Multi-catégorisation de textes juridiques et retour de pertinence

Vincent Pisetta, Hakim Hacid et Djamel A. Zighed

article paru dans
G. Ritschard et C. Djeraba (eds), *Extraction et gestion des Connaissances (EGC 2006), Lille, janvier 2006*, numéro spécial de la *Revue des Nouvelles Technologies de l'Information, RNTI*, E-6, 235–246.

Multi-catégorisation de textes juridiques et retour de pertinence

Vincent Pisetta, Hakim Hacid, Djamel A. Zighed

Laboratoire ERIC – 5, av. Pierre Mendès-France- 69767 Bron- France
vpisetta@etu.univ-lyon2.fr,
hhacid@eric-univ.lyon2.fr,
zighed@univ-lyon2.fr

Résumé. La fouille de données textuelles constitue un champ majeur du traitement automatique des données. Une large variété de conférences, comme TREC, lui sont consacrées. Dans cette étude, nous nous intéressons à la fouille de textes juridiques, dans l'objectif est le classement automatique de ces textes. Nous utilisons des outils d'analyses linguistiques (extraction de terminologie) dans le but de repérer les concepts présents dans le corpus. Ces concepts permettent de construire un espace de représentation de faible dimensionnalité, ce qui nous permet d'utiliser des algorithmes d'apprentissage basés sur des mesures de similarité entre individus, comme les graphes de voisinage. Nous comparons les résultats issus du graphe et de C4.5 avec les SVM qui eux sont utilisés sans réduction de la dimensionnalité.

1 Introduction

Le cadre général de l'apprentissage automatique part d'un fichier d'apprentissage comportant n lignes et p colonnes. Les lignes représentent les individus et les colonnes les attributs, quantitatifs ou qualitatifs observés pour chaque individu ligne. Dans ce contexte, on suppose également que l'échantillon d'apprentissage est relativement conséquent par rapport au nombre d'attributs. Généralement la taille de l'échantillon est de l'ordre de 10 fois le nombre de variables pour espérer obtenir une certaine stabilité, c'est-à-dire une erreur en généralisation qui n'est pas trop loin de l'erreur en apprentissage. De plus, l'attribut à prédire est supposé à valeur unique. C'est une variable à valeurs réelles dans le cas de la régression et c'est une variable à modalités discrètes, appelées classes d'appartenance, dans le cas du classement. Ces questions relatives aux rapports entre taille d'échantillon et taille de l'espace des variables sont étudiées de façon très approfondies dans les publications relatives à l'apprentissage statistique (Vapnik, 1995). Dans ce papier nous décrivons une situation d'apprentissage qui s'écarte significativement du cadre classique tel que décrit plus haut. En effet, le contexte expérimental ne nous permet pas de disposer immédiatement d'un ensemble d'apprentissage conséquent, chaque individu peut appartenir à plusieurs classes simultanément, et chaque individu, au lieu d'être décrit par un ensemble attributs-valeurs, l'est par un texte en langage naturel en anglais.

Avant de décrire l'approche que nous préconisons pour apprendre dans ce contexte, nous allons tout d'abord rappeler la problématique de l'application visée (section 2). En section trois, nous décrivons l'approche méthodologique retenue. Dans la section quatre, nous décrivons les étapes mises en œuvre pour mettre en forme les données et notamment, la stratégie d'analyse linguistique mise en œuvre pour extraire les principaux concepts qui vont jouer le rôle de variables. Nous décrivons ensuite, section 5, les modèles topologiques à base de graphes de proximité qui nous permettent de gérer le multi-classes. Dans un but comparatif, nous utilisons une méthode à base d'arbre de décision qui nous sert également à mieux identifier les concepts discriminants. En section 6, nous présentons les résultats issus de l'analyse linguistique et de l'apprentissage. Nous décrivons également le principe de l'apprentissage par boucle de pertinence (relevance feedback). Ce concept est central car il met l'usager dans la boucle visant à améliorer le modèle de prédiction. Nous détaillons les performances obtenues. En section 7, nous concluons et détaillons les perspectives de ce travail, notamment l'utilisation de méthodes de règles d'association.

2 Cadre expérimental

2.1 Problématique

Ce travail s'inscrit dans un projet en collaboration avec le Bureau International du Travail (BIT). Plusieurs pays ont signé des conventions avec le BIT qui les lient au droit du travail international. Plus concrètement, l'accord porte sur deux conventions élaborées par le BIT, La Convention n°87 et la Convention n°98.[1] Celles-ci contiennent une série d'articles de lois que le signataire s'engage à respecter. Ces derniers sont soumis, une fois par an, à une inspection ayant pour but de vérifier la bonne application de ces conventions. A la fin de chaque inspection, les experts du BIT délivrent un rapport au pays concerné. Le rapport fait état des règles non appliquées, des violations constatées à partir de faits concrets et souligne les efforts à mettre en place pour être en adéquation avec les conventions. Il est en texte libre sans codification rigide des violations. Tous les rapports d'experts sont stockés dans une banque de données accessible aux membres et aux experts qui effectuent régulièrement des analyses, définissent de nouvelles recommandations et étudient les évolutions du droit du travail dans les différents pays, etc.

L'objectif de notre travail est définir et de mettre en place des méthodes et des outils de data mining permettant de traiter plus efficacement et plus rapidement ces corpus qui deviennent inexploitables manuellement. Les experts du BIT souhaiteraient avoir des outils permettant le repérage automatique des textes signalant la violation d'une ou plusieurs règles par pays. La finalité étant la catégorisation automatique des textes non étiquetés. Les experts pourront alors synthétiser plus vite les difficultés que rencontrent les différents pays dans l'application de ces conventions, et, le cas échéant, identifier les moyens de les aider.

[1]Le contenu de ces conventions ainsi que la liste des signataires est accessible sur http://www.ilo.org/ilolex/cgi-lex/convde.pl?C087 et http://www.ilo.org/ilolex/cgi-lex/convde.pl?C098

2.2 Description du corpus

La base de données du BIT comprend 1325 textes. Chaque texte, correspond à un commentaire annuel adressé par les experts du BIT au pays concerné. Chaque texte décrit les règles relatives à chaque convention qui ont été violées et les modalités de cette violation. Les textes sont rédigés en anglais par des juristes mandatés par le BIT. La banque de données constituée par le BIT comporte 834 textes signalant des violations relatives à la Convention N°87 et 481 pour la Convention N°98. Les textes sont classés par Convention, date, pays. On peut ainsi étudier l'évolution dans le temps des textes relatifs à un même pays sur chaque convention. Ces trajectoires constituent des points important pour le travail des experts. Notons pour clore sur ce corpus qu'un texte peut relater la violation de plusieurs règles. Précisons également que l'identification des règles violées s'effectue a posteriori par interprétation des observations et des commentaires des enquêteurs. Ce travail d'étiquetage s'effectue à l'heure actuelle par des experts juriste du BIT. Les raisons de délais d'interprétation et de coût associé ne facilitent pas l'exploitation des enquêtes menées. Il existe 17 violations pour la Convention n°87 et 10 pour la Convention n°98.

3 Méthodologie

3.1 Principes généraux

Pour mettre au point un outil d'identification des violations dans un corpus, nous avons procédé par des techniques d'apprentissage automatique. Le choix s'est porté sur les méthodes d'apprentissage supervisé qui produisent des modèles de prédiction qui une fois évalués et jugés acceptables par l'utilisateur, peuvent alors être utilisés comme moyen automatique de catégorisation pour les nouveaux textes. Pour aboutir à un modèle de prédiction, le principe consiste à fournir à l'algorithme d'apprentissage des exemples de textes pré-classés que nous appellerons ensemble d'apprentissage. Dans notre cas, il s'agit d'un problème à classes multiples. Autrement dit, chaque texte de l'échantillon d'apprentissage peut appartenir à plusieurs classes, chacune est identifiée à une règle d'une convention qui serait violée.

Plus formellement, si on note ω un texte du corpus global Ω et par $c_i \; i = 1,\ldots,k$ les règles de la convention v_a susceptibles d'être violées, alors $C(\omega) = \{c_1, c_2, c_5\}$ exprime que le texte ω comporte des violations relatives aux règles $\{c_1, c_2, c_5\}$ de la convention v_a. Notons que dans ce contexte, il est difficile et extrêmement coûteux d'effectuer cette catégorisation manuellement en une seule traite. Nous avons suggéré aux experts d'annoter une soixantaine de textes par convention (71 pour la convention n°87 et 65 pour la convention n°98) qui serviront de base d'apprentissage initiale. Appelons ce premier corpus d'amorçage Ω_1.

Soit Ψ un algorithme d'apprentissage supervisé. Cela peut être un graphe d'induction (Zighed et Rakotomalala, 2000), un SVM (Vapnik, 1995), etc. Le résultat d'un apprentissage est un modèle noté M et un taux d'erreur ε en généralisation estimé sur échantillon test ou par cross validation.

$$\Psi(\Omega, C) = (M, \varepsilon)$$

L'application du modèle M sur un échantillon de textes anonyme Ω' de taille relativement modeste, disons une vingtaine de cas, permet de prédire pour chaque individu anonyme ω' les règles qui seraient violées, $\mathrm{M}(\omega') = \{c_i, c_k\}$ par exemple. Le contrôle de pertinence permet à l'utilisateur d'évaluer chaque étiquetage sur les individus anonymes. L'usager U peut alors valider totalement ou partiellement l'étiquetage proposé par le modèle. Dans le cas où pour un texte ω' la prédiction est jugée erronée par U alors, le texte concerné est extrait et remis dans l'échantillon d'apprentissage $\Omega_{i+1} = \Omega_i \cup \omega'$. Cette opération étant renouvelée à chaque fois qu'un individu est jugé mal étiqueté. A la fin, nous réitérons le processus d'apprentissage avec le nouvel échantillon ainsi construit. Nous obtenons un nouveau modèle M' dont on espère un taux d'erreur ε' plus faible $(\varepsilon' < \varepsilon)$. Un nouvel échantillon anonyme de taille modeste pour faciliter une vérification manuelle est constitué. La réitération de ce processus de recyclage des individus mal étiquetés par le classifieur dans un nouvel apprentissage après rectification manuelle des étiquettes devrait conduire à une amélioration itérative du modèle.

La question qui se pose dès lors est le choix d'un algorithme capable de gérer l'étiquetage multiple. Le graphes de proximité (Toussaint (1980)) qui font partie des méthodes d'apprentissage à base d'instance permettent cela. Toutes ces techniques supposent par ailleurs que les individus sont plongés dans un espace de représentation sur lequel on peut définir une métrique. Les textes doivent par conséquent être transformés en un ensemble de vecteur. Chaque texte pourra être alors considéré comme un point de $\mathbb{R}^p$. Les coordonnées d'un texte ω dans cet espace seront $\mathrm{X}(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_p(\omega))$. Que représente alors ces variables, comment sont elles extraites ? C'est l'objet de la partie analyse linguistique. L'objectif étant de trouver les concepts et les plus adaptés.

3.2 Application sur les données du BIT

Nous effectuons l'extraction de la terminologie sur l'intégralité du corpus. La finalité de cette extraction est la construction de concepts relatifs aux deux Conventions. Les concepts ainsi extraits constituent l'espace de représentation des documents. Les textes étiquetés par les experts du BIT (violations connues), nous servent ensuite de base d'apprentissage. Nous utilisons deux classifieurs : C4.5 (Quinlan, 1993) et les graphes des voisins relatifs (GVR)(Toussaint, 1980) dans le but de prédire les violations contenues dans les textes non étiquetés. La méthodologie est décrite par (figure 1).

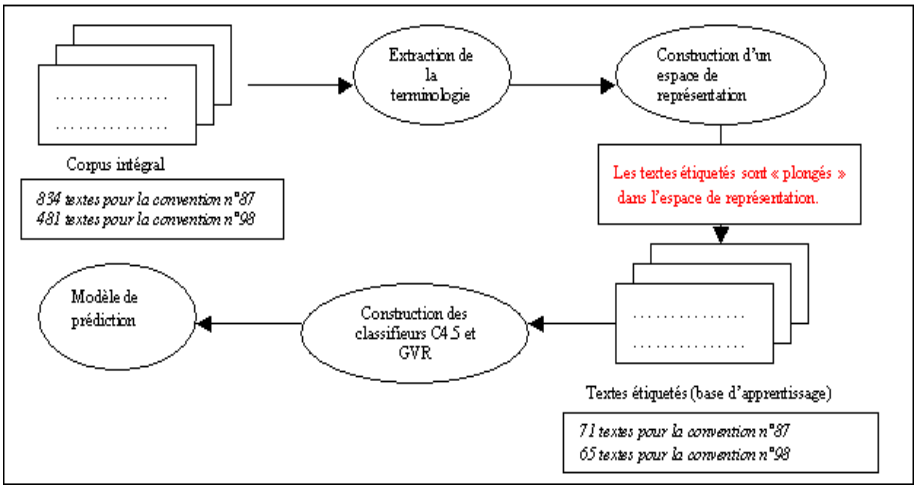

FIG. 1 – *Méthodologie d'analyse*

4 Espace de représentation des textes

4.1 Extraction de la terminologie

Nous choisissons de construire notre espace de représentation par extraction de concepts. Un des avantages de cette technique par rapport à des méthodes telles que les N-grammes, ou les matrices de co-occurrences de mots, est la réduction importante de la dimensionnalité, permettant notamment l'usage de classifieurs utilisant des mesures de similarité. Diverses applications basées sur ce principe ont données des résultats intéressants (Kumps et al., 2004). Deux méthodes différentes existent pour la construction de concepts : par apprentissage et par extraction.

La première (statistique) recherche les mots les plus discriminants selon un attribut à prédire. Les mots sont ensuite regroupés en concepts sur la base de leur co-occurrences ou à partir de règles d'association (Kumps et al., 2004). La seconde méthode (linguistique) consiste à extraire la terminologie du corpus et à regrouper les termes extraits selon leur proximité sémantique.

Notre préférence se porte vers les techniques d'analyse linguistique. Ce choix se justifie par le fait que l'analyse linguistique permet de lutter contre la polysémie et de lever certaines ambiguïtés liées au contexte (Flurh, 2000). Elle fonctionne également sur de petites unités textuelles (Pouliquen et al., 2002). De plus, notre base d'apprentissage comportant peu d'exemples, il nous semble difficile d'utiliser les techniques d'apprentissage décrites plus haut. Notre travail est effectué en collaboration avec des experts du domaine juridique, ce qui est une raison supplémentaire pour utiliser les techniques linguistiques. Nous utilisons la chaîne de traitement décrite en (figure 2).

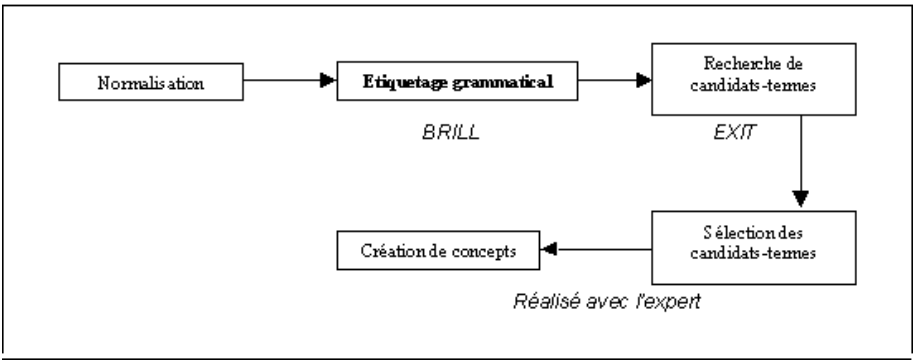

FIG. 2 – *Chaîne de traitement linguistique*

Après une phase de normalisation (traitement des noms propres, conservation ou non des majuscules, etc.), nous effectuons un étiquetage grammatical du corpus grâce au logiciel BRILL (Brill, 1995). Cette opération a pour but d'attribuer à chaque mot son étiquette grammaticale. Nous passons ensuite à l'extraction de la terminologie à l'aide d'EXIT (Roche et al., 2004). La méthodologie mise en place dans EXIT est avant tout basée sur une approche statistique, contrairement à d'autres approches (Bourigault et Jacquemin, 1999). L'extraction terminologique passe par la recherche de candidats-termes. Ces derniers sont des ensembles de deux ou plus unités adjacentes lexicales (mots), syntaxiques (mots étiquetés grammaticalement, ce qui est notre cas) ou sémantiques (mots étiquetés conceptuellement). Nous groupons ensuite les candidats-termes extraits selon leur proximité sémantique, de manière à repérer les concepts présents dans les textes.

4.1.1 Choix des candidats termes et création de concepts

Nous reprenons ici une méthodologie utilisée par (Baneyx et al., 2005). Nous distinguons deux étapes dans la sélection des candidats termes :

- Dans un premier temps, nous parcourons l'ensemble des résultats donnés par EXIT, et nous étudions tout d'abord les termes dont la fréquence d'apparition est supérieur à un seuil l. Dans un premier temps, nous fixons un seuil élevé. Cette étape préliminaire permet de repérer les grands axes conceptuels ;
- Nous regroupons ensuite les candidats termes sémantiquement proches à l'aide d'outils tels que WordNet (Miller et al., 2005). Le recours à l'expert est ici primordial. Certaines plates-formes proposent des outils plus sophistiqués comme Syntex-Upery (Bourigault et Jacquemin, 1999) qui permettent d'analyser la proximité distributionnelle entre les candidats-termes.

Les phases de construction 1 et 2 étant itératives, nous augmentons très rapidement la représentation en examinant les candidats-termes dont la fréquence d'apparition dans le corpus est inférieure au seuil l. L'augmentation progressive du nombre de candidats-termes possède deux issues :

- certains « nouveaux » candidats-termes viennent renforcer des concepts existants ;
- d'autres « nouveaux » candidats-termes créent de nouveaux concepts qui peuvent être des concepts fils de ceux existants.

4.2 Représentation vectorielle des documents

A ce niveau se pose le problème du choix du modèle de représentation. Nous avons choisi le modèle vectoriel (Salton, 1971) qui nous paraît plus adapté que le modèle booléen. La raison est qu'il semble simpliste d'appliquer une logique binaire à une recherche d'information. De plus, le modèle vectoriel permet de calculer des scores de similarités entre différents documents (Pouliquen et al., 2002).

Le modèle vectoriel propose de représenter un document sur les dimensions représentées par les mots. Nous l'avons adapté pour représenter un document par un vecteur de concepts. Et, plutôt que de le représenter en fonction de la fréquence du concept dans le document, nous utilisons la pondération TF x IDF (Salton et Buckley, 1988). Ce score permet de donner une importance au concept en fonction de sa fréquence dans le document (TF = Term Frequency) pondérée par la fréquence d'apparition du concept dans tout le corpus (IDF = Inverse Document Frequency). Ainsi un concept très spécifique au document aura un score correspondant à sa fréquence d'apparition, par contre, un concept apparaissant dans tous les documents du corpus aura une pondération maximale. Nous calculons donc, pour chaque concept dans un document, son score TF x IDF.

$$TF \times IDF_{X_i(\omega)} = TF_{X_i(\omega)} \log\left(\frac{n}{DF_{X_i}}\right)$$

Avec X_i un concept ; ω le document ; $TF_{X_i(\omega)}$: la fréquence absolue d'apparition du concept dans le document ω ; DF_{X_i} : le nombre de documents de la base d'apprentissage contenant le concept X_i ; n : le nombre de documents de la base d'apprentissage. A la fin de cette opération, nous disposons d'une base d'apprentissage que nous disposons sous forme tabulaire.

5 Modélisation et outils de généralisation

Nous avons à présent défini un espace de représentation pour les documents. L'objectif est maintenant d'utiliser des techniques d'apprentissage dans le but de classer automatiquement les documents. Dans notre étude, nous sommes amenés à classer des textes « multi-étiquettes », autrement dit, susceptibles de comporter plusieurs violations. Deux possibilités sont alors envisageables :

- une approche globale ;
- une approche binaire en une division en m sous-problèmes.

Nous effectuons les deux approches. Deux classifieurs différents sont utilisés. Nous utilisons les arbres de décision dans l'approche binaire et les graphes de voisinage dans l'approche globale.

5.1 Prédiction par le graphe des voisins relatifs

La représentation vectorielle de nos documents est d'une dimensionnalité très raisonnable et nous permet par conséquent d'avoir recourt à des classifieurs basés sur la notion de voisinage. Nous avons choisi les graphes de proximité provenant de la géométrie computationnelle (Preparata et Shamos, 1985) plutôt que les k-NN. Les graphes présentent plusieurs avantages par rapport aux k-NN et permettent de mieux définir la proximité entre des individus (Clech, 2004). Ils nécessitent une mesure de dissimilarité (Toussaint, 1980). Nous choisissons la distance euclidienne. Plusieurs modèles de graphes existent. Notre choix se porte sur le graphe des voisins relatifs qui est un bon compromis entre nombre de voisins et complexité algorithmique (en $O(n^3)$).

Dans le graphe des voisins relatifs $G_{rng}(\Omega,\varphi)$, deux points (α,β) de Ω^2 sont voisins si ils vérifient la propriété suivante. Soit $H(\alpha,\beta)$ l'hypersphère de rayon $\delta(\alpha,\beta)$ et centrée sur α, et $H(\beta,\alpha)$ l'hypersphère de rayon $\delta(\beta,\alpha)$ et centrée sur β. $\delta(\alpha,\beta)$ et $\delta(\beta,\alpha)$ sont des mesures de dissimilarité entre les deux points α et β. $\delta(\alpha,\beta) = \delta(\beta,\alpha)$. α et β sont voisins si et seulement si la lunule $A(\alpha,\beta)$ formée par l'intersection des deux hypersphères $H(\alpha,\beta)$ et $H(\beta,\alpha)$ est vide (Toussaint, 1980). De façon formelle:

$$A(\alpha,\beta) = H(\alpha,\beta) \cap H(\beta,\alpha)$$
$$(\alpha,\beta) \in \varphi \Leftrightarrow A(\alpha,\beta) \cap \Omega = \varnothing$$

Nous appliquons une fonction de décision simple. Un texte non étiqueté hérite des propriétés de ses voisins contenus dans la base d'apprentissage. Soit K le nombre de voisins du texte à étiqueter ω', c_i la ième règle. La probabilité que le nouveau texte ω' à étiqueter contienne une violation de c_i s'écrit :

$$P(c_i \mid \omega') = \frac{\textit{Nombre de voisins de } \omega' \textit{ contenant une violation de } c_i}{K}$$

L'application du GVR sur un texte anonyme ω' permet de prédire les règles qui seraient violées, par exemple :

$$GVR(\omega') = \{c_j, c_u\}$$

5.2 Prédiction par arbre de décision

Nous nous plaçons ici dans l'optique de prédire la présence ou l'absence de chaque violation. Nous construisons par conséquent autant d'arbres qu'il existe de règles. Plus formellement, nous considérons chaque règle comme étant un attribut booléen $c_i = \{0,1\}$. S'il existe k règles pour la violation v_α, nous construisons k arbres. Chaque arbre est alors un modèle M_i prévoyant la présence ou l'absence de chaque règle c_i. Nous obtenons ainsi k modèles qui renvoient c_i si la règle i est estimée violée, $\varnothing$ sinon. Notons que cette approche est valable dans la mesure où les violations sont a priori indépendantes.
Il suffit ensuite d'agréger les modèles pour obtenir un « méta-modèle » donnant la liste des violations détectées pour le texte ω'. La discrimination est effectuée par l'algorithme C4.5.

6 Résultats, méthodes et comparatif

A l'issue de l'analyse linguistique, nous obtenons 17 concepts pour la Convention n°87 et 11 concepts pour la convention n°98. Nous présentons les résultats observés sur la Convention n°98. Notre base de textes étiquetés est de taille modeste (65 textes). A ce jour, une étape du processus de relevance feedback a été réalisée. Elle concerne 20 textes qui ont été étiquetés par les experts du BIT et qui n'étaient pas présents initialement dans la base d'apprentissage. Nous présentons les résultats de la prédiction sur ces 20 textes issue de C4.5 et GVR. Dans un but comparatif, nous avons utilisé les SVM selon le même principe que pour C4.5. Les SVM sont des méthodes robustes résistant très bien à la forte dimensionnalité des données (Joachim, 1998). La différence essentielle réside dans le fait que les SVM sont utilisées sans pré-traitement des textes (excepté la normalisation). Les résultats sont présentés en 6.1.

6.1 Résultats obtenus

Nous présentons les résultats obtenus en terme de reclassement. Les résultats sont décrits dans (tableau 1).

On observe de bon taux de reclassement. Notons qu'il n'existe qu'une seule violation pour laquelle SVM fait mieux que C4.5 ou GVR. La non prise en compte de séquences de mots par SVM rend les prédictions parfois instables, ce qui se traduit par une mauvaise sensibilité ou spécificité. Nous observons des taux de sensibilité-spécificité parfois nuls pour GVR. Ceci est dû au fait que deux des dix violations (n°4 et n°10) sont peu fréquemment rencontrées dans le corpus d'apprentissage. Ainsi, il y a peu de chances que les quelques textes contenant ces violations soient en nombre suffisant pour être pris en compte dans le voisinage de l'individu à étiqueter. Ce problème peut éventuellement se résoudre par la technique de retour pertinent décrite précédemment.

	C4.5			GVR			SVM		
	bien classés	sensibilité	spécificité	bien classés	sensibilité	spécificité	bien classés	sensibilité	spécificité
Violation 1	0.9	0.9	0.9	0.9	0.9	0.9	0.85	1	0.7
Violation 2	1	1	1	1	1	1	1	1	1
Violation 3	1	1	1	1	1	1	1	1	1
Violation 4	0.85	0.5	0.94	0.8	0	1	0.95	0.75	1
Violation 5	1	1	1	0.9	1	0.85	0.9	1	0.85
Violation 6	0.8	0.8	0.8	1	1	1	0.5	0.6	0.4
Violation 7	0.85	0.875	0.83	0.8	0.75	0.83	0.9	1	0.83
Violation 8	0.7	0.66	0.73	0.8	0.81	0.77	0.6	0.56	0.64
Violation 9	0.85	0.88	0.82	0.85	0.77	0.83	0.7	0.6	0.72
Violation 10	1	1	1	0.9	0.33	1	0.9	1	0.88

TAB. 1 – *Résultats des trois méthodes de classification*

7 Conclusion et perspectives

La finalité de ce travail est de proposer un modèle de prédiction capable de déterminer les violations de plusieurs pays concernant deux convention de droit du travail. Une approche d'apprentissage automatique a été adoptée. Dans un premier temps (préparation des données), nous avons extrait, grâce aux techniques d'analyses linguistiques, un ensemble de candidats termes qui nous permettent ensuite de construire des concepts relatifs au corpus étudié. Cette opération a pour but de réduire la dimensionnalité de l'espace de représentation des textes du corpus. Nous avons été ainsi en mesure d'utiliser les graphes de voisinage, en plus d'une méthode plus classique (C4.5) pour la catégorisation automatique.

Les résultats semblent intéressants dans la mesure où les deux méthodes de prédiction que nous utilisons aboutissent à des taux de reclassement tout à fait acceptables en dépit d'une base d'apprentissage comportant peu d'exemples. Nous envisageons à présent d'augmenter la taille de celle-ci dans le but d'améliorer la prédiction et d'aboutir à des résultats plus robustes. La phase de test avec les experts du BIT est en cours. La liste des concepts extraits du corpus a été validée par ces derniers.

L'une des perspectives de ce travail est d'observer l'impact du relevance feedback sur la qualité de prédiction. En effet, cette dernière devrait augmenter au fur et à mesure du nombre d'interventions des experts. De plus, il serait intéressant de comparer de nouveau la qualité de prédiction de notre approche avec les SVM lorsque la base d'apprentissage sera plus conséquente. L'utilisation d'autres techniques de catégorisation textuelles, comme Winnow (Dagan et al., 1997) et éventuellement d'autres classifieurs peut aussi s'avérer intéressantes.

Références

Baneyx, A., J. Charlet, et M. C. Jaulent (2005). *Construction d'ontologies médicales fondée sur l'extraction terminologique à partir de textes : application au domaine de la pneumologie*. Journées Francophones d'Informatique Médicale (JFIM) 2005, 1-6.

Brill, E. (1995). *Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging*. Computational Linguistics, 21(4):543-565.

Bourigault, D. et C. Jacquemin (1999). *Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology* . Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'99), Bergen, pages 15-22.

Clech, J. (2004). *Contribution Méthodologique à la Fouille de Données Complexes*. Thèse. Laboratoire ERIC, Université Lumière Lyon 2.

Dagan, I., Y. Karov, and D. Roth (1997). *Mistake-Driven Learning in Text Categorization*. Proceedings of the Second Conference on Empirical Methods in NLP, 55-63.

Fluhr, C. (2000). *Indexation et recherche d'information textuelle*. Ingénierie des langues. Paris: Hermès.

Gaines, M. et B. Shaw (1989). *Comparing Conceptual Structures: Consensus, Conflict, Correspondence and Contrast*. Knowledge Science Institute, University of Calgary.

Hajek, P., T. Havranek and M. Chytil (1983). *GUHA Method*. Prague: Academia.

Joachim, T. (1998). Text categorization with support vector machines with many relevant *features*. Proceedings of ICML-99, 16th International Conference on Machine Learning (Bled, Slovenia, 1999), 200-209.

Kumps, N., P. Francq, et A. Delchambre (2004). *Création d'un espace conceptuel par analyse de données contextuelles*. JADT 2004, (International Conference on Statistical Analysis of Textual Data), 683-691.

Miller, A. G., C. Fellbaum, R. Tengi, S. Wolff, P. Wakefield, H. Langone, and B. Haskell (2005). *WordNet 2.1*, Cognitive Science Laboratory, Princeton University.

Pouliquen, B., D. Delamarre, et P. Le Beux (2002). *Indexation de textes médicaux par extraction de concepts, et ses utilisations*. JADT 2002, 617-627.

Preparata, F., et M. Shamos (1985). *Computational Geometry An Introduction*. New-York: Springer.

Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. San Mateo, CA.: Morgan Kaufman

Roche, M., T. Heitz, O. Matte-Tailliez, and Y. Kodratoff (2004). *EXIT: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés*. Proceedings of JADT 2004 (International Conference on Statistical Analysis of Textual Data), volume 2, 946-956.

Salton, G. (1971) *The SMART retrieval system. Experiment in automatic document processing*. New Jersey : Prentice Hall. Englewood Cliffs.

Salton, G. et C. Buckley (1988). *Term weighting approaches in automatic text retrieval*. Information Processing and Management, (l) 24, n° 5, 513-523.

Toussaint, G.T. (1980). *The relative neighborhood graphs in a finite planar set*. Pattern recognition, 12:261–268.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Zighed, D. A. et R. Rakotomalala (2000) *Graphe d'induction et Data Mining*. Paris: Hermès.

Summary

Text retrieval is an important field in automatic information and natural language processing. A large variety of conferences, like TREC, are dedicated to it. In this paper, we are interested in text retrieval and especially the automatic juridical texts classification as a way to facilitate their retrieval. We use linguistic tools (terminology extraction) in order to determine concepts presents in the corpus. Those concepts aim to create a low dimensionality space which enabling us to use automatic learning algorithms based on similarity measures as proximity graphs. We compare results extracted from the graph and from C4.5 with SVM which are used without reducing dimensionnality.