale	3	3
Lieu	1	1
Nom commun confondu avec prénom	1	1

FIG. 5 – Causes d'erreur ayant conduit à l'omission de noms par notre grammaire, au chevauchement avec des noms corrects et à l'identification de noms qui n'en sont pas.

sonne du corpus de test se réduisent à 63 formes différentes. Si l'on ne compte qu'une seule fois chacune des formes, la précision de notre grammaire est de 88 % et le rappel est de 73 %. Ces mesures sont davantage indicatrices de la *qualité* de la grammaire puisqu'elles considèrent qu'une forme ignorée a autant de poids qu'une forme bien identifiée. Les mesures brutes, quant à elles, sont révélatrices de l'*utilité* du système puisque l'économie d'efforts apportée par l'automatisation se compte en nombre d'occurrences.

Nous avons privilégié des règles conservatrices (la présence d'un prénom connu ou d'un titre de civilité étant habituellement obligatoire, comme nous l'avons expliqué ci-haut) pour limiter le nombre de faux positifs (figure 5): seulement 5 extraits (tous différents) ont été faussement étiquetés comme étant des noms de personne, contre 39 (se réduisant à 22 formes différentes) avec la grammaire ANNIE. En contrepartie, le rappel en souffre et il devra être amélioré en priorité. Les références à des articles de loi sont une particularité des textes juridiques et nous craignions qu'elles soient identifiées à tort comme étant des noms propres puisqu'elles sont formées de combinaisons d'initiales et de mots débutant par une majuscule ; l'approche conservatrice que nous avons adoptée a permis d'éviter ces pièges avec succès, contrairement à la grammaire ANNIE conçue pour des textes journalistiques. La majorité des entités faussement annotées comme noms de personne par notre grammaire sont en réalité des noms de lieu, d'organisme ou d'entreprise. Ces cas seront examinés plus tard au cours du projet, lorsque nous écrirons les grammaires de reconnaissance de lieux et de raisons sociales et que nous résoudrons les ambiguïtés avec les noms de personne.

Les omissions de la grammaire comptent pour la plus grande part des erreurs. Trois raisons expliquent la majorité des omissions: des noms de famille apparaissent sans marqueurs (titre, prénom connu ou initiale), des prénoms sont inconnus du lexique et des noms sont écrits entièrement en majuscules dans les entêtes. Heureusement, beaucoup de ces individus sont nommés

autrement ailleurs dans la décision et sous une forme reconnue par la grammaire: réanalyser la décision avec les nouvelles informations permettra de récupérer beaucoup d'omissions.

Les deux erreurs de chevauchement d'annotations sont en fait des noms trop longs: des mots adjacents débutant par une majuscule ("`Ontario`" dans le premier cas et le premier mot d'une phrase dans le deuxième cas) ont été considérés à tort comme faisant partie du nom de la personne.

4 Travaux futurs

4.1 La sélection des noms propres confidentiels

Nous sommes présentement en mesure de reconnaître les noms de personne et l'étape suivante sera de déterminer lesquels sont confidentiels. Nous prévoyons utiliser les titres (pour identifier d'emblée les juges et les avocats, lesquels n'ont pas à être anonymisés), l'entête si possible (pour identifier le plaignant et l'accusé) et le contexte (pour identifier les enfants et les proches).

4.2 L'analyse des dates

La date de naissance d'un individu dont l'identité est protégée doit être supprimée. Nous évaluons qu'un nombre limité de règles suffisent à identifier les dates: Grouin en fournit quelques-unes pour le français et GATE en utilise d'autres pour l'anglais. La difficulté résidera dans la distinction entre une date de naissance et une date quelconque. L'examen de plusieurs textes révèlera peut-être que les dates de naissance ne figurent que dans un nombre limité de formulations, comme par exemple "`Jamie, born 1998.`"

4.3 L'analyse des coordonnées et des noms de lieux

Les coordonnées sont faites d'une combinaison du numéro de porte, du nom de la rue, de la municipalité, du code postal, du numéro de téléphone, du numéro de télécopieur, du courriel et de l'URL d'une page Web. Tout comme pour les dates, nous pourrons nous inspirer des règles de Grouin et de GATE, les adapter aux conventions canadiennes et ajouter des règles pour reconnaître les courriels et les URL.

Le nom d'un lieu peut être assez facilement identifié lorsqu'il fait partie d'une adresse complète. Dans les cas où l'adresse est réduite à la ville seulement, ou lorsque l'on fait référence au lieu de travail ou de naissance, nous pouvons faire appel à un lexique. Nous disposons d'une liste de 3 millions de villages, villes et agglomérations du monde entier reconnues par la commission de toponymie américaine (USGS) et l'agence de cartographie et d'imagerie américaine (NIMA).

Tout comme pour les noms de personne, la difficulté sera de distinguer les municipalités à anonymiser des autres.

4.4 L'analyse des raisons sociales

Pour identifier les raisons sociales d'entreprises, d'organismes et d'établissements scolaires, l'utilisation d'un lexique sera moins efficace. D'une part, la liste des entreprises privées change fréquemment: le risque de ne pas identifier une nouvelle entreprise est élevé. D'autre part, les raisons sociales sont souvent longues et donc sujettes à des variations, surtout pour les abréger: *Shell International Limited* ou *Shell International ltée* sont aisés à identifier, surtout à cause des marqueurs *Limited* et *ltée*, mais l'entreprise est mieux connue sous l'appellation *Shell* qui, elle, peut être confondue avec le nom commun *shell* (coquillage).

Le fait que les entreprises doivent être anonymisées seulement lorsqu'il s'agit d'un employeur pourrait faciliter les choses. En effet, une étude de corpus ou un apprentissage automatique permettra peut-être de déceler des séquences de mots qui précèdent habituellement le nom d'un employeur: "`employed by`", "`working for`", "`cashier for`" et autres, toutes ces séquences étant suivies d'un ou plusieurs mots débutant par une majuscule. Le même raisonnement s'applique à l'identification des établissements scolaires.

5 Travaux connexes

La recherche en anonymisation a été initiée il y a déjà plusieurs dizaines d'années pour satisfaire les besoins des instituts de statistiques. Dans ce domaine, le fait que les données soient rigoureusemement structurées permet de formaliser les algorithmes d'anonymisation, ce que Sweeney a été la première à faire (Sweeney, 2001). La plupart des techniques et algorithmes d'anonymisation de bases de données s'appliquent à des données numériques, ce qui ne constitue pas l'essentiel du problème dans le cas des décisions de justice.

Des systèmes d'anonymisation de textes ont vu le jour mais seulement dans le domaine médical. Sweeney (1996) et Grouin (2002) se sont penchés sur la suppression des noms de patient dans les rapports médicaux à l'aide de lexiques et de mots déclencheurs, techniques dont nous nous sommes inspirés pour l'identification des noms propres de personne. Sweeney a obtenu une précision de 100 % avec un seuil optimal de sensibilité des règles ; nous n'avons cependant pas accès à un échantillon du corpus de test et nous croyons que les rapports médicaux sont plus faciles à traiter à cause de leur forme et leur vocabulaire plus restreints que dans les décisions de justice. Grouin a quant à lui obtenu une précision de 87 % et un rappel de 83 %.

Ruch *et al*. (2000) obtiennent une précision de 97 % en n'utilisant que les titres devant les noms ; cette stratégie fait partie de la nôtre mais elle n'est pas suffisante pour les décisions car elle ne permet pas de détecter le nom des enfants, rarement précédé d'un titre. Taira *et al*. (2002) ont entraîné un modèle d'entropie maximale pour repérer un certain nombre de relations logiques faisant intervenir un patient (exemple de relation *Patient-procédure*: "`John received therapy`"). Les paramètres du modèle sont les séquences d'étiquettes grammaticales, les n-grammes, l'ordre des mots, la proximité de certains mots, etc. Le modèle a permis d'isoler les noms de patient avec une précision de 99 % et un rappel de 94 % lorsque le seuil d'acceptabilité du modèle est optimal. Cette approche est intéressante mais la difficulté réside dans le choix des paramètres à modéliser et il n'est pas certain que les relations logiques liant un individu et un contexte particulier soient aussi prononcées dans une décision que dans un rapport médical. Nous examinerons toutefois l'apprentissage automatique comme stratégie complémentaire à celle basée sur des lexiques et des grammaires.

6 Conclusion

L'anonymisation des décisions de justice est un problème complexe qui ne peut être résolu que si les aspects juridiques et informatiques sont traités de front, ce qui explique que peu de travaux aient jusqu'ici porté directement sur le sujet. Du point de vue juridique, il faut bien jauger le niveau d'anonymat requis par la Loi tout en s'assurant que les décisions de justice demeurent lisibles et utiles une fois les informations nominatives supprimées. Du point de vue informatique, le problème est double: il faut d'abord que le système analyse le texte de la décision pour repérer les entités nommées (noms, adresses, lieux, numéros de téléphone, dates, etc.) et ensuite qu'il détermine automatiquement lesquelles sont confidentielles. Le système d'anonymisation automatique que nous développons en est à la phase de reconnaissance des noms propres de personne. Grâce à l'environnement de développement GATE, nous avons rapidement mis sur pied un prototype utilisant des lexiques de prénoms et de titres et une grammaire de reconnaissance des noms propres écrite sur mesure pour les textes juridiques. Nous avons privilégié une approche symbolique plutôt que statistique afin de limiter le bruit, donc le nombre de corrections manuelles à apporter au document, quitte à ce que le rappel requière d'être amélioré. Nous nous attaquerons prochainement à la sélection automatique des individus requérant l'anonymat.

Remerciements

Nous désirons remercier l'instigateur du projet, Daniel Poulin, professeur-chercheur et directeur du laboratoire LexUM au Centre de recherche en droit public de l'Université de Montréal.

Références

CUNNINGHAM H., MAYNARD D., BONTCHEVA K., TABLAN V., URSU C. & DIMITROV M. (2003). *Developing Language Processing Components with GATE (a User Guide)*. Sheffield Natural Language Processing Group. http://www.gate.ac.uk/sale/tao/index.html (visité le 11 janvier 2004).

GROUIN C. (2002). Chaîne de traitements pour la constitution automatique de corpus: application au domaine médical pour le projet corpus CLEF. Master's thesis, INaLCO – Institut National des Langues et Civilisations Orientales, Paris, France. http://grouin.free.fr/projets/clef/sources/memoire/ (visité le 26 août 2003).

PELLETIER F. (2003). Protecting Identities in Published Case Law. Document interne, version préliminaire, 3e révision. LexUM.

RUCH P., BAUD R. H., RASSINOUX A.-M., BOUILLON P. & ROBERT G. (2000). Medical Document Anonymization with a Semantic Lexicon. In J. M. OVERHAGE, Ed., *Proceedings of the 2000 AMIA Annual Symposium*, p. 729–733: American Medical Informatics Association.

SWEENEY L. (1996). Replacing Personally-Identifying Information in Medical Records, the Scrub System. In J. J. CIMINO, Ed., *Proceedings of the 1996 AMIA Annual Symposium*, p. 333–337, Washington, DC: American Medical Informatics Association Hanley & Belfus.

SWEENEY L. (2001). *Computational Disclosure Control, A Primer on Data Privacy Protection*. Rapport interne, Carnegie Mellon University. http://www.swiss.ai.mit.edu/classes/6.805/articles/privacy/sweeney-thesis-draft.pdf (visité le 26 août 2003). Note: version préliminaire d'un livre en instance de publication.

TAIRA R. K., BUI A. A. T. & KANGARLOO H. (2002). Identification of Patient Name References within Medical Documents Using Semantic Selectional Restrictions. In I. S. KOHANE, Ed., *Proceedings of the 2002 AMIA Annual Symposium*, p. 757–761: American Medical Informatics Association.