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

Introduction

La recherche accorde, ces dernières années, beaucoup d'importance au traitement des données textuelles [Lebart, 2001, Kodratoff, 2001] et en particulier aux données multilingues. Ceci pour plusieurs raisons : un nombre croissant de collections mises en réseau et distribuées au plan international, le développement de l'infrastructure de communication et de l'Internet, la progression constante du nombre de personnes connectées au réseau mondial et dont la langue maternelle n'est pas l'anglais [Peters and Sheridan, 2001]. Ceci a créé de nouveaux besoins pour organiser et traiter ces immenses volumes de données. Les traitements manuels de ces données (systèmes experts, ou à base de connaissances) s'avèrent très coûteux en temps et en personnel, ils sont peu flexibles et leur généralisation à d'autres domaines sont quasi impossibles ; c'est pourquoi on cherche à mettre au point des méthodes automatiques [Moulinier, 1996, Sebastiani, 2002].

Dans ce travail nous nous intéressons à l'utilisation de l'apprentissage automatique pour catégoriser (ou classer) les textes. L'apprentissage automatique est un processus d'induction général qui permet la construction automatique de classifieurs [Mitchell, 1997]. Ici, il s'agit d'affecter une ou plusieurs catégories à des documents : l'objectif est de trouver une liaison fonctionnelle, que l'on appelle également modèle de prédiction, entre les textes à classer et l'ensemble des catégories. Pour estimer le modèle de prédiction, il faut disposer d'un ensemble de textes préalablement étiquetés, dit ensemble d'apprentissage, à partir duquel on estime les paramètres du modèle de prédiction le plus performant possible, c'est-à-dire qui produit le moins d'erreurs en prédiction.

L'objectif de la catégorisation de texte est donc d'associer automatiquement une étiquette à tout nouveau texte à classer. Une application typique est le classement automatique de dépêches de presse en différents thèmes (par exemple, "les actualités du monde", "l'économie", "les sports", *etc.*), à l'aide de leurs contenus textuels [Lewis, 1992a].

Dans ce travail, nous étendons la catégorisation de textes à la catégorisation de textes *multilingues*. La nouveauté apportée est la possibilité d'inférer pour un texte ré-

digé dans une langue quelconque. En reprenant l'exemple précédent, nous apprenons à sélectionner, parmi l'ensemble des dépêches, celles qui concernent l'*économie* ; ceci quelque soit leur langue.

Comme dans le cas monolingue habituel, la phase d'apprentissage s'effectue à partir d'un corpus d'apprentissage étiqueté. Bien entendu, comme pour la catégorisation de textes monolingue, le texte à classer doit appartenir aux mêmes domaines que ceux utilisés lors de l'apprentissage. On ne saurait, par exemple, essayer de classer un article scientifique à partir d'un modèle construit sur un ensemble d'apprentissage constitué d'articles de journaux de mode [Sebastiani, 2002].

Cette extension au cas multilingue introduit des contraintes supplémentaires. Il faut adapter le processus habituellement mis en œuvre pour classer les nouveaux textes ; et certaines techniques à base linguistique, utilisées en monolingue, deviennent alors inopérantes [Biskri and Delisle, 2001].

Plan de la thèse

Notre travail s'intéresse à l'application de méthodes issues de l'apprentissage automatique à la catégorisation de textes multilingues. Il comporte deux parties.

Une première partie donne une présentation générale de la catégorisation de textes :

- les définitions, objectifs généraux et domaines d'application (chapitre 1 page 7).
- l'adaptation des algorithmes d'apprentissage aux spécificités des textes ; représentation vectorielle d'un texte ; choix des unités d'information (mots, lemme, phrase, n-grammes, *etc.*) ; choix des valeurs numériques associées (présence/absence, fréquences, information mutuelle, *etc.*) ; réduction de l'espace de représentation, par sélection de termes, ou bien par extraction (chapitre 2 page 21).
- la méthode de sélection de termes multivariée (chapitre 3 page 33).
- le codage en n-grammes (chapitre 4 page 39) :
  - les avantages et inconvénients des n-grammes, en particulier dans le contexte multilingue,
  - le lien entre les n-grammes et les mots.
- les méthodes d'apprentissage et la mesure de leurs performances (chapitre 5 page 51).
- les tests réalisés pour comparer les algorithmes d'apprentissage sur les textes monolingues : les méthodes des "k-plus proches voisins", des machines à vecteurs de support (SVM), les fonctions à bases radiales (RBF) et les graphes d'induction (sections 5.8.3 page 77 et 7.4 page 107).

La deuxième partie s'intéresse à l'apprentissage de textes multilingues en comparant deux chaînes possibles (chapitres 6 page 85 et 9 page 117) :

- **chaîne 1 :** reconnaissance de la langue, puis utilisation de règles de classement construites pour chaque langue ; il faut alors avoir construit un modèle adapté à chacune des langues.

– **chaîne 2 :** utilisation de la traduction automatique dans le processus de catégorisation ; cette solution permet d'utiliser un seul ensemble de règles de classement. Ici, il y a deux options :

  1. construire un modèle unique sur l'ensemble d'apprentissage d'une langue donnée ; ensuite, pour classer un nouveau texte, (i) reconnaissance de sa langue, (ii) traduction de ce texte vers la langue d'apprentissage, (iii) application du modèle de prédiction sur le texte traduit ; ici la phase de traduction n'intervient que dans la phase de classement.
  2. faire intervenir la traduction automatique dès la phase d'apprentissage : à partir d'un ensemble étiqueté de textes en différentes langues, traduction automatique de tous ces textes vers une langue cible et apprentissage sur cet ensemble de textes traduits ; ensuite, pour classer un nouveau texte, la procédure est la même qu'en 1.

Ces deux chaînes dépendent de la qualité du traducteur automatique (chapitre 8 page 111).

Nous testons nos algorithmes sur des corpus multilingues pour pouvoir évaluer et comparer la performance de ces deux chaînes de traitement (section 9.3 page 120). Nos principales contributions portent sur :

– l'exploration des raisons de l'efficacité de la représentation des textes par le codage en n-grammes ; ce travail a été présenté aux JADT-2002 [Jalam and Chauchat, 2002] ; le chapitre 4 page 39 développe ce travail.
– la proposition d'une nouvelle méthode de sélection de termes multivariée pour l'apprentissage basées sur la statistique χ^2, publiée dans SFdS-2003 [Clech et al., 2003] ; cette proposition est détaillée dans le chapitre 3 page 33.
– une nouvelle utilisation des SVM et des RBF avec le χ^2 comme noyau, publiée dans EGC-2001 et IJCNN-2001 [Jalam and Teytaud, 2001, Teytaud and Jalam, 2001] ; les sections 5.8.3 page 77 et 7.4 page 107 présentent les résultats auxquels nous sommes parvenus avec ces méthodes.
– la proposition d'un nouveau cadre pour la catégorisation de textes multilingues, dans les chapitres 6 page 85 et 9 page 117.
– l'adaptation du corpus CLEF pour disposer d'un échantillon de textes multilingues étiquetés en vue de l'apprentissage ; ce corpus adapté est, à notre connaissance, le premier jeu de données dans ce domaine. Il est présenté au chapitre 9 page 117.