

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, Examineur	Professeur, Université Claude Bernard-Lyon 1
Ludovic LEBART, Examineur	Directeur de recherche, CNRS, ENST Paris
Jean-Hugues CHAUCHAT, Directeur de thèse	Professeur, Université Lumière-Lyon 2

Table des matières

Introduction	1
I. Catégorisation de textes monolingues	5
1. Catégorisation de textes	7
1.1. Introduction	8
1.2. Définition de la catégorisation de texte	8
1.3. Comment catégoriser un texte ?	9
1.3.1. Représentation, le codage, des textes	10
1.3.2. Choix de classifieurs	11
1.3.3. Évaluation de la qualité des classifieurs	12
1.4. Applications de la catégorisation de texte	12
1.4.1. Catégorisation de textes : une fin en soi	13
1.4.2. Catégorisation de textes : un support pour différentes appli- cations	13
1.5. Difficultés particulières de la catégorisation de textes	13
1.5.1. Grandes dimensions	14
1.5.2. Imprécision des fréquences	15
1.5.3. Déséquilibre	15
1.5.4. Ambiguïté	15
1.5.5. Synonymie	15
1.5.6. Subjectivité de la décision	16
1.6. Lien avec la recherche documentaire	16
1.7. Jeu de données utilisé pour l'évaluation	18
1.8. Conclusion	19
2. Approches pour la représentation de textes	21
2.1. Introduction	22

2.2.	Choix de termes	22
2.2.1.	Représentation en « sac de mots »	22
2.2.2.	Représentation des textes par des phrases	23
2.2.3.	Représentation des textes avec des racines lexicales et des lemmes	24
2.2.4.	Méthodes basées sur les n-grammes	24
2.3.	Codage des termes	26
2.3.1.	Codage $TF \times IDF$	27
2.3.2.	Codage TFC	27
2.4.	Réduction de la dimension	28
2.4.1.	Réduction locale de dimension	29
2.4.2.	Réduction globale de dimension	29
2.4.3.	Sélection de termes	29
2.4.4.	Extraction de termes	30
2.5.	Conclusion	30
3.	Sélection multivariée de termes	33
3.1.	Introduction	34
3.2.	Méthode du χ^2 univariée	34
3.3.	Méthode du χ^2 multivarié	36
3.4.	Expérimentation	36
3.5.	Conclusion	38
4.	Pourquoi les n-grammes fonctionnent	39
4.1.	Introduction	40
4.2.	Intérêt du codage en n-grammes	40
4.3.	Étapes de la recherche des mots caractéristiques	41
4.3.1.	Recherche des n-grammes caractéristiques et des mots qui les contiennent	41
4.3.2.	Filtrage des mots « parasites »	42
4.3.3.	Algorithme complet	42
4.4.	Exemple d'application	42
4.4.1.	Données indexées de Reuters	42
4.4.2.	Quelques résultats	44
4.4.3.	Discussion des résultats sur la collection Reuters	44
4.5.	Conclusion	46
5.	Techniques pour la construction de classifieurs	51
5.1.	Introduction	52
5.1.1.	Manière de construction du classifieur	52
5.1.2.	Caractéristique du modèle	53
5.2.	Méthode de Rocchio	53
5.3.	Arbres de décision	55

5.3.1.	Phase d'apprentissage	56
5.3.2.	Phase de classification	61
5.3.3.	Critiques de la méthode	61
5.4.	Classifieurs à base d'exemples	62
5.4.1.	K-plus proches voisins	63
5.5.	Fonctions à bases radiales	67
5.6.	Machine à Vecteurs de Support	68
5.6.1.	Cas des classes linéairement séparables	69
5.6.2.	Cas des classes non séparables	70
5.7.	Évaluation de classifieurs de textes	71
5.7.1.	Évaluation des classifieurs, l'approche « binaire »	72
5.7.2.	Évaluation des classifieurs, l'approche « multi-classes »	76
5.8.	Contributions personnelles	77
5.8.1.	Nouvelle utilisation des SVM	77
5.8.2.	Nouvelle utilisation des réseaux RBF	77
5.8.3.	Nos expérimentations	77
5.9.	Conclusion : quel est le meilleur classifieur ?	79
 II. Catégorisation de textes multilingues		83
 6. Catégorisation multilingue : les solutions proposées		85
6.1.	Introduction	86
6.2.	Intérêt accru aux traitements multilingues	86
6.2.1.	Davantage de collections numériques	86
6.2.2.	Plus de personnes connectées en ligne	87
6.2.3.	Plus de globalisation et de pays unifiés	87
6.2.4.	Réseau plus rapide et plus souple	88
6.3.	Recherche documentaire multilingues	88
6.3.1.	Approches basées sur la traduction automatique	89
6.3.2.	Thésaurus multilingues	90
6.3.3.	Utilisation de dictionnaires	90
6.4.	Nos solutions pour catégoriser des textes multilingues	91
6.4.1.	Premier schéma : le schéma trivial	92
6.4.2.	Deuxième schéma : choisir une seule langue d'apprentissage	93
6.4.3.	Troisième schéma : mélanger les ensembles d'apprentissage	93
6.5.	Conclusion	94
 7. Identification de la langue		97
7.1.	Introduction	98
7.2.	Approches linguistiques	99
7.2.1.	Présence de certains chaînes de caractères spécifiques	99
7.2.2.	Présence de certains mots	100

7.2.3. Approche lexicale	101
7.2.4. Approche plus linguistique	101
7.3. Approches statistiques et probabilistes	102
7.3.1. Mots les plus fréquents	102
7.3.2. Méthodes basées sur les n -grammes	103
7.4. Expériences pour la reconnaissance de langue	107
7.5. Conclusion	109
8. Traduction automatique	111
8.1. Introduction	112
8.2. Premières approches historiques de la traduction automatique	112
8.2.1. Décryptage	112
8.2.2. Analyse par micro-contexte	112
8.2.3. Imiter la traduction humaine	113
8.3. Nouvelles approches, plus modestes	113
8.3.1. Mémoire de traduction	113
8.3.2. Sous-langages et langages contrôlés	114
8.4. Évaluer la traduction automatique	114
8.5. Conclusion	115
9. Cadre pour la catégorisation de textes multilingues	117
9.1. Introduction	118
9.2. Méthodes pour la catégorisation de textes multilingues	118
9.2.1. Nouveau cadre pour la catégorisation multilingue	118
9.2.2. Détection de la langue du texte à classer	119
9.2.3. Traduction du texte à classer	119
9.3. Application sur les corpus CLEF	120
9.3.1. Constitution du corpus	120
9.3.2. Représentation des textes	122
9.3.3. Algorithmes d'apprentissage	123
9.3.4. Reconnaissance de la langue	123
9.3.5. Catégorisation des articles	123
9.4. Discussion	128
9.5. Conclusion	129
Conclusion et perspectives	133
Index des auteurs cités	137
Bibliographie	141

Conclusion et perspectives

Dans ce travail, nous avons 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_{uni}^2 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 $\check{\Phi} : \mathcal{D} \times \mathcal{C} \rightarrow \{V, F\}$, 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 sur-représentation 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.