UNIVERSITÉ LUMIÈRE LYON2
Année 2003

THÈSE
pour obtenir le grade de
DOCTEUR
en
INFORMATIQUE

présentée et soutenue publiquement par

Radwan JALAM
le 4 juin 2003

Apprentissage automatique et catégorisation de textes multilingues

---

préparée au sein du laboratoire ERIC
Equipe de Recherche en Ingénierie des Connaissances

sous la direction de
Jean-Hugues CHAUCHAT

DEVANT LE JURY, COMPOSÉ DE :

Annie MORIN, Rapporteur	Maître de conférences habilitée, IRISA, Rennes
Yves KODRATOFF, Rapporteur	Directeur de recherche, CNRS, LRI Orsay
Martin RAJMAN, Rapporteur	Professeur, Ecole Polytechnique Fédérale, Lausanne
Geneviève BOIDIN-LALLICH, Examinateur	Professeur, Université Claude Bernard-Lyon 1
Ludovic LEBART, Examinateur	Directeur de recherche, CNRS, ENST Paris
Jean-Hugues CHAUCHAT, Directeur de thèse	Professeur, Université Lumière-Lyon 2

Table des matières

Conclusion et perspectives

Dans ce travail, nous avions comme objectif l'adaptation des techniques de l'apprentissage automatique au problème de la catégorisation de textes multilingues.

Nous présentons ici un cadre général pour catégoriser automatiquement des textes multilingues.

La catégorisation de textes dite "monolingue" suit habituellement le schéma suivant :

– **représentation des textes** dans un format adapté aux algorithmes d'apprentissage ; on utilise souvent la représentation vectorielle proposée par [Salton and McGill, 1983] mais d'autres modes de représentation existent aussi tels la représentation probabiliste (modèle de *Robertson et Sparck-Jones* [Robertson and K.Sparck-Jones, 1976] ou le modèle *Okapi* [Robertson et al., 1996]), où l'information concernant la position de termes dans les phrases peut être conservée, contrairement à ce que fait le modèle vectoriel de [Salton and McGill, 1983]. La représentation englobe trois éléments : le choix des termes, le choix des poids associés et le choix des méthodes de sélection de termes. Nous apportons deux contributions à ce sujet :
  – pour le choix de ce qui est un terme, nos travaux [Jalam and Chauchat, 2002] montrent les raisons de l'efficacité des n-grammes comme méthode de représentation. Ces travaux ont montré par exemple que les n-grammes, comme choix de représentation, capturent les connaissances contenues dans les mots. Ainsi, à partir des n-grammes, nous sommes capables d'extraire automatiquement des candidats mots-clés, sans utiliser aucune connaissance linguistique ; ceci est intéressant car l'indépendance de la méthode par rapport à la langue est nécessaire pour traiter les textes écrits en plusieurs langues.
  – pour la sélection de termes, nos travaux [Clech et al., 2003] proposent une nouvelle utilisation de la statistique du χ^2 (multivarié) calculée sur le tableau complet (termes$\times$classes). Cette méthode prend en compte l'interaction entre les termes eux-même, et entre les termes et les classes. Nos premiers résultats

sont encourageants et montrent une amélioration des performances par rapport à l'utilisation du χ^2 (univarié), utilisé pour la sélection des descripteurs par les auteurs précédents comme [Schütze et al., 1995, Wiener et al., 1995, Yang and Pedersen, 1997, He et al., 2000] ; ce χ^2_{uni} mesurait l'écart à l'indépendance entre un seul descripteur t_k (présent ou absent) et un seul thème c_i (présent ou absent).

– **choix d'une méthode d'apprentissage** pour construire un modèle de prédiction. Il s'agit de proposer une fonction $\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{V, F\}$ qui associe une ou plusieurs étiquettes (catégories) à un document d_j telle que la décision donnée coïncide « le plus possible » avec la fonction $\breve{\Phi} : \mathcal{D} \times \mathcal{C} \rightarrow \{V, F\}$, la vraie fonction qui retourne pour chaque vecteur d_j la valeur c_i de sa classe réelle. Plusieurs critères de choix sont possibles : si les résultats du classifieur sont destinés à des humains (experts ou décideurs), il est souhaitable de privilégier les méthodes "explicatives" comme les arbres de décision ; si, au contraire, le résultat produit est à intégrer dans un processus automatique, on peut alors choisir la méthode qui donne les meilleures performances.
  Dans [Jalam and Teytaud, 2001] et [Teytaud and Jalam, 2001] nous avons proposé de nouvelles utilisations des méthodes SVM et RBF en introduisant un nouveau noyau, fondé sur le χ^2, qui donne de bons résultats et qui confirment les résultats obtenus par [Yang, 1999, Sebastiani, 2002].
– **évaluation le modèle** appris afin de s'assurer qu'il est généralisable à d'autres textes.

Nous avons ensuite étendu la catégorisation aux **textes multilingues**. Dans ce cas, les méthodes utilisant des analyses linguistiques fines deviennent impraticables. Nous avons proposé une méthode générale, automatique et largement indépendante des langues.

La phase d'apprentissage s'effectue toujours de manière classique, à partir d'un corpus d'apprentissage étiqueté, rédigé dans une langue donnée. Pour classer un texte rédigé dans une langue quelconque, il faut d'abord identifier automatiquement la langue utilisée ; ensuite, il y a deux voies possibles (voir figure 9.1 page 119) :

– soit on applique un modèle propre à chaque langue ; ceci exige de disposer d'ensembles d'apprentissage (préalablement étiquetés manuellement) suffisamment vastes et variés dans chaque langue, ce qui est souvent hors de portée ;
– soit on utilise des traducteurs automatiques vers une langue fixe (disons l'anglais), puis on construit (par apprentissage automatique) un modèle unique de catégorisation. Il y a encore deux façons de construire le modèle, en fonction du moment où l'on place l'étape de traduction, on apprend
  – soit sur des textes écrits en anglais,
  – soit sur les traductions vers l'anglais de textes écrits dans différentes langues.

Nous avons proposé trois schémas et nous en avons expérimenté deux.

Les perspectives

La thèse de doctorat n'est qu'une étape dans la vie d'un chercheur. A l'issue de ce travail, je vois de nombreuses pistes qui restent à explorer.

Expérimentation de l'apprentissage direct sur les résultats des traductions automatiques

On peut sans doute améliorer le modèle de classement en estimant ses paramètres sur les résultats du traducteur automatique pour les textes de l'échantillon d'apprentissage, et non pas sur les textes originaux écrits dans la langue cible. On exploiterait ainsi jusqu'aux erreurs du traducteur. C'est la dernière méthode que nous avons proposée, sans avoir encore eu le temps de l'expérimenter.

Adaptation du corpus CLEF pour en faire un jeu d'essai multilingue étiqueté

La catégorisation de texte multilingue est un domaine nouveau qui nécessite des corpus de référence pour tester les algorithmes. Après de nombreuses discussions avec des spécialistes en catégorisation de textes et dans la recherche documentaire multilingues (parmi lesquels David Lewis, Fabrizio Sebastiani, Douglas Oard, Carol Peters, et bien d'autres), nous avons conclu à l'absence de tels corpus. Nous avons donc créé un corpus multilingue à partir des collections proposées aux concurrents du concours CLEF de l'année 2002 (*The Cross-Language Evaluation Forum*, voir : `http://clef.iei.pi.cnr.it`). Les responsables du concours CLEF sont intéressés par le nouveau domaine d'application que je propose. Je vais donc développer une coopération avec eux pour contribuer à étendre CLEF à la classification multilingue.

Échantillonnage dans les textes

L'échantillonnage dans les textes est à explorer : au lieu de travailler sur la totalité d'un texte, on peut se limiter à travailler sur une partie. Actuellement, il y a peu de recherches sur ce sujet. Au moins deux problèmes se posent :

1. la détermination de "l'individu statistique" à échantillonner : un N-grammes, un mot, une phrase, un paragraphe, *etc.*
2. la méthode d'échantillonnage : échantillonnage aléatoire simple, ou bien surreprésentation de certaines parties des textes comme les titres, les résumés, les introductions, les conclusions, *etc.*

La recherche dans cette voie peut être fructueuse car un échantillon bien choisi peut être porteur de presque toute l'information du texte initial, et l'échantillonnage devrait accélérer les traitements (si la procédure d'échantillonnage elle-même n'est pas trop longue).

Applications à la veille technologique

La veille est l'une des clefs de succès des entreprises, comme des centres de recherche. Le problème central de la *veille* est de détecter *l'information cruciale* le plus tôt possible. La fouille de textes (ou *Text Mining*) est l'un des outils permettant d'analyser rapidement une grande quantité d'information. La nature multilingue des textes à analyser implique la généralisation des méthodes qui se développent en ce moment. La veille est particulièrement difficile car il faut détecter des « signaux faibles » dans un océan de « bruit ». Il y a donc un défi à relever pour les méthodes automatiques.

Approfondissement et généralisation du LSI (Latent Semantic Indexing) avec l'ensemble des méthodes d'analyse factorielle

Il nous semble que les techniques connues sous le nom de « LSI » peuvent être améliorées en utilisant l'ensemble des méthodes statistiques d'analyse des données multidimensionnelles, souvent mal connues de la communauté de « l'apprentissage automatique ». C'est l'une des voies à explorer.

Application en biologie sur le code génétique

Le code génétique est une sorte de texte, composé avec un alphabet de 4 caractères. Les codes des individus, ou des espèces sont comme de très longs textes, avec quelques fautes d'orthographe. Les méthodes d'analyse automatique de textes à partir du codage en n-grammes peuvent donc s'appliquer à la recherche sur les séquences génétiques.