Chapitre 1 Introduction

1.1 Problématique générale

En raison de l'augmentation constante du volume d'information accessible électroniquement, la conception et la mise en œuvre d'outils efficaces, permettant notamment à l'utilisateur de n'avoir accès qu'à l'information qu'il juge pertinente, devient une nécessité absolue. Comme la plupart de ces outils sont destinés à être utilisés dans un cadre professionnel, les exigences de fiabilité et de convivialité sont très importantes ; les problèmes à résoudre pour satisfaire ces exigences sont nombreux et difficiles. L'accès à cette information pertinente peut se faire en fournissant à un utilisateur des documents pertinents ou en lui proposant des passages de documents pertinents (ou des réponses à des questions). Le premier cas relève du domaine de la *recherche de textes* et le second du domaine de l'*extraction d'informations*.

C'est dans le domaine très actif de la recherche de textes que se place le présent travail, réalisé dans le cadre d'une collaboration entre Informatique CDC, filiale de la Caisse des Dépôts et Consignations, et le Laboratoire d'Électronique de l'ESPCI.

1.2 La recherche de textes

Le domaine de la recherche de textes se divise en deux grandes disciplines :

- la recherche d'informations (également appelée recherche *ad hoc*)
- la catégorisation de textes.

La recherche d'informations consiste à trouver, dans une importante base de documents, les documents pertinents correspondant à des requêtes *ad hoc* (par mots clefs) ou posées en langage naturel. L'utilisation de moteurs de recherches sur le web et la recherche informatisée de documents dans un fonds bibliothécaire sont deux exemples d'applications de ce domaine.

La recherche d'informations est généralement effectuée en indexant préalablement tous les documents de la base selon les mots qu'ils contiennent ; la recherche consiste à trouver, le plus rapidement possible, les documents ayant des mots communs avec la requête de l'utilisateur. On montrera au chapitre 2 comment certaines méthodes modifient automatiquement la requête initiale pour améliorer le résultat de la recherche.

La catégorisation de textes, appelée également filtrage, consiste à trouver, dans un flux de documents (comme un fil de dépêches d'agence de presse), ceux qui sont relatifs à un sujet défini par avance. L'une des applications consiste à fournir à un utilisateur, en temps réel, toutes les informations importantes pour l'exercice de son métier. Dans ce cas, l'utilisateur n'exprime pas son intérêt par une requête, mais par un ensemble de documents pertinents. Cet ensemble de documents pertinents définit ce que l'on appelle, dans la suite de ce mémoire, un *thème* ou une *catégorie*. Pour un thème donné, la catégorisation consiste donc à résoudre un problème de classification supervisée à deux classes ; celui-ci peut être résolu, entre autres, par des méthodes à base d'apprentissage numérique comme les réseaux de neurones, les arbres de décision, les réseaux bayesiens, les machines à vecteurs supports, ou les modèles de Markov cachés.

La distinction entre la recherche d'informations et la catégorisation de textes peut être schématisée de la manière suivante : dans le premier cas, la base de documents est fixe et l'interrogation est variable, alors que, dans le deuxième cas, la source de documents est variable et l'interrogation est fixe.

Dans la pratique, la catégorisation de textes bénéficie de deux facilités considérables par rapport à la recherche d'information : la stabilité dans le temps de la thématique filtrée et la "faible" quantité de documents à traiter dans le temps. La stabilité de la thématique laisse le temps de construire des modèles sophistiqués permettant de rechercher la façon dont l'information est codée dans un texte. Le fait de filtrer un à un les documents au fil de leur arrivée, au lieu de s'attaquer à une base importante de documents, soumet le système à une contrainte de performance plus faible, et rend possible l'utilisation de modèles plus complexes.

1.3 L'extraction d'informations

L'extraction d'information cherche à analyser de manière précise le contenu d'un document, contrairement à la recherche de texte qui étudie sa thématique générale. Il s'agit donc d'une tâche qui ne peut être accomplie qu'après une sélection préalable des documents, et qui est considérée comme plus ardue que la catégorisation de textes. Pendant plusieurs années, les modèles d'extraction d'information ont été évalués grâce à la conférence Message

Understanding Conference (MUC) [Muc7, 1999]. Pour cette compétition, les participants doivent proposer des systèmes qui remplissent automatiquement des formulaires (ou patrons). Par exemple, pour les textes traitant des changements de poste au sein des grandes entreprises, il faut compléter automatiquement les champs suivants : nom, date d'arrivée, date de départ, changement, …. .

Beaucoup de systèmes d'extraction d'informations reposent sur l'utilisation de méthodes issues du traitement naturel du langage ; par exemple, [Vichot *et al.*, 1999] proposent une application, opérationnelle au sein de la Caisse des Dépôts, qui permet de visualiser simultanément une dépêche sur des prises de participations et le graphe de participations des sociétés qui y sont mentionnées.

Des méthodes d'apprentissage ont également été utilisées pour cette tâche, le lecteur intéressé pourra trouver des informations sur ces systèmes dans [Zaragoza, 1999] ainsi que des approches utilisant des réseaux de neurones ou des modèles de Markov cachés.

1.4 Les sources d'informations

Le filtrage de documents présuppose une source de documents, c'est-à-dire un canal par lequel sont délivrés de nouveaux documents au fil du temps.

L'avènement des documents numérisés, avec les traitements de textes puis l'Internet, a vu une croissance très rapide du nombre de sources de documents. Nous en mentionnons quatre pour leur importance :

- les agences de presse,
- les *news groups*,
- les courriers électroniques,
- le web.

Ce qui distingue particulièrement les agences de presse des autres sources, c'est la qualité, l'homogénéité et la régularité éditoriale avec lesquelles elles sont produites. Les autres sources citées ne subissent pas les contraintes éditoriales traditionnelles.

Parmi les grandes agences de presse de dimensions internationales figurent Reuters[1], Associated Press[2] et l'Agence France-Presse[3](AFP).Toutes trois fournissent depuis longtemps les médias (journaux, radios, télévisions) en dépêches, délivrées en temps réel sur les téléscripteurs de leurs clients. Aujourd'hui, ces agences touchent une gamme beaucoup plus étendue de clients allant des *traders* aux services de communications des grands groupes ; le support de transmission privilégié actuellement est le protocole TCP/IP, via l'Internet et les grands navigateurs du marché (Internet Explorer et Netscape Navigator).
Dans notre application opérationnelle, présentée au chapitre 9, la source d'information est l'AFP.

Parallèlement à ces sources alimentées par des professionnels de l'information, l'essor de l'Internet a vu se développer les *news groups,* dans lesquels l'information n'est pas travaillée, vérifiée, certifiée, authentifiée selon les règles de l'art dans une salle de rédaction de journalistes appliquant des règles éditoriales. Au contraire, l'information y est produite par ceux qui veulent faire connaître leur expérience. Ces *news groups* sont le moyen, pour des utilisateurs ayant une communauté d'intérêts, de s'informer mutuellement et de débattre de questions extrêmement spécialisées au fil de l'actualité ou des marottes de ses membres. Les faiblesses de ce média sont l'inégale qualité des contributions, la résurgence cyclique de certaines thématiques, la prise de parole excessive et, en définitive, le nombre de messages que le *news group* peut engendrer chaque jour. Compte tenu, à la fois, de la gratuité, de l'indéniable intérêt, et de l'inorganisation de cette source d'informations, il n'est pas étonnant que de nombreux travaux de filtrage de documents s'y soient attaqués [Zarogoza, 1999].
Le nombre croissant de messages électroniques, et l'envoi de courriers publicitaires, peuvent nécessiter l'utilisation d'un système de filtrage. Par exemple, [Cohen, 1996] propose de classer automatiquement les courriers reçus dans des répertoires prédéfinis et [Sahami *et al.*, 1998] proposent de filtrer automatiquement les courriers indésirables (appelés souvent *spam*).

[1] http://www.reuters.com

[2] http://www.ap.org

[3] htpp://www.afp.com

Le web est évidemment une source importante de documents dont la richesse potentielle demande qu'ils soient traités. Il s'agit d'un domaine très porteur pour le filtrage ; on peut imaginer disposer d'un outil à qui l'on spécifie des sites de référence ainsi qu'une thématique de filtrage afin de détecter les nouvelles pages. Il s'agit là d'applications en dehors de notre champ d'investigation.

1.5 La conférence TREC

La conférence annuelle Text REtrieval Conference[1] est organisée chaque année sous l'égide du National Institute of Standards and Technology (NIST) sous le patronage de la DARPA ; elle est ouverte à toutes les équipes ayant préalablement participé à la compétition.

Elle offre un forum d'évaluation et de discussions pour la communauté scientifique qui se consacre au traitement automatique des textes en général, et au filtrage en particulier.

Un ensemble de tâches différentes est proposé aux différents participants qui soumettent des résultats à autant de tâches qu'ils le souhaitent. Certaines tâches font uniquement appel à des approches issues du traitement automatique du langage naturel et d'autres, comme la tâche de filtrage, nécessitent l'utilisation de méthodes à base de statistiques. Une description générale de la huitième édition de cette conférence (TREC-8) peut-être trouvée dans [Voorhees et Harman, 2000] ; cette huitième édition de la conférence a regroupé soixante six équipes différentes venant de seize pays différents.

Parmi les multiples tâches proposées dans cette compétition, la recherche *ad hoc* (recherche d'informations) était la tâche principale jusqu'à l'édition TREC-9. Cette tâche a notamment permis d'étiqueter une très grande quantité de textes pour un grand nombre de thèmes différents. Cet ensemble constitue un corpus de référence, qui peut être utilisé par la communauté scientifique pour comparer des méthodes d'apprentissage et les faire progresser.

La tâche de filtrage proposée à TREC se décompose en trois sous-tâches :

1. Le filtrage adaptatif (*adaptive filtering*) consiste à construire un premier modèle grâce à une requête formulée en langage naturel, puis à simuler un flux de documents. Le système peut tirer parti de la pertinence ou de la non-pertinence des documents sélectionnés pour s'améliorer au fil du temps.

[1] Toutes les informations et publications relatives à cette conférence sont disponibles sur : http://trec.nist.gov

2. Le filtrage par lots (*batch filtering*) consiste à utiliser une base de documents préalablement étiquetés pour construire un modèle. Pour chaque document d'un flux, le système doit prendre une décision binaire et peut utiliser, comme précédemment, la classe des documents sélectionnés pour s'améliorer.
3. Pour le routage (routing), le système dispose également d'une base de documents étiquetés pour l'apprentissage, puis est ensuite figé dans le temps. Les documents de la base de test doivent être ensuite ordonnés, du plus pertinent au moins pertinent. Le système ne doit donc pas effectuer une décision binaire, mais il doit être capable de calculer un score de pertinence.

Il est toujours possible de passer du routage au filtrage par lots en considérant que les documents dont le score est au-dessus d'un certain seuil sont pertinents. Il est nécessaire de choisir un "bon" seuil, ce qui n'est pas trivial : un système performant pour le routage peut être médiocre pour le filtrage par lots si le seuil n'est pas correctement choisi.

1.6 Notre travail

Le but de nos travaux est de développer un modèle fondé sur l'apprentissage numérique pour la catégorisation de textes ou, plus précisément, pour ce qui correspond à la tâche de routing dans le découpage de TREC.

Notre approche propose l'utilisation d'un réseau de neurones avec une architecture qui prend en considération le contexte local des mots. Malgré un nombre de paramètres élevés et un nombre d'exemples souvent limité, l'utilisation d'une méthode de régularisation permet d'effectuer correctement les apprentissages.

Nos résultats ont été validés d'une part grâce au corpus Reuters-21578[1] qui est souvent utilisé par la communauté de la catégorisation de textes, et d'autre part, par la participation aux sous-tâches de routing de TREC-8 [Stricker *et al*., 2000b] et TREC-9 [Stricker *et al.*, 2001] qui ont permis d'effectuer des comparaisons chiffrées avec d'autres approches.

Tous ces corpus sont composés d'articles d'organes de presse les dépêches de l'AFP dont le filtrage est décrit dans le chapitre 9.

[1] Ce corpus est publiquement accessible sur le site : http://www.att.research.com/~lewis/reuters21578.html

Nos travaux ont été intégrés dans l'application ExoWeb développé à la Caisse des Dépôts [Landau *et al.*, 1993] [Wolinski et Vichot, 2001] pour y ajouter des fonctionnalités opérationnelles originales. Cette application offre, sur l'intranet du groupe, un service de catégorisation de dépêches AFP en temps réel ; cette catégorisation s'effectue grâce à des modèles à bases de règles.

La première fonctionnalité développée offre un outil pour l'administrateur du système pour surveiller automatiquement le vieillissement de filtres construits sur des modèles à base de règles [Wolinski *et al.*, 2000]. L'idée de cette application est de fabriquer une "copie" d'un filtre à base de règles avec un filtre utilisant un réseau de neurones. Comme, le réseau de neurones produit une probabilité de pertinence et non une réponse binaire, il est possible d'analyser les plus grandes divergences entre les deux filtres : les documents considérés comme pertinents par la méthode à base de règles, mais obtenant une probabilité proche de zéro avec le réseau de neurones, et les documents considérés comme non pertinents avec le premier et obtenant une probabilité de pertinence proche de un avec le second.

Nous proposons ensuite les bases d'une deuxième application pour qu'un utilisateur puisse fabriquer lui-même un filtre à sa convenance avec un travail minimum. Pour réaliser cette application, il est nécessaire que l'utilisateur fournisse une base de documents pertinents. Cela peut se faire grâce à l'utilisation d'un moteur de recherche conjointement avec un réseau de neurones [Stricker *et al.*, 2000a] ou uniquement grâce au moteur de recherche.

1.7 Plan du mémoire

Le chapitre 2 est une présentation des modèles couramment utilisés en recherche d'informations, comme le modèle vectoriel ou le modèle probabiliste. Ces modèles sont présentés, car ils sont à l'origine de beaucoup de modèles construits pour la catégorisation de textes. Des approches mettant en œuvre des méthodes d'apprentissage numérique pour la catégorisation de textes sont également présentées, en insistant sur les approches "neuronales".

Le chapitre 3 présente les corpus utilisés tout au long de cette étude : le corpus Reuters-21578, et le corpus de la tâche de routing de TREC-8. Cette présentation expose les caractéristiques de chacun de ces deux corpus afin de mettre en avant leurs spécificités. Ces

corpus sont disponibles gratuitement et ont été utilisés par d'autres auteurs, ce qui facilite les comparaisons.

Le chapitre 4 introduit les différentes mesures utilisées pour évaluer les performances des systèmes. Au-delà des définitions, ce chapitre met en évidence le bruit inhérent à ces mesures et l'impossibilité qu'il existe à faire une mesure exacte. Les performances absolues n'ont pas un grand sens, et seules les performances relatives sont importantes.

Le chapitre 5 montre comment les textes sont transformés pour pouvoir être utilisés par les méthodes d'apprentissage numérique. Ce chapitre met en évidence la nécessité et la difficulté d'effectuer une sélection de descripteurs. Nous proposons une méthode entièrement automatique en deux étapes. La première étape détermine le vocabulaire spécifique des documents pertinents par rapport à l'ensemble du corpus ; la deuxième étape est une procédure d'orthogonalisation selon la méthode de Gram-Schmidt qui présente l'avantage d'être adaptée à la classification. Les comparaisons avec d'autres méthodes comme l'information mutuelle ou la méthode du chi-2 prouvent que, malgré des approches différentes, il existe peu de différences significatives entre toutes ces méthodes pour les problèmes qui nous concernent.

Le chapitre 6 est une présentation succincte des réseaux de neurones. Ce chapitre insiste sur la notion de surapprentissage pour les problèmes de classification, et montre que les méthodes de régularisation comme le *weight decay* apporte une solution à ce problème tout en ajoutant de nouveaux paramètres appelés hyperparamètres. Ces hyperparamètres peuvent être théoriquement déterminés grâce à l'approche bayésienne qui est présentée avec les approximations nécessaires à sa mise en œuvre.

Le chapitre 7 présente les premières expériences effectuées sur le corpus Reuters ainsi que la description de notre participation à TREC-8. Ce chapitre étudie l'impact des différents paramètres intervenant dans la sélection des descripteurs sur les performances (nombre de documents non pertinents, choix de ces documents, nombre de descripteurs initiaux). L'étude

du nombre optimal de neurones cachés montre qu'il est nécessaire d'améliorer la représentation des textes avant de complexifier l'architecture des réseaux de neurones en ajoutant des neurones cachés.

Le chapitre 8 présente une méthode originale pour déterminer automatiquement le contexte caractéristique d'un mot pour effectuer une désambiguïsation partielle de ce mot. L'architecture neuronale est modifiée pour prendre en considération cette nouvelle représentation. Avec celle-ci, l'utilisation d'une méthode de régularisation est indispensable ; malheureusement les résultats de l'approche bayesienne n'ont pas permis d'obtenir des résultats satisfaisants pour la détermination des hyperparamètres.
Finalement les résultats obtenus sur le corpus Reuters et sur le corpus de TREC-8, montrent une amélioration notable des résultats par rapport au chapitre 7.
Ce chapitre se termine par la description de notre participation à TREC-9.

Enfin le chapitre 9, montre comment ces résultats sont intégrés dans une application existante de filtrage de dépêches de l'AFP en temps réel. Les méthodes d'apprentissage numérique permettent de proposer de nouvelles fonctionnalités à cette application. Une première application utilise la sortie d'un filtre construit sur des méthodes à base de règles conjointement avec la sortie d'un filtre neuronal pour surveiller le premier filtre. Une deuxième application utilise un moteur de recherche couplé aux méthodes neuronales pour permettre à un utilisateur de définir son propre filtre.