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

Deuxième partie .

Catégorisation de textes multilingues

Chapitre 6

Catégorisation multilingue : les solutions proposées

Sommaire

6.1. Introduction

La recherche accorde, ces dernières années, beaucoup d'importance au traitement de données multilingues. Les utilisateurs ne se contentent plus d'accéder aux informations et de les manipuler dans leurs langues maternelles, mais ils tendent de plus en plus de franchir le pas vers les autres langues.

La recherche documentaire multilingue est l'une des solutions proposée pour répondre à ces nouveaux besoins. Son objectif est de permettre la formulation d'une requête dans une langue et d'extraire les documents correspondants en différentes langues. Les solutions proposées par les équipes de recherches ont prouvé la faisabilité de l'approche mais aucun moteur de recherche ne l'a offerte à ses clients.

Nous, de notre côté, nous allons proposer et décrire trois solutions pour catégoriser les textes multilingues. Pour ce faire nous faisons appel à deux étapes supplémentaires : l'indentification de la langue et la traduction automatique.

Dans ce chapitre nous allons aborder les facteurs qui ont contribué à l'intérêt accru de traitement de données multilingues, pour ensuite présenter les solutions proposées par d'autres pour le traitement de ces données multilingues ; enfin, nous proposerons notre propre solution qui sera testé dans la section 9.3 page 120.

6.2. Intérêt accru aux traitements multilingues

Plusieurs raisons ont été à l'origine de l'intérêt accru pour les traitements de données multilingues : la disponibilité de plus en plus large des documents mis en réseau et distribués au plan international, le nombre croissant de « non-anglophone » qui se connectent en ligne, la globalisation, la création de zones de coopération entre des pays (Union Européenne, ALENA, Forum Asie-Pacifique, etc.), le développement de l'infrastructure de communication et de l'Internet.

6.2.1. Davantage de collections numériques

La société globale de l'information a radicalement transformé la façon d'acquérir la connaissance, de la disséminer et de l'échanger, provoquant une révolution dans le monde des bibliothèques. La disponibilité des collections, mises en réseau et distribuées au niveau mondial, a créé chez les utilisateurs de nouveaux besoins de trouver, de retrouver et de comprendre une information pertinente, quelle que soient la langue et la forme de stockage [Peters and Sheridan, 2001]. Il est, par exemple, impensable pour l'utilisateur final qui désire interroger une collection multilingues de formuler des requêtes dans toutes les langues.

	Anglais	Non-anglais	Total langu. europe (sauf l'anglais)
Accès à l'Internet (M)	230,6	403,5	224,1
%âge pop. en ligne	36,5%	63,5%	35,5%
2004 (est. in M)	280	657	328
Total pop. (M)	508	5.633	1.218
Produit national brut ($B)	13.812	27.590	14.112
%âge de l'économie mondial	33,4%	66,6%	33,9%

TAB. 6.1.: Répartition de la population mondiale connectée à internet en 2003

6.2.2. Plus de personnes connectées en ligne

Le nombre de langues parlées sur cette planète s'élève à 6.700 dans 228 pays. L'anglais est la langue maternelle de 6% de la population mondiale et pourtant elle est la langue dominante pour les collections et les ressources sur le réseau mondial internet. Néanmoins cette domination est en train de reculer pour ouvrir la voie vers un réseau mondial multilingue [Nunberg, 2000].

En 1998, d'après les chiffres publiés par les organismes internationaux comme L'UNESCO (http://www.unesco.org) et le NUA (http://www.nua.com), environ 60% de la population connectée en ligne parlait (ou lisait) l'anglais et 30% communiquaient par les autres langues européennes. Plus précisément, 147 millions de personnes étaient connectées et, parmi elles, 87 millions aux États Unis et au Canada, 33,5 millions de l'Europe, 22 millions des pays d'Asie, et seulement 800.000 dans les pays d'Afrique.

En 2002, la situation est en train de changer. La table 6.1 montre les nouvelles statistiques confirmant la progression constante des personnes connectées au réseau mondial des autres pays ; nous pouvons constater que la population « non anglaise » représente 63,5 % (403 millions) et la population dont l'anglais est la langue maternelle ne représente désormais que 36,5% (230,6 millions)[1].

6.2.3. Plus de globalisation et de pays unifiés

La globalisation du monde s'accentue. Plusieurs zones de coopération, voire d'union, sont en cours de création. L'unification et l'élargissement de l'Europe, qui a pour objectif d'ouvrir le marché entre les pays et de créer un espace de coopération politique européen, est à l'origine de plusieurs projets de recherches sur le multilinguisme. Parmi ces projets nous pouvons citer le projet EMIR (*European Multilingual Information Retrieval*) qui a lancé le système commercial SPIRIT supportant les

[1] Ces statistiques sont extraites en grande partie à partir de cette page : http://www.nua.com/surveys/index.cgi?f=VS&art_id=905358509&rel=true

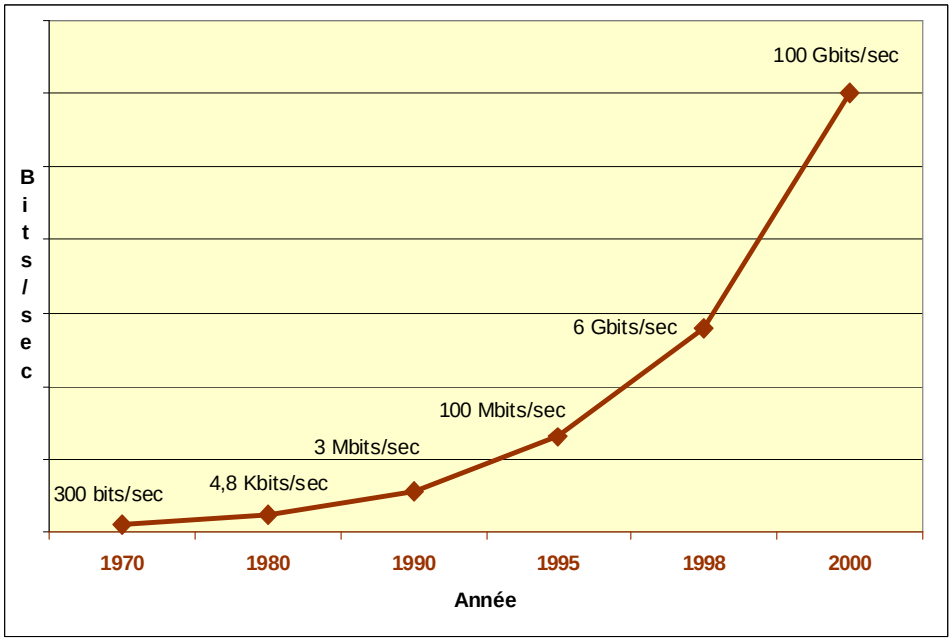


FIG. 6.1.: Évolution de la vitesse de transmission des réseaux

langues française, anglaise, allemande, hollandaise et russe. L'UNESCO a également financé des projets comme le projet MEDLIB, pour démocratiser l'accès au patrimoine culturel mondial de la Méditerranée.

6.2.4. Réseau plus rapide et plus souple

La vitesse de transmission des réseaux, comme le montre la figure 6.1, est une vraie révolution. L'accès distant à des masses considérables de données pour un coût quasi dérisoire conduit à une duplication des données par copie. En une décennie, nous sommes passés d'une ère où les données étaient rares, et où leur regroupement dans une base de données leur donnait une valeur économique intrinsèque, vers une ère d'abondance de données où la valeur de celles-ci s'est complètement diluée au profit de l'information ou de la connaissance potentielle qu'elles renferment [Zighed and Rakotomalala, 2002].

6.3. Recherche documentaire multilingues

Les facteurs présentés ci-dessus ont aidé à la disponibilité de l'information multilingue en format électronique et ont provoqué de nouveaux besoins. Ces besoins

concernent à la fois : - la manière de rechercher une information en langue étrangère et - la façon d'accéder à cette information écrite dans différents alphabets et/ou différents codages de caractères. Ici, nous nous limitons à présenter les solutions proposées par les chercheurs pour la recherche documentaire multilingue.

Comme il est dit dans la section 1.6 page 16, la recherche documentaire est le processus par lequel l'utilisateur cherche à localiser les documents dont le sujet correspond à ses besoins, exprimés par une requête.

Beaucoup d'utilisateurs ont une connaissance partielle de langues étrangères, mais leur capacité est insuffisante pour formuler une requête dans ces langues [Oard and Dorr, 1996]. Il y a bien d'autres circonstances dans lesquelles l'utilisateur, non familier avec la langue principale d'une collection, peut utiliser la recherche d'information multilingue ; par exemple :

1. la recherche dans une collection d'images indexées et annotées dans une langue non familière,
2. la recherche des organismes, ou des individus, qui s'intéressent à un certain domaine (veille technologique),

La disponibilité de moyens de traduction automatique permettant à l'utilisateur de traduire des textes sélectionnés développe l'intérêt de la recherche d'information multilingue.

La recherche documentaire multilingue a pour objectif de permettre à l'utilisateur de formuler une requête dans sa langue maternelle afin de trouver les documents dans plusieurs langues.

A ce jour, trois familles d'approches pour la recherche documentaire multilingue ont été proposées [Fluhr, 1995] : les approches basées sur la traduction automatique, les approches basées sur les thésaurus multilingues et les approches basées sur les dictionnaires.

6.3.1. Approches basées sur la traduction automatique

Les objectifs d'un système de recherche documentaire et d'un système de traduction automatique ne sont pas les mêmes. La recherche multilingue vise à trouver les documents, dans une langue cible, les plus proches au sens d'une mesure de similarité, de la requête dans la langue d'origine. La traduction automatique, de son côté, vise à produire une version lisible et fiable d'un document d'une langue d'origine vers la langue cible.

La similarité et la lisibilité sont deux objectifs différents ; pour reproduire le sens, le traducteur automatique n'utilise pas forcement le même vocabulaire que celui utilisé dans la requête.

La traduction d'une collection de documents dans la langue de requête (la langue cible) est coûteuse. Les recherches fondées sur un système de traduction automatique se sont donc concentrées sur la traduction des requêtes plutôt que celle des documents

[Peters and Sheridan, 2001]. Les requêtes sont en général un ensemble de mots avec peu ou pas de structure syntaxique. De ce fait, l'analyse grammaticale et les méthodes d'identification de polysémie ne peuvent s'appliquer sur ces requêtes car celles-ci ne sont pas sémantiquement cohérentes. De nombreux travaux ont montré que les techniques fondées simplement sur des dictionnaires peuvent être plus efficaces que les systèmes de traduction automatique [Ballesteros and Croft, 1998].

6.3.2. Thésaurus multilingues

Le vocabulaire contrôlé est fréquemment utilisé pour indexer et décrire un document. Il s'agit d'un ensemble fini de concepts que l'on doit associer aux textes d'une collection. Ainsi, un document sera représenté par un ou plusieurs concepts.

Un thésaurus multilingue est une extension de la notion de vocabulaire contrôlé monolingue. Un thésaurus multilingue peut être vu comme un ensemble de thésaurus monolingues qui englobent tous un même système de concepts. Les utilisateurs, pour rechercher une information écrite dans d'autres langues, formulent leurs requêtes dans leurs langues maternelles puis le système, automatiquement, réalise en interne la relation entre terme et descripteur [Soergel, 1997].

Notons que cette approche souffre des limites relatives au vocabulaire contrôlé : très chers à construire, coûteux à maintenir et difficile à mettre à jour ; il faut ajouter les efforts nécessaires pour former les utilisateurs qui doivent utiliser de tels systèmes [Peters and Sheridan, 2001].

6.3.3. Utilisation de dictionnaires

L'utilisation de traducteurs automatiques pour traduire les requêtes a montré ses limites car l'absence de structure syntaxique de requête empêche l'analyse grammaticale et n'aide pas, par conséquent, à lever l'ambiguïté. D'autres solutions, comme l'utilisation de dictionnaires bilingues informatisés, sont proposées. Ces outils sont de plus en plus disponibles en ligne. Certains auteurs ont utilisé les dictionnaires pour traduire les requêtes et ont obtenu de résultats acceptables mais ils sont bien inférieurs à ceux obtenus par la recherche monolingue. [Ballesteros and Croft, 1998, Hull and Grefenstette, 1996] obtiennent, à titre d'exemple, des performances de 40% à 60% inférieurs à celles obtenue par une recherche monolingue. [Peters and Sheridan, 2001] reportent trois raisons principales qui expliquent cette situation : (i) les dictionnaires généraux n'incluent pas tous les vocabulaires spécialisés ; (ii) échec de la traduction de termes constitués de plusieurs mots (exemples : sécurité sociale, Armée du Salut, chemin de fer) ; (iii) problème de l'ambiguïté.

Finalement, le problème de la recherche documentaire multilingue a-t-il trouvé sa solution ? D'après [Oard, 2002], la réponse est oui et non. La réponse est oui puisque plusieurs équipes de recherches ont montré la faisabilité de l'approche. La réponse est non puisque aucun moteur de recherche, ou service de recherche commerciale

(comme Dialog), n'offre ce service à ses clients. En effet, nombre de problèmes sont encore sans solution, tels que :

- **La généralité** : la recherche documentaire multilingue a été testée avec succès sur certains types de textes, et dans certains domaines, mais le problème est loin d'être réglé pour d'autres domaines comme les brevets, les résumés de papiers scientifiques, *etc*.
- **L'efficacité** : peut-on fournir ces services avec des coûts comparables à ceux de la recherche monolingue ? La réponse est non car ceci impose des coûts importants de temps d'accès aux donnée et de traduction de requêtes et/ou de documents.
- **L'interaction** : lorsqu'un système de recherche monolingue fournit une liste ordonnée de réponses pour une requête, cette liste fixe les frontières de recherche pour un utilisateur et cette liste devient utilisable telle quelle. Mais fournir une liste qui pointe vers des pages multilingues est beaucoup plus difficile à utiliser. Ainsi, il faut viser à proposer des moyens permettant plus d'interaction entre utilisateur et système de recherche multilingue.

6.4. Nos solutions pour catégoriser des textes multilingues

Nous proposons maintenant un cadre pour la catégorisation de textes multilingues. L'objectif est d'apprendre à inférer sur des textes rédigés dans une langue quelconque, à partir d'un modèle de prédiction construit sur un corpus de textes rédigés dans un langue donnée. Par rapport à la généralisation classique, la phase d'inférence comprend deux étapes supplémentaires :

1. la détection automatique de la langue du texte,
2. puis la traduction automatique du texte vers la langue de référence.

Nous supposons avoir $\mathfrak{L}$ corpus de textes, chaque corpus $\mathfrak{L}_l$ comportant des textes écrits dans la langue $\mathfrak{L}_l$ (nous utiliseront $\mathfrak{L}_l$ pour désigner le corpus et la langue du corpus) avec $l = 1, ..., |\mathfrak{L}|$; d_{jl} désigne un texte d_j qui fait partie du corpus $\mathfrak{L}_l$ (dont la langue est L_l) ; chaque texte d_{jl} peut être associé à une ou plusieurs classes $c_i \in \mathcal{C}$. Notre objectif est de pouvoir associer une classe c_i à chaque texte d_{xy}, dont la langue et la classe sont a priori inconnues. Il est évident que les classes des corpus doivent êtres comparables c'est pour cette raison que nous utilisons un seul ensemble de classes $\mathcal{C}$.

Nous allons présenter trois schémas.

- Le premier, nommé le schéma « trivial », représente une extension naïve du schéma de catégorisation monolingue habituel ; Il consiste en l'apprentissage de $|\mathfrak{L}|$ modèles (un modèle pour chaque langue).
- Le deuxième est un schéma permettant l'apprentissage d'un seul modèle.

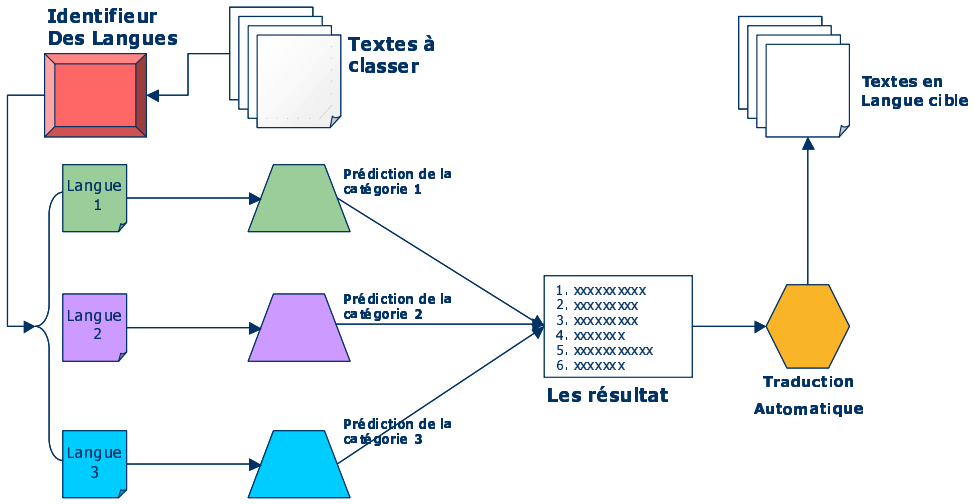


FIG. 6.2.: Phase de classement selon le schéma trivial (extension de schéma monolingue) : apprendre un modèle pour chaque langue et ainsi il faut apprendre $|\mathfrak{L}|$ modèles

– Le troisième consiste en la traduction des textes de plusieurs ensembles d'apprentissage vers une langue cible pour ensuite apprendre un seul modèle « mixte ».

Nous allons maintenant détailler ces trois schémas.

6.4.1. Premier schéma : le schéma trivial

Nous appelons ce schéma « trivial » car il s'agit d'une extension directe de la catégorisation « monolangue ». La figure 6.2 décrit le processus de catégorisation. Pour chacun des $\mathfrak{L}$ corpus (langues) nous apprenons un modèle, soit $|\mathfrak{L}|$ modèles à construire. Pour chaque texte à classer 1) nous identifions sa langue puis 2) nous appliquons le modèle de prédiction appris pour cette langue. Si le texte est identifié comme « intéressant » nous le passons à un traducteur automatique afin de le traduire vers la langue cible (par exemple le français). L'avantage avec ce schéma est l'intervention de la traduction à la fin de processus : le traducteur n'intervient pas dans l'apprentissage et, par conséquent, aucune distorsion d'information ni perte n'est commise à cette étape.

Mais ce schéma présente de sérieuses limites :

– il exige de faire $|\mathfrak{L}|$ apprentissages, un par langue,
– et ceci suppose d'avoir des quantités suffisantes de textes étiquetés dans chaque langue et dans chaque classe. Ceci est difficile surtout pour les langues peu présentes sur le Web.

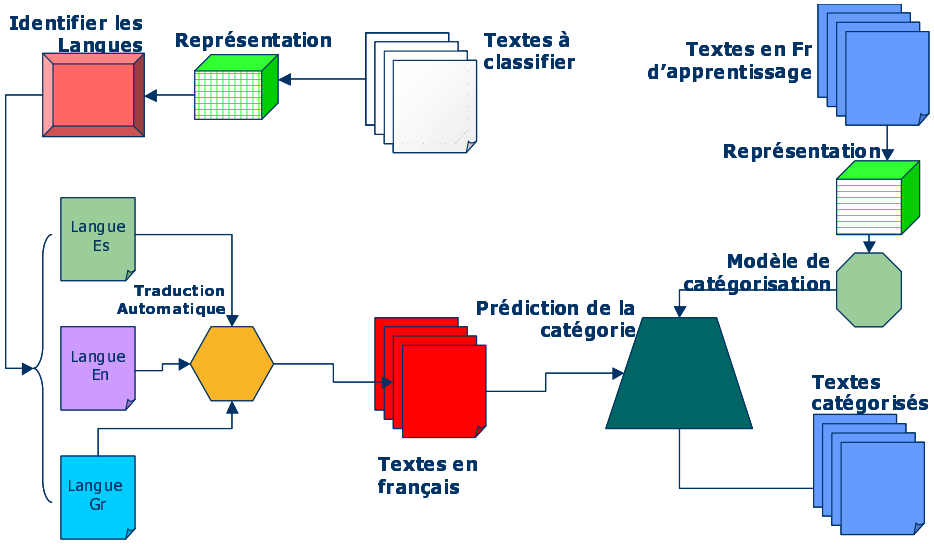


FIG. 6.3.: Deuxième schéma : choix d'une langue d'apprentissage et utilisation d'un seul modèle, dans cette langue

Mais ce schéma sera notre référence pour valider les autres solutions proposées maintenant.

6.4.2. Deuxième schéma : choisir une seule langue d'apprentissage

Le deuxième schéma que nous proposons pallie en partie les deux critiques précédentes, car nous utilisons ici un seul modèle de prédiction. Pour chaque texte à classer nous identifions d'abord sa langue ; si cette langue est supportée par le traducteur automatique, nous traduisons le texte dans la langue d'apprentissage (la langue du modèle de prédiction) ; enfin, nous appliquons le modèle de la langue d'apprentissage sur ce texte pour le classer. Dans ce schéma le traducteur joue un rôle primordial dans la phase de classement (voir la figure 6.3).

6.4.3. Troisième schéma : mélanger les ensembles d'apprentissage

Dans le deuxième schéma, la traduction intervient uniquement dans le processus de classement. Il nous semble intéressant de tester une solution où la traduction intervient dans les deux processus : apprentissage et classement.

Disposant d'un ensemble de textes (étiquetés) écrits dans différentes langues, on les traduits tous dans une langue d'apprentissage $\mathfrak{L}_{app}$ unique, puis on les réunit en un corpus unique, sur lequel on procédera à l'apprentissage des étiquettes (classes).

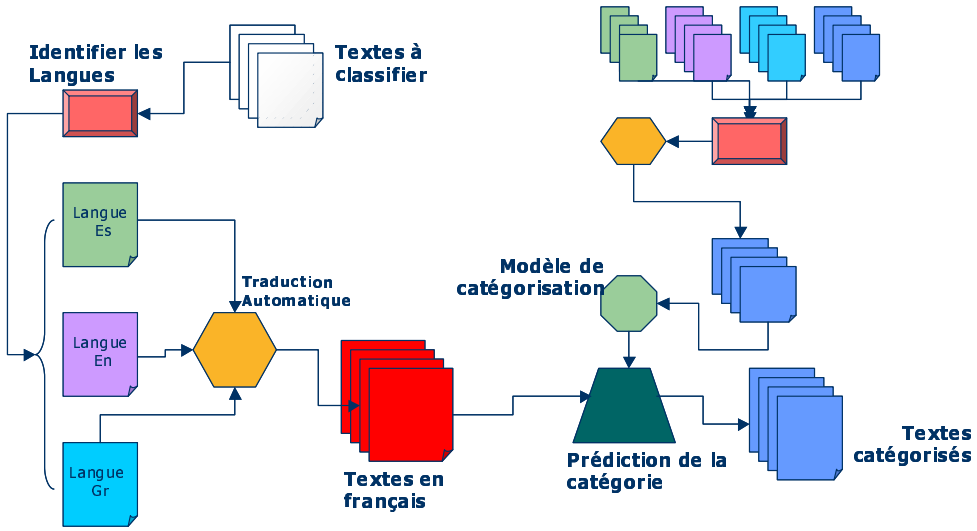


FIG. 6.4.: Troisième schéma : réunion des ensembles d'apprentissage après avoir traduit tous les corpus vers une seule langue, et apprentissagr d'un seul modèle, adapté aux textes traduits

Ainsi le modèle sera appris à partir des textes produits par les traducteurs automatiques et il sera plus à même de classer un nouveau texte, traduit par le même outil.

Ensuite, pour chaque texte à classer, 1) nous identifions sa langue ; 2) si la langue est prise en charge par le traducteur, nous le traduisons vers la langue cible $\mathfrak{L}_{app}$, et 3) nous appliquons le modèle (voir la figure 6.4).

Nous somme actuellement en train de tester ce schéma. Nous espérons qu'il donnera de bons résultats car les traducteurs ont tendance à « agréger » les vocabulaires et ainsi à fournir un corpus homogène.

6.5. Conclusion

De nombreux facteurs concourent à la disponibilité de textes en plusieurs langues. Les utilisateurs ne se contentent plus d'utiliser leur langue maternelle, mais ils s'intéressent de plus en plus aux textes écrits dans d'autres langues. De nouveaux outils sont apparus pour répondre à ces besoins émergeants. La recherche documentaire multilingue, à titre d'exemple, doit de permettre à l'utilisateur de formuler une requête dans sa langue d'origine et de lui fournir des textes écrits dans d'autres langues.

Nous proposons des méthodes de catégorisation de textes, queles que soient leurs langues ; ces méthodes peuvent sembler naïves, mais elles donnent d'assez bons résultats comme nous le verrons au chapitre 9 page 117. Les trois schémas proposés font appel à deux étapes supplémentaires pour le classement et/ou l'apprentissage de

textes : l'identification de la langue et la traduction automatique. Ces deux étapes sont importantes et les bonnes performances des modèles d'apprentissage sont directement affectées par ces deux étapes.

Dans les deux chapitres suivants page 97 et page 111, nous allons présenter les méthodes d'identification automatique de la langue d'un texte puis de traduction automatique. Le chapitre 9 page 117 présentera nos expérimentations sur les deux premiers schémas discutés dans les sections 6.4.1 et 6.4.2 ci-dessus.

Chapitre 7

Identification de la langue

Sommaire

7.1. Introduction

L'identification de la langue consiste à attribuer une unité textuelle, supposée monolingue, à une langue. Cette identification est devenue importante puisque les données textuelles, en différentes langues, trouvent de plus en plus leur chemin sur le réseau mondial. En outre la bioinformatique fournit de vastes problèmes traitables par des outils similaires.

Dans ce chapitre nous allons identifier les facteurs à considérer, puis nous allons présenter les techniques utilisées et proposer deux nouvelles techniques, la première basée sur la mesure de distance de χ^2 et la deuxième sur les RBFs.

L'identification automatique de la langue est possible car (i) les langues naturelles écrites sont extrêmement non-aléatoires ; (ii) elles présentent chacune des régularités dans l'utilisation des caractères ou séquences de caractères ; (iii) l'alphabet de chaque langue est soit unique, soit très caractéristique de cette langue. Les informations sur la stabilité et la constance des fréquences de lettres et séquences de lettres ne sont pas nouvelles ; elles sont connues depuis des centaines d'années ; Kahn, dans son survey sur la cryptographie, mentionne que ces régularités ont été reportées dans une encyclopédie arabe écrite par Qalqashandi en 1412 qui, à son tour, attribue la majorité de ces informations à Ibn ad-Duraihim qui vivait entre 1312 et 1361 [Beesley, 1988].

Il est prouvé statistiquement que, pour chaque langue, les nombres d'occurrences des séquence de deux, trois, quatre ou cinq lettres sont stables et différents des autres langues. Par exemple, en anglais, dans un texte quelconque, la fréquence de la lettre « E » est d'environ 13%, la fréquence de la lettre « U » d'environ 3% et la fréquence de la lettre « Z » d'environ 0.1%. Pour les séquences de deux caractères ou bi-grammes, on trouve par exemple que la probabilité d'avoir la chaîne « TH » en anglais est relativement grande ; en espagnol et portugais, cette probabilité s'approche de zéro. Dans le même ordre d'idées, la probabilité d'avoir « SZ » en hongrois et polonais est grande ; la chaîne « TION » caractérise le français et l'anglais.

Partant de ces probabilités d'apparition des lettres et séquences de lettres, on peut concevoir un algorithme capable d'identifier la langue d'un texte.

Plusieur approches peuvent être utilisées lors la conception de méthodes d'identification. A titre d'exemple, [Sibun and Reynar, 1996] considèrent les distinctions suivantes :

1. Le type des éléments significatifs : un caractère, un n-gramme, ou d'autres ? Utilise-t-on des connaissances linguistiques ou non (possibilité d'adaptation à d'autres problèmes, par exemple ceux issus de la bioinformatique) ?
2. La forme de l'analyse : quel algorithme faut-il utiliser pour identifier la langue ? Certains algorithmes demandent une intervention humaine alors que d'autres sont totalement automatisés. Certains sont basés sur l'existence, ou non, de certaines séquences (mots courts), tandis que d'autres s'intéressent aux fréquences des séquences.

3. La forme de codage : quelle est la meilleure représentation du point de vue des critères de rapidité, robustesse et précision ? Doit-on utiliser un codage compréhensif ou un codage simplifié (en supprimant les caractères accentués) ?
4. Le choix de l'univers des langues à identifier.
5. La taille du texte : il est souvent nécessaire d'identifier la langue à partir d'un court extrait (une phrase, voire quelques mots). Dans les systèmes de recherche documentaire, il est important de détecter les passages multilingues afin de construire un index pour chacune des langues. Si l'identification sur de courts passages est satisfaisante, alors, pour les documents longs on a intérêt à analyser quelques échantillons de texte au lieu d'analyser le texte tout entier.
6. La forme des entrées : textes en ligne, images en ligne ou les deux.
7. Le choix de la méthode statistique : on peut les comparer selon divers critères. La performance ne doit pas se dégrader si on ajoute plus de données pour l'apprentissage, ou si le texte à identifier devient plus long (par exemple, l'identification d'une langue ne doit pas être infectée par la présence ou l'absence d'autres langues. On doit pouvoir tenir compte d'informations de fréquences ; [Cavnar and Trenkle, 1994] utilisent une méthode qui calcule la distance en fonction de la position des n-grammes selon leur fréquence, si deux langues possèdent les mêmes rangs alors la méthode de Cavnar n'arrivera pas à les distinguer bien que les fréquences des n-grammes ne soient pas les mêmes.

Plusieurs éléments significatifs peuvent être considérés pour l'identification de la langue : la présence de certains caractères [Mustonen, 1965] [Souter et al., 1994], la présence de certains mots [Souter et al., 1994] [Giguet, 1998] ou la fréquence de *n*-grammes [Beesley, 1988] [Cavnar and Trenkle, 1994] [Dunning, 1994] [Grefenstette, 1995].

Nous distinguerons deux familles d'approches : les approches linguistiques (en section 7.2) et les approches statistiques et probabilistes (en section 7.3).

7.2. Approches linguistiques

Ces approches d'identification nécessitent une connaissance linguistique préalable et ces connaissances linguistiques sont intégrées dans le programme informatique.

7.2.1. Présence de certains chaînes de caractères spécifiques

On repère des chaînes de caractères qui sont uniques pour chaque langue. Par exemple, le tableau 7.1 montre une liste proposée par plusieurs auteurs (cité dans [Dunning, 1994]).

Cette approche est valide intuitivement, mais pas assez pour établir des règles de décisions sûres [Souter et al., 1994] et [Dunning, 1994] : on ne peut pas attribuer la

Langue	Chaîne
Néerlandais	« vnd »
Anglais	« ery_ »
Français	« eux_ »
Gaélique	« mh »
Allemand	« _der_ »
Italien	« cchi_ »
Portugais	« _seu_ »
Serbo-croate	« lj »
Espagnol	« _ir_ »

TAB. 7.1.: Quelques chaînes de caractères spécifiques à une langue [Dunning, 1994]

langue italienne à tout texte qui parle de « Zucchini » ou de « Pinocchio », ni associer la langue française à un texte anglais qui cite le mot français « milieux ». De plus, ces chaînes de caractères uniques sont relativement rares et par conséquent ils peuvent être absents dans des textes courts (taille 50 ou 100 octets par exemple). De toute façon, comme nous le verrons plus loin (voir section 7.3.2 page 103), les approches basées sur les *n*-grammes capturent aussi ce type de connaissances.

7.2.2. Présence de certains mots

Une autre approche consiste à choisir les mots grammaticaux comme discriminants pour l'identification de la langue [Grefenstette, 1995]. L'utilisation de ces mot grammaticaux (prépositions, conjonctions, déterminants, pronoms, adverbes en liste close, auxiliaires) se justifie par le fait qu'ils permettent un diagnostic sûr : 1) ils représentent en moyenne 50% des mots d'une phrase dans la plupart des langues et que leur présence est indispensable [Giguet, 1998], 2) ces mots grammaticaux ont la propriété d'être courts et en nombre limité ce qui permet la construction de listes exhaustives.

Cette approche est assez efficace : elle obtient des scores supérieurs à ceux obtenus par l'approche précédente (la présence de chaînes de caractères uniques) et elle est rapide par rapport à d'autres méthodes basées sur les *n*-grammes que nous aborderons plus loin ; elle demande moins de calculs. Mais les performances sont mauvaises sur les énoncés très courts (comme les titres des sections) qui ne contiennent pas forcément ces mots courts [Grefenstette, 1995]. Et il faut évidement que les langues à reconnaître soient segmentées en mots (nous abordons ce point en section 7.3.1).

L'identification de la langue n'est pas un objectif en soi ; elle fait souvent partie d'un processus plus complexe comme la recherche documentaire, ou l'indexation automatique des documents multilingues. Les mots grammaticaux dans un tel pro-

cessus ne rapportent pas beaucoup d'informations ; au contraire, les chaînes significatives sont généralement les plus rares et les plus longues. Les mots courts, employés pour fournir une structure syntaxique de la langue, sont utilisés indépendemment du contenu et les systèmes de recherche documentaire et d'indexation automatique éliminent ces mots. La recherche de mots courts pour identifier la langue implique donc une étape de traitement supplémentaire. Un système basé sur les n-grammes n'aura pas ce genre de problème (voir section 2.2.4 page 25).

7.2.3. Approche lexicale

Si on définit la langue d'un énoncé par la langue des mots qui le composent, le simple recours à des lexiques de mots est suffisant pour identifier la langue : il suffit de reconnaître la langue du segment comme étant celle dont le lexique contient tous les mots [Giguet, 1998].

Mais cette approche ignore l'incomplétude des lexiques : de nombreux termes techniques, particuliers à certaines domaines, ne figurent jamais dans des lexiques généraux. Cette approche suppose par ailleurs l'absence de fautes dans les énoncés, fautes d'orthographe, fautes de frappe, qui sont très courantes et qui perturbent la reconnaissance des mots.

7.2.4. Approche plus linguistique

Cette approche s'intéresse aux propriétés positives intrinsèques des langues : par ce que l'on trouve obligatoirement ou très souvent dans une langue, et non par opposition aux autres langues. Elle choisit les mots grammaticaux, l'alphabet, les affixes fréquents, comme outils discriminants pour l'identification de la langue en les présentant comme porteurs de la langue de l'énoncé [Giguet, 1998].

Dans la section 7.2.2, nous avons présenté les avantages de l'utilisation des mots grammaticaux. L'utilisation de l'alphabet se justifie par le fait que si le mot ne peut s'écrire dans l'alphabet d'une langue alors il n'en fait pas partie. La vérification d'appartenance de toutes les lettres d'un mot à un alphabet n'est pas suffisante pour identifier une langue, mais cela peut renforcer d'autres techniques, en écartant certaines langues.

L'utilisation des affixes (préfixes et suffixes) peut contribuer à discriminer une langue ; ceci nous rappelle l'approche étudiée en section 7.2.1. Comme la technique précédente, la technique des affixes ne suffit pas pour identifier la langue, mais elle peut être utilisée en combinaison avec d'autres techniques.

Ces techniques (mots grammaticaux, l'alphabet, et les affixes fréquents) se distinguent de l'utilisation des mots les plus fréquents (technique illustrée plus loin) qui s'acquièrent sur des corpus d'apprentissage (lourds à construire) et qui sont tributaire du choix du nombre de mots à considérer comme faisant partie des plus fréquents [Péry-Woodley, 1995]. [Déjean, 1998] propose une méthode multilingue

d'extraction automatique de ces mots sur des corpus non annotés, pour la constitution de ces ressources linguistiques (les mots grammaticaux, les suffixes).

Toutes ces approches demandent une acquisition non triviale des connaissances linguistiques sur toutes les langues, tâche coûteuse et impossible à réaliser pour toutes les langues. Ces approches ne sont justifiées que si l'identification de la langue constitue une étape d'un processus d'analyse linguistique plus complexe, qui s'intéresse à la compréhension et la modélisation des phénomènes linguistiques et vise à maîtriser leur application.

Finalement, il nous semble que cette approche (l'approche plus linguistique) n'est qu'une version plus développée des autres approches présentées en sections 7.2.1 et 7.2.2 car les mots grammaticaux, l'alphabet et les affixes qui discriminent une langue peuvent être vues comme des chaînes uniques appartenant à cette langue.

7.3. Approches statistiques et probabilistes

Ces approches utilisent des ressources construites automatiquement à partir d'un corpus textuel représentatif de la langue. L'objectif est de capturer au moyen de modèles statistiques ou probabilistes certaines régularités formelles des langues et d'estimer leurs probabilités d'apparitions. Ces régularités empiriques jouent le rôle de connaissances linguistiques. L'identification consiste à calculer la probabilité pour un énoncé d'appartenir aux différentes langues, en fonction des régularités observées. Deux principaux types de régularités sont couramment exploités : les mots les plus fréquents, et les séquences de *n*-grammes les plus fréquentes.

7.3.1. Mots les plus fréquents

Cette approche est plus générale que celle étudiée en section 7.2.2. L'idée de base est proche : détecter empiriquement la présence de certains mots très présents dans une langue. Mais ici la démarche est purement statistique et la construction des listes ne demande aucune connaissance linguistique préalable ; la détection de ces mots se fait en calculant les fréquences d'apparition dans un corpus de textes et on retient les mots dépassant une fréquence donnée. Cette approche semble intuitivement valide : il est de bon sens d'attribuer la langue anglaise à un passage de texte qui contient beaucoup de mots comme *are*, *that*, *the*, *and*, *of*, ... et d'attribuer la langue espagnole à un passage de texte qui comporte des mots communs comme *los*, *del*, *por*, *para*, ..., etc. La table 7.2 montre les mots les plus fréquents, pour dix langues européennes, calculées à partir du corpus ECI [Grefenstette, 1995].

Ensuite, le principe d'identification de la langue est simple : à partir de la liste des mots retenus pour chaque langue, la fréquence d'apparition de chaque mot *m* est transformée en fréquence relative. Pour identifier la langue d'un texte, on découpe ce texte en mots. Pour ceux qui apparaissent dans la liste on leur attribue les probabilités correspondantes et pour ceux qui n'apparaissent pas on leur attribue des

Dan	Dut	Eng	Fre	Ger	Ita	Nor	Por	Spa	Swe
i	de	the	de	des	di	.	de	de	och
af	van	and	la	die	e	og	a	la	i
og	het	to	le	und	il	det	que	que	att
at	een	of		den	che	,	o	el	som
ğ	en	a	et	in	la	han	e	en	en
til	in	in	des	von	a	i	do	y	r
for	dat	was	les	.	in	er	da	a	p
en	is	his	du	zu	per	"	no	los	det
om	te	that	"	dem	del	p	um	del	av
der	op	I	en	,	un	til	em	se	fr
er	voor	he	un	fr		at	para	por	med
U	met	as	que	mit	non	som	com	las	den
ikke	die	had	a	das	i	var	se	con	till
eller	De	with	qui	des	si	jeg		un	har
som	zijn	it	dans	ist	le	med	os	para	de

TAB. 7.2.: Mots les plus fréquents d'après le « ECI Multilingual Corpus »

probabilités minimales puis on calcule le produit des probabilités. La langue de ce texte est celle pour laquelle ce produit est maximal. Cette technique est utilisée par [Grefenstette, 1995] et, bien avant, par [Beesley, 1988] (voir page 104).

Cette approche est simple, intuitive et relativement efficace, mais elle présente quelques inconvénients.

Premièrement, ses performances se dégradent lorsque les textes à identifier deviennent courts du fait qu'ils ne contiennent pas forcément des mots les plus fréquents de leur langue. [Cowie et al., 1998] montre que les performances des approches basées sur les mots se dégradent considérablement lorsque la taille du texte devient petit : les taux d'erreurs sont autour de 20% pour des textes de 50 octets et de 40% pour des textes de 20 octets, lors de tests sur 34 langues. Deuxièmement, cette approche repose sur l'hypothèse du découpage de texte en mots, or, pour certaines langues, comme le chinois, il est difficile de faire la segmentation de texte en mots ; dans le cas de séquences génétiques ADN, on ne peut pas non plus utiliser la notion de mots.

7.3.2. Méthodes basées sur les *n*-grammes

Nous avons présenté la notion de n-grammes et les avantages de l'utilisation de n-grammes dans la section 2.2.4 page 24. Rappelons que le « profil *n*-grammes » d'un document est la liste des contiguïtés de *n* caractères les plus fréquents, classées par ordre décroissant de leurs fréquences d'apparition dans le document, ainsi que ces fréquences ; un document est ainsi caractérisé par son profil *n*-grammes

[Cavnar and Trenkle, 1994].

Les systèmes d'identification de langues basés sur les *n*-grammes suivent en général le même schéma :

- Dans la phase d'acquisition automatique des connaissances linguistiques, on choisit un corpus représentatif pour chaque langue puis on génère un profil caractéristique qui fera office de référence, c.-à-d. calculer les nombres d'occurrences des différents *n*-grammes (souvent *n* est compris entre 1 et 5) et les transformer en fréquences.
- Dans la phase de diagnostic, pour chaque texte à identifier, on construit son profil *n*-grammes et on cherche le profil de référence le plus similaire. Plusieurs méthodes sont proposées pour mesurer cette similarité.

Méthodes de plus proche voisin

Plusieurs algorithmes de catégorisation de texte sont basés sur une distance ou, plus généralement, sur une similarité ou une dis-similarité. L'idée générale est de chercher le texte, parmi l'ensemble d'apprentissage, qui soit le plus proche du texte à classer pour ensuite attribuer la classe de ce texte au texte dont la classe est inconnue. La difficulté est de définir une distance. En pratique, on utilise des pseudo-distances ; voir la section 5.4 page 62, pour une description de cette technique.

Distance de Beesley La méthode de [Beesley, 1988] est basée sur des modèles mathématiques linguistiques qui, à l'origine, étaient utilisés pour décoder les textes cryptés. L'identification comporte deux phases :

- La phase d'acquisition des connaissances : établir, pour chaque langue, un profil bi-grammes (dans l'ordre sans pourtant utiliser une lettre plus d'une fois - il ne s'agit pas du profil au sens usuel) qui fera office de référence, puis calculer les probabilités d'apparitions de chaque bi-gramme dans chaque langue.
- La phase de diagnostic : rechercher, pour le texte à identifier, le profil de référence le plus proche en utilisant une mesure de similarité. Cette phase est composée de trois étapes :
  - Segmenter le texte à identifier en mots.
  - Pour chaque mot, calculer tous les bi-grammes, puis, *pour chaque langue candidate*, attribuer à chaque bi-gramme sa probabilité dans le profil de référence de cette langue et faire le produit de toutes ces probabilités.
  - La langue du texte est celle pour laquelle le produit est maximum.

Exemple : Pour identifier la phrase « LES ORDINATEURS SONT APPELÉS A JOUER UN RÔLE », [Beesley, 1988] segmente cette phrase en mots puis calcule les bi-grammes de chaque mot (en blocs consécutifs, chaque lettre est utilisée une fois seulement) ; si un bi-gramme apparaît dans le profil de référence d'une langue alors sa probabilité sera celle observée dans le profil référence. Le tableau 7.3 montre les scores obtenus ainsi que la décision prise pour le mot « ORDINATEUR ».

Langue	_O	RD	IN	AT	EU	RS	Décision
Anglais	1.8905	0.1492	1.9568	1.0945	0	0.1492	éliminé
Espagnol	0.1860	0.1691	0.7102	0.2705	0	0.0338	éliminé
Français	0.4437	0.0657	0.9202	0.8052	0.4601	0.2136	retenu

TAB. 7.3.: Probabilités (multipliées par 100 pour faciliter la lecture) et la décision finale pour le mot « ordinateur »

Par conséquent, il suffit qu'un seul bi-gramme d'un mot soit absent d'un profil de langue pour éliminer l'appartenance du mot à cette langue. Le tableau 7.4 montre les décisions finales.

Cette méthode suppose la segmentation du texte en mots ce qui empêche, par conséquent, son utilisation à des langues dont les frontières entre les mots ne sont pas claires (c.f., page 103). Elle se limite également à prendre en compte uniquement les bi-grammes alors qu'il est important de travailler aussi avec des tri-grammes ou quadri-grammes pour mieux conserver la spécificité de chaque langue : par exemple, la séquence « tion » est typique du français et de l'anglais, si on la découpe en « ti » et « on » alors le système aura du mal à pénaliser les autres langues comme l'espagnol et le portugais qui utilisent « ti » et « on » mais pas « tion ». Autre remarque : la génération des profils ne correspond pas à ce qu'on entend par profil (voir la page 104). Finalement, on remarque que cette méthode est très sensible aux fautes d'orthographes et aux déformations causées pendant les OCR, il suffit qu'un mot soit mal saisi pour pénaliser la langue concernée.

Langue	Retenu	$\simeq$ Retenu	$\simeq$ Éliminé	Éliminé
Anglais	1	6	0	1
Espagnol	0	1	4	3
Français	7	1	0	0

TAB. 7.4.: Décisions finales concernant les mots du texte

Distance de Cavnar et Trenkle (CT) Ces deux auteurs [Cavnar and Trenkle, 1994] proposent une méthode d'identification de la langue en deux phases :

Phase d'acquisition : Établir pour chaque langue un profil tri-grammes caractéristique acquis automatiquement qui fera office de référence.

Phase de diagnostic : Rechercher, pour chaque texte à classer, le profil de référence le plus proche en utilisant une mesure de similarité. Cette étape se fait en trois étapes :

– Le profil tri-gramme du nouveau texte est calculé, comme s'il s'agissait d'un corpus de référence ;
– Pour chaque langue, une distance est calculée entre son profil caractéristique et celui du nouveau texte. Cette distance repose sur la définition d'une mesure de similarité : elle correspond à la somme des écarts de rangs entre chaque tri-gramme du nouveau profil et ce même tri-gramme dans le profil de référence, s'il est présent (un écart maximal est attribué si le tri-gramme est absent du profil de référence) ;
– La langue diagnostiquée est celle pour laquelle la distance est la plus petite.

Formellement, la distance entre deux profils P_1 et P_2 est définie comme

$$CT(P_1, P_2) = \sum_{w \in P_1,\, R_{P_1}(w) < NMAX} min(|R_{P_2}(w) - R_{p_1}(w)|, DMAX)$$

où $|x|$ est la valeur absolue de x ; et où $R_P(w)$ (avec w un n-gramme et P un profil de langue) est le rang de w dans le profil P, si w fait parti de P, et $DMAX$ dans le cas contraire (exemples usuels : $NMAX = 500$ et $DMAX = 1000$).

Distance de Kullbach-Leibler(KL) [Sibun and Reynar, 1996] proposent une méthode qui utilise l'entropie relative de Kullbach et Leibler comme mesure de distance. Une entropie relative entre deux distributions de probabilité reflète la quantité d'informations supplémentaires nécessaire pour coder la deuxième distribution en utilisant un code optimal généré pour la première. Techniquement, cette mesure n'est pas une distance car elle n'est pas symétrique.

Formellement,

$$KL(T_1, T_2) = \sum_{N_g} f_2(N_g).\log(\frac{f_2(N_g)}{f_1(N_g)})$$

Ici, la somme est prise sur tous les n-grammes, avec T_1 et T_2 des textes, $f_i(N_g)$ est la fréquence du n-gramme N_g dans le texte T_i. Si le n-gramme N_g est absent dans T_i, une demi-fréquence est alors ajouté pour éviter que le score tombe vers moins l'infini.

Distance du χ^2 Cette distance est présentée dans [Benzecri, 1973] et rappelée dans notre domaine par [Rajman and Lebart, 1998]. Ici p_{ig} est la fréquence du terme (n-gramme) *g* dans le texte *i*, et $p_{.g}$ la fréquence du terme *g* dans l'ensemble de tout le corpus.

$$\chi^2(T_1, T_2) = \sum_{N_g} \frac{1}{p_{.g}} \left(\frac{p_{1g}}{p_{1.}} - \frac{p_{2g}}{p_{2.}} \right)^2 \tag{7.1}$$

Cette distance possède la propriété d'équivalence distributionnelle soulignée par [Benzecri, 1973] : si deux textes T_1 et T_2 ont le même profil n-gramme, on peut les fondre en un seul texte sans que la distance du χ^2 soit modifiée, ni entre les textes, ni entres les descripteurs g.

7.4. Expériences pour la reconnaissance de langue

Dans les expérimentations que nous allons présenter dans cette section, nous nous concentrons sur les nouvelles utilisations de deux algorithmes : les *RBF*s, pour leurs bonnes performances signalées en section 5.8.3 page 77, et les 1-*plus proches voisins*, pour leur simplicité d'utilisation et leur fréquente utilisation dans le domaine du texte. Tout le travail a été réalisée avec des implémentations Java utilisant le package Jama (voir `http://math.nist.gov/javanumerics`).

La tâche consiste à reconnaître la langue d'un texte. Nous travaillons sur 5 langues : le français, l'arabe, l'anglais, l'espagnol et l'allemand[1]. La tâche étant facile, nous la compliquons en utilisons des chaînes très courtes pour tester les algorithmes.

Avec des ensembles-tests composés d'échantillon de longueur 100, 50 ou 20 bytes ; la table 7.5 donne les taux de succès (TS) selon la longueur des échantillons à classer. Le taux de succès est évalué par test sur un ensemble disjoint de l'ensemble d'apprentissage (ou *training set*).

Notons que le taux de succès coïncide avec les micro- et macro-moyen pour le rappel et précision puisque le problème consiste à classer de textes appartenant chacun à une et une seule langue (voir la section 9.3.5 page 124).

Algorithme	TS (100 bytes)	TS (50 bytes)	TS (20 bytes)
1-NN (KL)	100 %	99.4 %	92.8 %
1-NN (χ^2)	98.8 %	96.6 %	87.92 %
RBF ($\sigma^2 = 10$)	37.6 % (100 %)		
RBF ($\sigma^2 = 100$)	98.8 %	93 %	71.04 %

TAB. 7.5.: Résultats obtenus en reconnaissance des langues

Les algorithmes avec des caractéristiques entre parenthèse, exemple : RBF (2-regroupés prof.), correspondent à des résultats obtenus en coupant l'ensemble d'apprentissage en 50 sous-ensembles au lieu de déterminer les profils sur l'ensemble du training set. Ceci améliore parfois l'apprentissage RBF mais n'a apporté aucune amélioration dans certains cas d'échantillons très courts. Nous travaillons maintenant sur 250 échantillons de 100 bytes comme ensemble d'apprentissage, pour étudier

[1] L'utilisation de nos algorithmes pour l'identification d'autres langues est très facile à mettre en œuvre. En effet, il nous suffit de quelques textes pour chaque langue afin de lui calculer et associer un profil représentatif. Cette opération est quasi immédiat (moins d'une seconde).

l'influence de ce "regroupement" pour les RBFs ou les 1-plus proches voisins. Les résultats sont donnés en table 7.6.

Algorithme	Hyperparamètres	TS (100)	TS (50)	TS (20)
RBF	$\sigma^2 = 10$	99.2 %	84.8 %	31.52 %
	$\sigma^2 = 100$	98 %	93.2 %	71 %
RBF (2-regroupés prof.)	$\sigma^2 = 100$	97.2 %	88 %	68.56 %
RBF (5-regroupés prof.)	$\sigma^2 = 1000$	98.8 %	94 %	80.88 %
RBF (10-regroupés prof.)	$\sigma^2 = 1000$	99.2 %	95.2 %	76.72 %
RBF (25-regroupés prof.)	$\sigma^2 = 1000$	98.8 %	94.8 %	82.4 %
RBF (regroupés prof.)	$\sigma^2 = 100$	88.4 %	80.6 %	
	$\sigma^2 = 100000$	87.6%	77.4 %	61.36 %
1-PPV	χ^2	99.2 %	96.6 %	88.4 %
1-PPV	KL			47.2 %
1-PPV (2-regroupés prof.)	χ^2	99.6 %	96.8 %	88.8 %
1-PPV (5-regroupés prof.)	χ^2	100 %	97.6 %	90 %
1-PPV (10-regroupés prof.)	χ^2	99.2 %	97.2 %	88.56 %
1-PPV (10-regroupés prof.)	KL			89.84 %
1-PPV (25-regroupés prof.)	χ^2	100 %	96.8 %	87.2 %
1-PPV (regroupés prof.)	χ^2	100 %	93 %	84.56 %
1-PPV (regroupés prof.)	KL	99.7 %	97.4 %	89.4 %

TAB. 7.6.: Résultats obtenus en reconnaissance des langues, en testant différents degrés de regroupement. "m-regroupés profils" signifie que les échantillons d'apprentissage ont été regroupés m par m ; "regroupés", que tous les textes d'une même classe sont regroupés. Une petite valeur de m préserve la variabilité de l'ensemble d'apprentissage, un m plus grand conduit à des profils mieux définis. La dissimilarité de Kullbach-Leibler semble meilleure que celle du χ^2 pour des petits échantillons en test, mais supporte mal les petits échantillons en apprentissage (comme on s'y attend en examinant le comportement de cette dissimilarité pour de petites fréquences)

L'hyperparamètre σ^2 pour l'apprentissage RBF était facilement choisi dans notre jeu d'essai (voir section 5.6 page 79), car le taux de succès était stable sur une grande plage de valeurs de σ ; mais dans le cas de très courtes chaînes, l'efficacité variait beaucoup selon σ et en fonction du "regroupement". Ceci amène à deux hyperparamètres difficiles à calculer.

7.5. Conclusion

L'objectif de ce chapitre était d'examiner les approches existantes pour l'identification de la langue. Nous avons présenté les deux familles d'approches : celles basée sur des connaissance linguistique et celle basée sur l'approche purement statistique. Nous avons préféré l'approche statistique basée sur les n-grammes, car elle est indépendante de la langue, tolérante aux fautes d'orthographe et déformations, et capture des connaissances linguistiques tel que les mots les plus fréquents.

Les expériences menées nous montrent la fiabilité de l'approche de n-grammes basée sur la distance du χ^2 puisque le taux de succès est de 100% lorsque les textes à identifier sont de taille supérieure à 100 caractères (entre 8 et 15 mots). Les expériences ont été menées sur des textes très courts.

Dans le cas de fréquences moins bien définies, les 1-plus proches voisins semblent meilleurs que les RBF, en particulier avec de très petits échantillons d'apprentissage. Les résultats de [Dunning, 1994] utilisant des modèles de Markov, avec deux langues au lieu de 5 ici, suggèrent que les modèles de Markov entraînés avec $50Ko$ de données ont environ les mêmes performances que les 1-plus proches voisins avec $25Ko$. Le taux de succès en cas de réponse aléatoire étant de 20% dans le cas de 5 langues et de 50% dans le cas de 2 langues. En rappelant que les deux langues utilisées par Dunning sont testées sur le même échantillon, nous supposons que les 1-plus proches voisins sont plus adaptés pour ce problème.

Chapitre 8

Traduction automatique

Sommaire

8.1. Introduction

Comme nous utilisons un traducteur automatique dans nos chaînes de traitement, nous plaçons ici un court chapitre faisant le point sur ce sujet difficile : histoire, espoirs passés, solutions actuelles et critères d'évaluation. On verra que nous n'avons pas besoin d'un traducteur parfait, qui n'existera sans doute jamais, car les outils actuellement disponibles nous suffisent pour retrouver les thèmes d'un texte et donc pour catégoriser automatiquement les textes multilingues, de façon assez correcte.

La plupart des grands projets de traduction automatique sont nés entre 1958 et 1966 des besoins de traduction à partir du russe engendrés par la guerre froide. Le système Systran, utilisé par la *Foreign Technology Division* de l'armée de l'air américaine, est l'un des systèmes utilisés pour cette tâche. Ce système utilisait un énorme dictionnaire russe-anglais couvrant un grand nombre de domaines. La sortie de ce système était ensuite examinée par des spécialistes qui étaient chargés d'isoler les textes ayant une valeur stratégique ou scientifique pour les passer ensuite à des traducteurs humains.

En 1966, la recherche en traduction automatique a connu un coup d'arrêt avec la sortie du rapport *ALPAC de la National Science Foundation* qui concluait à l'impossibilité d'une traduction automatique de qualité. La plupart des chercheurs américains se sont alors tournés vers une science émergente : l'intelligence artificielle [Chandioux, 1998].

8.2. Premières approches historiques de la traduction automatique

[Chandioux, 1998] présente un panorama des approches utilisées pour la traduction automatique. Historiquement, trois approches sont expérimentés :

8.2.1. Décryptage

Les développeurs de la fin des années cinquante concevaient la traduction automatique comme un simple problème de décryptage. A chaque mot du texte incompréhensible correspondait un mot de la langue connue.

Cette approche a montré ses limites de fait que la traduction n'était pas biunivoque, qu'un même mot pouvait avoir plusieurs catégories grammaticales et qu'à chaque catégorie grammaticale pouvaient correspondre plusieurs sens.

8.2.2. Analyse par micro-contexte

Des tests de compréhension ont révélé que si un locuteur humain lisait un texte en déplaçant une feuille de carton avec un trou qui laissait paraître quelques mots contigus, beaucoup d'ambiguïtés pouvaient être levées. Ceci a donné naissance à

l'analyse par micro-contexte, qui est le principe de fonctionnement des systèmes de traduction automatique dits de première génération comme le système Systran.

8.2.3. Imiter la traduction humaine

Les systèmes de traduction automatique dites de deuxième génération s'inspirent de la traduction humaine. La traduction humaine se décompose habituellement en trois étapes : la compréhension du message en langue source, la transposition de la teneur du message en langue cible et la formulation du message selon les règles de la langue cible. Parallèlement, un système de traduction automatique de deuxième génération opère au niveau de la phrase et la traite en trois étapes : l'analyse, le transfert et la génération. Les chercheurs du GETA (Groupe d'étude pour la traduction automatique) à Grenoble ont été les premiers à jeter les bases de cette nouvelle génération.

Malheureusement, on n'est pas parvenu à développer un système général de traduction automatique de deuxième génération, même en se limitant aux textes scientifiques et techniques [Chandioux, 1998].

8.3. Nouvelles approches, plus modestes

Les nouvelles approches sont moins ambitieuses et tentent de se limiter à certains types de textes, à certains domaines, ou encore en réduisant l'ambiguïté des textes lors de leur rédaction. Dans les prochains paragraphes, nous présentons les solutions proposées :

8.3.1. Mémoire de traduction

Il s'agit d'une base de données qui stocke les phrases issues de deux langues, source et cible. Tout ce qu'on a traduit est contenu dans deux entités : le texte original dans la langue source et sa traduction antérieure dans la langue cible. Lorsque on travaille sur une nouvelle traduction, le système de mémoire de traduction recherche chaque phrase du texte original dans la base de données et, si la recherche est fructueuse, la phrase est traduite et il ne reste qu'à confirmer la traduction.

De nombreux systèmes commerciaux de mémoire de traduction sont proposés comme *Translation Manager 2* d'IBM, *Translator's Workbench* de Trados ou *Star Transit* qui mémorisent toutes les phrases au fur et à mesure de leur traduction.

Un système à base de mémoire de traduction ne peut réaliser lui même aucune traduction ; c'est un support qui aide le traducteur humain dans son travail. Bien que les systèmes de mémoire de traduction soient réellement utiles sur du texte répétitif comme les contrats d'assurances, les conventions collectives ou les descriptions de tâches, ils sont moins utiles dans des textes généraux. [Volk, 1998] compare les systèmes à base de mémoire de traduction avec les systèmes de traduction automatique.

8.3.2. Sous-langages et langages contrôlés

On appelle sous-langage un sous-ensemble de la langue ayant ses propres règles syntaxiques et un vocabulaire limité dont les relations sémantiques sont cernables. Le traducteur automatique METEO est un exemple connu d'application de la traduction automatique à un sous-langage. Ce système a pour vocation de traduire les prévisions publiques, agricoles et maritimes émises par Environnement Canada [Chandioux, 1998].

D'une façon générale, cette approche part de l'idée simple suivante : si on ne peut pas traduire n'importe quel texte, pourquoi ne pas écrire en fonction de la machine quand on sait que le texte devra être disponible en plusieurs langues.

« *Les premières recherches sur les langages contrôlés ont porté sur l'anglais technique, le but étant d'obtenir des documentations non ambiguës tant pour des utilisateurs anglophones que pour des utilisateurs dont l'anglais n'est pas la langue maternelle. Un bon exemple est le "Simplified English" développé par l'AECMA (Association européenne de constructeurs de matériel aéronautique) et dont l'usage est recommandé par l'ATA (American Transport Association). Le GIFAS (Groupement des Industries Françaises Aéronautiques et Spatiales) mène depuis quelques années des travaux similaires sur le « français rationalisé » »* [Chandioux, 1998].

8.4. Évaluer la traduction automatique

Le problème de l'évaluation de la traduction est difficile. Il y a beaucoup de littérature sur cette question [Hovy et al., 2002]. Le gouvernement américain finance depuis quelque temps des recherches sur les évaluations comparatives, du genre des TREC (*The Text REtrieval Conference*, voir `http://trec.nist.gov/`).

Dans les années antérieures, on a expérimenté des méthodes d'évaluation complexes et coûteuses qui font appel à des tests objectifs de compréhension ainsi qu'à des évaluations subjectives de la fluidité du texte traduit. On compare plusieurs traductions humaines et machines sur ces plans. Ces évaluation tentent de juger si un système de traduction automatique est fiable ou non. Pour ce faire elles s'intéressent à la qualité des traductions que le système produit, sa rapidité d'exécution, son coût, le temps qu'il exige pour la post-édition, sa convivialité, ou encore son potentiel de développement. D'autres questions se posent : que faut-il utiliser pour tester les systèmes : des corpus réels ou des corpus fabriqués ? Serait-il utile d'élaborer une méthodologie générale d'évaluation ? Voir [Cormier, 1992, King, 1992] pour une étude complète sur ce sujet. Notons que, vu le nombre limité de systèmes, ces évaluations ne concernaient qu'un seul système à la fois.

Récemment, suite à la disponibilité de plusieurs systèmes de traduction, une étude comparative de ces système est devenue possible. A titre d'exemple, [Volk, 1997] propose une évaluation de plusieurs systèmes commerciaux de traduction automatique pour les langues anglais-allemand. Il évalue la performance de six systèmes à

savoir :

1. German Assistant in Accent Duo V2.0 (développer par MicroTac/Globalink ; distributeur : Accent)
2. Langenscheidts T1 Standard V3.0 (développer par GMS ; distributeur : Langenscheidt)
3. Personal Translator plus V2.0 (développer par IBM ; distributeur : von Rheinbaben & Busch)
4. Power Translator Professional (développeur et distributeur : Globalink)
5. Systran Professional for Windows (développer par Systran S.A. ; distributeur : Mysoft)
6. Telegraph V1.0 (développeur et distributeur : Globalink)

Cette évaluation est basée sur des corpus artificiels (en anglais *test suites*), qui regroupent des phrases dont la construction syntaxique est représentative des textes d'un client ; ils sont très utiles au développeur qui peut, grâce à eux, vérifier la force réelle de son système.

Une tendance assez forte à l'heure actuelle consiste à utiliser des mesures extrêmement simples qui semblent avoir une très bonne corrélation avec les méthodes coûteuses. [Papineni et al., 2002], par exemple, présente une technique simple et rapide pour évaluer les performances de traducteurs automatiques. Cette technique est basé sur l'idée simple que si l'on souhaite évaluer la performances d'un système de traduction (ST), on doit comparer les résultats de traduction du système avec ceux produits par un traducteur humain (TH). Plus le texte du ST est proche de celui TH plus la traduction est bonne. La proximité est mesurée par une métrique. [Papineni et al., 2002] propose une mesure appelé BLEU (pour *BiLingual Evaluation Understudy*) qui consiste à calculer le nombre de mots communs entre les phrases fournit par le ST et le TH. Leurs résultats semblent consistants et les jugements donnés par leur système correspondent à ceux proposés par les THs. Ceci est très intéressant car il permet aux développeurs de systèmes de traduction d'évaluer leurs systèmes et tester leurs hypothèses rapidement avec les moindres coûts possibles.

8.5. Conclusion

Les traducteurs automatiques présentent une fiabilité toute relative. Aucun n'offre de résultats parfaits, même si plusieurs procédés différents sont mis en oeuvre. Les traductions sont approximatives.

Ce chapitre a pour objectif de montrer les approches successives proposées pour la traduction automatique. Nous avons montré que l'objectif initial était trop ambitieux : dans les années 60, on rêvait d'une machine rendant des textes parfaitement traduits. De nouvelles approches, qui tentent de se limiter à certaines taches et à certains types de textes, sont alors proposées. Un autre objectif de ce chapitre était de

montrer comment les traducteurs automatiques sont évalués. En fin, nous terminons ce chapitre par une phrase qui résume, à notre sens, les tendances actuelles pour la traduction automatique :

« *Si l'ordinateur ne comprend pas ce que nous écrivons, écrivons dans un langage qu'il comprend et laissons la littérature aux humains ...* » [Chandioux, 1998].

i

Chapitre 9

Cadre pour la catégorisation de textes multilingues

Sommaire

9.1. Introduction

Dans ce chapitre, nous proposons des solutions pour étendre la catégorisation de textes aux corpus multilingues. Ceci introduit des contraintes supplémentaires dans la catégorisation de textes : il faut reconnaître automatiquement la langue d'un texte puis procéder à une traduction automatique. Notre approche semble naïve, mais elle a le mérite d'être 1) la première solution automatique proposée, à notre connaissance, 2) opérationnelle, comme le montrent les premières expérimentations sur un corpus d'articles de journaux allemands, anglais et français.

La phase d'apprentissage s'effectuera, comme en monolingue, à partir d'un corpus d'apprentissage étiqueté rédigé dans une langue donnée $\mathfrak{L}_{app}$. Mais l'inférence sera possible pour un texte rédigé dans une langue quelconque, dès qu'un traducteur automatique sera disponible de cette langue vers $\mathfrak{L}_{app}$. Notons que notre approche exclut les méthodes qui utilisent de manière explicite des informations spécifiques à chaque langue.

Bien entendu, à l'instar de la catégorisation de textes monolingue, le texte à classer doit appartenir au même domaine que les textes 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 constitués d'articles de journaux à scandale.

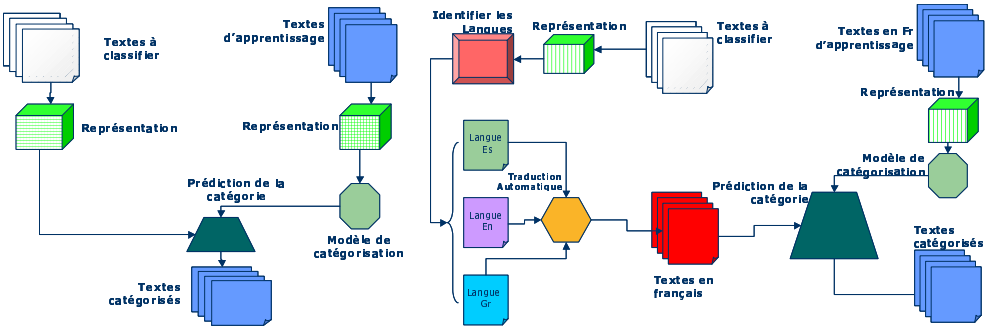
Ce chapitre est organisé de la manière suivante : dans la section 9.2, nous exposerons l'approche que nous avons pour étendre la catégorisation de textes au cas multilingue. Dans la section 9.3, nous mettrons en oeuvre le cadre proposé sur un exemple réel de catégorisation de journaux. Dans la section 9.4, nous discutons des résultats obtenus, et nous essayerons de mettre en perspective notre démarche afin de la faire évoluer. Nous conclurons alors dans la section 9.5.

9.2. Méthodes pour la catégorisation de textes multilingues

9.2.1. Nouveau cadre pour la catégorisation multilingue

Dans le cadre de la catégorisation de textes multilingues, le processus comporte deux nouvelles exigences : le corpus de textes étiquetés utilisé pour l'apprentissage est disponible dans une langue $\mathfrak{L}_{app}$ donnée ; chaque nouveau texte à classer est dans une langue que l'on doit d'abord déterminer avant de pouvoir lui associer son étiquette.

Pour répondre à ces nouvelles contraintes, nous aménageons le processus de catégorisation exposé dans la section 1.3 page 9. La phase d'apprentissage n'est pas modifiée, en revanche la phase de classement comporte deux étapes supplémentaires (figure 9.1.b) :



(a) Catégorisation de textes monolingue (b) Catégorisation de textes multilingue

Figure 9.1.: Schémas généraux de catégorisations mono et multilingue

1. nous devons tout d'abord détecter la langue dans laquelle le texte d_j à classer est rédigé ;
2. si la langue est reconnue par le traducteur, il est traduit vers la langue $\mathfrak{L}_{app}$ et le texte à classer devient $d_{j_{app}}$.

Nous recherchons alors les occurrences des termes $(w_{1j}, w_{2j}, ..., w_{|\mathcal{T}|j})$ dans $d_{j_{app}}$afin de pouvoir appliquer le modèle prédictif et ainsi associer une catégorie c_i au texte à classer.

9.2.2. Détection de la langue du texte à classer

Il est important de détecter avec précision la langue dans laquelle le texte à classer est rédigé, car une erreur à ce niveau voue à l'échec les étapes suivantes.

Il existe deux familles d'approches dans l'identification de la langue : linguistique ou statistique. Afin de rester cohérent avec nos choix pour la catégorisation de textes, nous avons privilégié l'approche statistique. Cette approche capture automatiquement certaines régularités statistiques des langues. Comme pour la catégorisation de textes, nous avons utilisé les 3-grammes qui sont des séquences de 3 caractères consécutifs extraits du texte à classer. Nos précédents travaux ont montré qu'un texte de longueur de 100 octets permettait d'obtenir une reconnaissance d'excellente qualité, à près de 99% [Teytaud and Jalam, 2001]. Et pour des textes plus longs, la reconnaissance est parfaite.

9.2.3. Traduction du texte à classer

La traduction du texte à classer dans la langue du corpus d'apprentissage $\mathfrak{L}_{app}$ est également une étape primordiale. L'objectif ici n'est pas de produire un texte traduit retraçant fidèlement les propriétés sémantiques de l'original, mais de fournir un texte

assurant une qualité de classement suffisante. Il est évident que le résultat obtenu dépendra du traducteur utilisé, une des perspectives immédiates de notre travail actuel consistera à analyser de manière approfondie le traducteur pour évaluer son efficacité lors de la catégorisation de textes.

Afin de défricher le terrain, nous avons utilisé un traducteur en ligne disponible sur Internet (`http://babelfish.altavista.com/`, qui utilise la technologie Systran). Ce traducteur n'est sans doute pas le meilleur, mais comme il est publiquement disponible, nous aurons la possibilité par la suite de comparer nos résultats avec d'autres études. L'utilisation d'un meilleur traducteur ne peut qu'améliorer la catégorisation des textes.

9.3. Application sur les corpus CLEF

9.3.1. Constitution du corpus

La constitution de ce corpus a été un travail long et difficile ; malgré nos recherches, nous n'avons pas trouvé de corpus multilingues de textes étiquetés en classes comparables. Nous avons dû adapter pour cela les documents proposés par les organisateurs du concours CLEF (voir la page web : `http://clef.iei.pi.cnr.it`).

Les corpus de documents utilisés dans la campagne d'évaluation CLEF proviennent de différents journaux tels que le Los Angeles Times (États-Unis), Le Monde (France), La Stampa (Italie), Der Spiegel et Frankfurter Rundschau (Allemagne), d'agences de presse comme EFE (Espagne) ou des dépêches de l'agence télégraphique suisse (disponibles en allemand, français et italien). Les documents de ces corpus sont tous extraits de l'année 1994 et les thèmes abordés sont approximativement comparables.

Ces corpus sont destinés aux tâches de la recherche documentaire (RI) monolingues, bilingues et multilingues. Les fichiers de ces corpus sont au format SGML, par exemple, le fichier `lemonde_19940604.sgml` concerne les articles apparus dans le journal Le Monde (LM) lors de la journée du 4 juin 1994. Chaque article de ce fichier contient des balises sgml qui décrivent son contenu (table 9.1). Il est facile d'y distinguer son numéro d'identification, son titre, et son corps.

La taille des corpus varie fortement entre les langues, avec des volumes plus restreints pour le français et l'italien. Le nombre de mots par article reste assez similaire (environ 130), avec une moyenne un peu plus élevée pour la collection anglaise (167). Par contre, la variabilité de cette longueur demeure assez forte (écart-type d'environ 120), sauf pour les langues espagnole, où l'écart-type est de 60, et italienne, où l'écart-type est de 97 [Savoy, 2002].

Comme on le sait, la catégorisation de textes nécessite d'avoir un échantillon d'apprentissage composé de textes associés à leurs étiquettes (ou classes). Ces exemples serviront, dans le processus d'apprentissage, à apprendre un classifieur (ou modèle) qui sera ensuite appliqué sur les textes à catégoriser. Une première approche peut

```
<DOC>
<DOCNO>LEMONDE94-000386-19940604</DOCNO>
<DOCID>LEMONDE94-000386-19940604</DOCID>
<ACCOUNT>339369</ACCOUNT>
<GENRE>RECTIF</GENRE>
<DATE>19940604</DATE>
<LMDOC>MHB</LMDOC>
<FAB>06031011</FAB>
<SUBJECTS>DEMENTI,DESSIN</SUBJECTS>
<GO21>PUBLICATION</GO21>
<NAMES>SERGUEI</NAMES>
<PUM1>QUO</PUM1>
<REFERENCE1>2-002-06</REFERENCE1>
<SEC1>IDE</SEC1>
<PAGE>13</PAGE>
<TITLE>Serguei précise</TITLE>
<TIO1>PAS DE PANIQUE A BORD</TIO1>
<TEXT>Un lecteur s'est étonné de constater qu'une publication satirique d'extrême droite, Pas de
panique à bord, citait, parmi les noms de ses collaborateurs, celui du dessinateur Sergueï. Etonne-
ment encore plus grand _ pour ne pas dire plus _ de Sergueï, celui que nos lecteurs connaissent
bien et qui ne saurait être le même que son homonyme de Pas de panique à bord, s'il existe.
&gt ;</TEXT>
</DOC>
```

TAB. 9.1.: Exemple extrait de la collection CLEF et qui montre un article du journal Le Monde publié le 4 juin 1994

consister à utiliser les termes-index associés à chaque article comme classes d'appartenance. En général, l'indexation est une opération qui *décrit* et *caractérise* un document, ou un fragment de document, en repérant les thèmes présents dans ce document [Afnor, 1993]. Malheureusement, les résultats de nos expérimentations utilisant les termes-index, sur les corpus français et allemand, sont médiocres : le taux d'erreur est proche de celui du hasard.

Il est difficile d'apprendre si l'on utilise les termes-index proposés en tant que classes pour différentes raisons :

- Les termes-index associés aux articles français et allemand sont trop nombreux, plus de 16200 termes (et donc classes) pour le corpus français et plus de 32000 termes pour le corpus SDA allemand. Il n'y a pas eu de règle d'indexation basée sur des vocabulaires contrôlés et beaucoup de termes sont des synonymes (mort, morts ; France, fr ; German, gr).
- Les termes-index proposés ne représentent vraiment pas les thèmes abordés dans les articles ; par exemple, on associe à la dépêche de la table 9.1 le mot clé dessin alors qu'elle parle d'un démenti.
- Les termes-index dans les dépêches de l'agence télégraphique suisse (allemand et français) ne sont pas séparés par des signes de ponctuation ainsi, il est difficile d'extraire les termes composés comme "*conseil de sécurité*". Le corpus de Los Angeles Times, à la différence des autres corpus, ne fournit pas de tout de termes-index, ces termes-index aident normalement à décrire le contenu d'un

Classes	CLEF id	CLEF topic	#LAT	#LM	#SDA
C_1	92	U.N. sanctions against Iraq	27	24	23
C_2	95	Conflict in Palestine	96	89	66
C_3	103	Conflict of Interests in Italy	10	24	70
C_4	108	Southern Yemen Secession	18	19	63
C_5	119	Destruction of Ukrainian nuclear weapons	54	33	55
C_6	122	North American car industry	27	23	6
C_7	124	Common foreign and security policy (CFSP)	32	48	24
C_8	131	Intellectual Property Rights	40	43	43
C_9	133	German Armed Forces Out-of-area	10	21	17
C_{10}	140	Mobile phones	70	95	23
Total			**384**	**419**	**390**

TAB. 9.2.: Description des catégories utilisées

article.

Dans nos expérimentations, nous choisissons donc de considérer les thèmes proposés dans la campagne CLEF 2002 comme classes à prédire. Ces dernières sont très variées ; on trouve par exemple : “U.N. sanctions against Iraq”, “Conflict in Palestine” ou “Leaning Tower of Pisa”. Nous avons travaillé sur trois corpus de langues anglaise, française et allemande : le Los Angeles Times (LAT), Le Monde (LM) et l’agence télégraphique suisse (SDA). Les thèmes utilisés dans nos évaluations sont décrits dans la table 9.2.

9.3.2. Représentation des textes

La représentation des textes est une étape critique. Nous avons choisi d’utiliser les mots et les 3-, 4- et 5-grammes. Nous avons appliqué notre algorithme de sélection de termes χ^2_{multi} présentées dans la section 3 page 33 [Clech et al., 2003] en sélectionnant 100 mots avec pour seuls prétraitements l’uniformisation de la casse et la suppression des StopWords. En outre, nous avons sélectionnés 200 3-, 4- et 5-grammes avec pour seul prétraitement l’uniformisation de la casse. Nous avons choisi de sélectionner moins de mots que de n-grammes puisqu’un mot est composé de plusieurs n-grammes ; après plusieurs expériences, le choix de 100 mots et 200 n-grammes apparaît comme un bon compromis conservant la structure informationnelle du corpus.

L’étude des termes sélectionnés révèle la présence importante de noms propres, tels les noms des pays ou de leurs ressortissants, ou encore les noms de personnalités. L’étude indique également certaines difficultés de traduction. Par exemple, l’expression française "téléphone portable" est traduite en anglais par "portable phone" ce qui n’a pas le sens voulu. Les noms propres ne sont pas non plus épargnés par des difficultés de traduction ; ainsi le terme français "Koweït" est laissé tel quel au lieu

	10-V.C. LAT		10-V.C. LM		10-V.C. SDA	
taux d'erreur	**C4.5**	**3-NN**	**C4.5**	**3-NN**	**C4.5**	**3-NN**
100 mots	16%	8%	24%	5%	15%	3%
200 3-grammes	23%	9%	20%	9%	11%	2%
200 4-grammes	16%	7%	16%	5%	8%	2%
200 5-grammes	14%	7%	16%	5%	9%	2%

TAB. 9.3.: Taux d'erreur, en validation croisée (V.C.), pour les articles de journaux dans leur langue originelle

d'être traduit par le terme "Kuwait".

9.3.3. Algorithmes d'apprentissage

Dans notre application, nous utilisons deux méthodes renvoyant à des paradigmes d'apprentissage très différents (voir les sections 5.3 page 55 et 5.4 page 62) :

- une méthode arborescente : C4.5 [Quinlan, 1993]. Cette algorithme glouton a la particularité de sélectionner les variables et effectue des découpages parallèles aux axes ;
- une méthode d'estimation des probabilités locales : les 3 plus proches voisins (3-ppv) [Aha et al., 1991]. Cet algorithme est sensible aux variables non pertinentes ainsi qu'aux espaces de représentations creux [Mitchell, 1997].

9.3.4. Reconnaissance de la langue

Dans nos expérimentation, nous avons utilisé la distance de χ^2 pour identifier la langue de texte parmi les trois langues français, anglais et allemand. Nos nouveaux tests confirment nos résultats précédents [Teytaud and Jalam, 2001] : pour des textes de taille égale ou supérieure à 100 caractères le taux de reconnaissance de la langue est de 100%.

9.3.5. Catégorisation des articles

Afin de pouvoir évaluer notre processus de catégorisation de textes dans un corpus multilingue, nous avons effectué plusieurs expériences de catégorisation. Pour chacune d'elles, nous avons évalué le taux d'erreur pour toutes les configurations de représentations des textes (100 mots, 200 3-grammes, 4-grammes et 5-grammes) et des modèles utilisés (C4.5 et les 3 plus proches voisins).

Usuellement, en catégorisation de textes, un document peut appartenir à n classes. Cependant, dans l'application présentée ici (voir la table 9.2), chaque document n'appartient qu'à une classe, et une seule. Il en résulte que les rappel et précision "micro-

	LM (An)		SDA (An)	
Taux d'erreur	**C4-5**	**3-NN**	**C4-5**	**3-NN**
100 mots	25%	6%	12%	3%
200 3-grammes	16%	6%	9%	3%
200 4-grammes	12%	6%	9%	2%
200 5-grammes	13%	5%	12%	3%

TAB. 9.4.: Taux d'erreur, en validation croisée, pour les articles de journaux traduits en anglais. LM (An) désigne le corpus français Le Monde (LM) traduit en anglais (An). SDA (An) désigne le corpus allemand de l'agence télégraphique suisse (SDA) traduit en anglais (An)

moyennes" sont identiques et égaux au taux de succès ; pour cette raison nous allons aussi utiliser le taux d'erreur pour mesurer les performances.

Nous présentons tout d'abord les résultats en catégorisation monolingue (table 9.3), ils ont servi de référence pour juger de la qualité de ceux obtenus dans un contexte multilingue.

Catégorisation monolingue

Afin d'évaluer le niveau intrinsèque de difficulté de catégorisation des articles, nous avons mesuré l'erreur de classement pour nos 3 corpus (LAT, LM et SDA) dans leur langue d'origine.

Les résultats (par *10-validation croisée* [Stone, 1974]) sont présentés dans la table 9.3) ; nous en dégageons 4 principaux résultats :

- Il y a un apprentissage effectif puisque le taux d'erreur est largement inférieur aux taux d'erreur du classifieur par défaut ;
- Il y a un avantage important pour la méthode du 3-ppv (un écart supérieur à 10 points) par rapport à C4.5 qui souffre de la fragmentation des données ;
- Le bon apprentissage du 3-ppv laisse penser que les termes sélectionnés sont pertinents ;
- Le corpus allemand est plus facile à catégoriser que les deux autres.

Catégorisation multilingue

Comme décrite précédemment, la catégorisation multilingue nécessite des étapes intermédiaires complémentaires, chacune pouvant générer du biais pour le processus même de la catégorisation. En effet, si l'étape de traduction est de qualité médiocre, l'apprentissage sera difficile et les résultats de la catégorisation seront de mauvaise qualité. Par ailleurs, comme nous avons dit dans la section 9.3.2, les noms propres sont très présents dans les mots sélectionnés, et de ce fait nous pourrions nous demander si de bons résultats de catégorisation ne seraient pas simplement dus à ces noms

propres. Ainsi, dans l'objectif de valider notre processus de catégorisation multilingue, nous évaluons d'abord l'effet de la traduction, puis l'effet potentiel des noms propres et enfin la capacité même du modèle à être appliqué sur des textes traduits.

Les effets du traducteur Pour mesurer l'effet du traducteur sur le contenu informationnel des documents, nous avons traduit vers l'anglais le corpus français LM et le corpus allemand SDA. Nous avons appliqué le schéma d'apprentissage monolingue (voir la figure 9.1.a) et évalué l'erreur en validation croisée (table 9.4). Les résultats montrent que le contenu informationnel des corpus, du moins ce qui est nécessaire à la catégorisation, est très peu dégradé par la traduction. En effet, les différences des taux d'erreurs obtenus après traduction (table 9.4) comparées à ceux obtenus dans la langue d'origine (table 9.3) ne sont pas significatives.

Les effets de noms propres Pour évaluer comment les noms propres peuvent influencer les résultats de l'étape d'apprentissage, nous avons estimé un modèle à partir du corpus LAT dans sa langue d'origine (anglais) que nous appliquons directement sur les corpus LM et SDA laissés dans leur langue d'origine (respectivement français et allemand). Nous supposions que, si les résultats ne sont pas significativement différents de ceux obtenus après traduction, alors, la contribution des noms propres lors de l'étape de catégorisation est supérieure à la contribution des mots communs traduits.

Enfin, nous avons étudié les résultats de catégorisation après traduction en anglais des corpus à classer (LM et SDA) en appliquant le modèle élaboré à partir du corpus du LAT en anglais.

Nous présentons seulement les résultats concernant la représentation utilisant 100 mots et celle utilisant 200 4-grammes. Le tableau 9.5 regroupe les résultats obtenus pour le corpus LM et le tableau 9.6 ceux de SDA. Nous en dégageons 3 résultats principaux :

- Le premier concerne la viabilité de notre approche : même si le taux d'erreur s'accroît quand on passe d'un apprentissage sur les traductions anglaise au lieu des textes originaux, la qualité de prédiction surpasse largement celle du classifieur par défaut.
- Le second concerne la faible variabilité des résultats obtenus en fonction des représentations de texte utilisées (mots ou n-grammes) : on ne peut pas dire si l'une est plus robuste que l'autre.
- La troisième concerne le faible biais introduit par les noms propres ; les tableaux 9.5 et 9.6, montrent que l'écart entre les résultats avant et après traduction sont suffisamment significatifs : nous observons pour le modèle 3-PPV un saut de plus de 10 points pour le corpus LM (Tableau 9.5) et un saut de plus de 70 points pour le corpus SDA (Tableau 9.6) ; rappelons que nous supposions que, si les résultats du modèle estimé pour l'anglais mais appliqué à des textes en français ou allemand, n'étaient pas significativement inférieurs de ceux obtenus

(a) représentation avec 100 mots

	Appris et appliqué sur LM (Fr)		**Appris sur LAT (An) et appliqué sur**			
	LM (Fr)		**LM (Fr)**		**LM (An)**	
	3-NN	**C4.5**	**3-NN**	**C4.5**	**3-NN**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	95%	76%	78%	12%	89%	60%
ρ^{M}	94%	68%	78%	14%	92%	55%
π^{M}	92%	71%	80%	10%	88%	52%

(b) représentation avec 200 4-grammes

	Appris et appliqué sur LM (Fr)		**Appris sur LAT (An) et appliqué sur**			
	LM (Fr)		**LM (Fr)**		**LM (An)**	
	3-NN	**C4.5**	**3-NN**	**C4.5**	**3-NN**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	95%	84%	86%	55%	90%	68%
ρ^{M}	96%	80%	81%	43%	89%	59%
π^{M}	95%	79%	90%	32%	91%	50%

TAB. 9.5.: Précision et Rappel (micro et macro-moyen), pour le corpus Le Monde (LM) représenté par 100 mots (table a) et pour 200 4-grammes (table b). Les deux tables montrent les performances obtenues en validation croisée. On applique le modèle LM (Fr) sur des textes écrits en français et les résultats obtenus sont comparés avec ceux du modèle LAT anglais sur le corpus “LM écrit en français” et sur le corpus “LM traduit en anglais”

(a) représentation avec 100 mots

	classifieur SDA (Ger) appliqué sur		**classifieur LAT (An) appliqué sur**			
	SDA (Ger)		**SDA (Ger)**		**SDA (An)**	
	3-NN	**C4.5**	**3-NN**	**C4.5**	**3-NN**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	97%	85%	20%	17%	97%	56%
ρ^{M}	93%	77%	13%	14%	95%	50%
π^{M}	95%	70%	12%	25%	94%	62%

(b) représentation avec 200 4-grammes

	classifieur SDA (Ger) appliqué sur		**classifieur LAT (An) appliqué sur**			
	SDA (Ger)		**SDA (Ger)**		**SDA (An)**	
	3-NN	**C4.5**	**3-NN**	**C4.5**	**3-NN**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	98%	92%	21%	17%	97%	54%
ρ^{M}	94%	80%	14%	14%	97%	52%
π^{M}	94%	78%	12%	25%	96%	54%

TAB. 9.6.: Précision et Rappel (micro et macro-moyen) pour le corpus allemand SDA représenté par 100 mots (table a) et pour 200 4-grammes (table b). Les deux tables montrent les performances obtenues en validation croisée : On applique le modèle SDA (Ger) sur des textes écrits en allemand et les résultats obtenus sont comparés avec ceux du modèle LAT anglais sur le corpus “SDA écrit en allemand” et sur le corpus “SDA traduit en anglais”

sur ces textes traduits, alors, la contribution des noms propres lors de l’étape de catégorisation serait supérieure à la contribution des mots communs traduits.

Par ailleurs : le taux d’erreur de C4.5 s’explique largement par la difficulté de prédire les classes c_3, c_6 et c_9 dont les taux de rappels et de précisions sont nuls (voir les tables 9.7 et 9.8). Pour c_3 et c_9 cela est dû au seuil d’élagage (fixé à 10) correspondant au nombre de textes du corpus d’apprentissage (LAT) composant ces classes (voir la table 9.2). Pour c_6, C4.5 produit une seule règle, basée sur la présence du mot ’auto’, qui provient des expressions “*auto manufacter*s” ou “*auto shows*”, dans le corpus du LAT ; or les expressions françaises “*salon de l’auto*” et “*industrie automobile*” des articles du Monde (LM) sont respectivement traduites par “*car show*” et “*car industries*” : le terme “*auto*” étant traduit par “*car*”, le terme “*auto*” est absent des traductions de LM ; C4.5 ne peut donc appliquer son (unique) règle apprise.

9.4. Discussion

Les résultats obtenus par nos modèles sur notre corpus sont des résultats encourageants. Cette section propose de discuter les différents choix effectués pour chacune des étapes du processus afin de moduler la signification de nos résultats.

La première étape du processus consiste à définir une représentation du corpus par des termes. Nous avons choisi ici la représentation basée sur les mots et celle basée sur les n-grammes. Ce dernier choix était motivé par la capacité des n-grammes à capturer aisément les structures informationnelles basiques en s'affranchissant des problèmes de séparation des mots, de coquilles et tout autre aspect linguistique. Nos expériences n'ont pas montré une différentiation marquée entre les résultats issus d'une sélection de mots et d'une sélection de n-grammes ; ceci montre. Nous attribuons ces similarités à la qualité contrôlée des corpus. En effet, ces derniers sont destinés à la presse écrite, qui est exigeante envers les fautes d'orthographes et coquilles.

La deuxième étape consiste en la sélection des termes. Les coprésences des mots sélectionnés à partir du corpus du LAT et de celui du LM traduit en anglais se situent au niveau de 50 %. La moitié d'entre-eux est constituée de mots communs. Nous obtenons des résultats similaires lors de la comparaison des termes sélectionnés à partir du corpus du LAT et de celui du SDA traduit en anglais. Ceci montre la similitude informationnelle de nos 3 corpus. Par ailleurs, la forte quantité de noms propres n'est pas un problème en lui-même, et nous avons vu que son apport était faible. Cependant, lorsque le traducteur ne connaît pas un terme il le laisse tel quel. De plus, les noms propres ont une faible variabilité d'écriture dans les différentes langues. Ainsi, nous pensons que les noms propres empêchent l'évaluation complète de l'effet de la traduction.

Enfin, la sélection des termes est qualitativement intéressante puisqu'elle permet une séparabilité aisée de nos 10 classes. Là encore, nous nous interrogeons sur la constitution de notre corpus. Le corpus CLEF contient 96 000 articles. De ces 96 000, seuls 8 000 ont été assigné à 50 sujets (classes). Pour améliorer notre corpus et confirmer nos résultats, nous envisageons la généralisation en travaillant sur l'ensemble des 96 000 textes (8 000 assignés à des classes définies et les 88 000 restants assignés à une classe "autre") ceci rendra plus difficile la séparabilité des sujets.

La dernière étape concerne l'apprentissage ; elle a montré les bons résultats du '3-ppv' sur ces corpus. Ce résultat est en opposition avec la difficulté de prédiction des k-ppv dans un espace creux et/ou avec des variables non pertinentes. Nous attribuons donc ces bons résultats des "3-ppv" à la qualité de notre espace de représentation (bon choix des descripteurs par notre méthode du χ^2_{multi} multivarié). Enfin, le C4.5 donne des résultats convenables mais est moins performant que le 3-ppv. Comme nous l'avons vu, cela est du au faible effectifs de certaines classes et au paramétrage de la méthode.

9.5. Conclusion

L'objet de ce chapitre est la définition d'un processus pour la catégorisation multilingue. Nous avons introduit deux nouvelles étapes par rapport au processus monolingue : la détection de la langue du texte à catégoriser et sa traduction dans la langue du corpus d'apprentissage. Nous avons illustré notre procédé par une application sur des corpus réels de journaux écrits en 3 langues (anglaise, française et allemande). Nous avons décrit chaque étape de notre processus et présenté les résultats de nos expériences. Nous concluons à la l'efficacité de notre approche.

Nous envisageons de perfectionner notre cadre pour la catégorisation de textes multilingues en proposant un schéma plus général dans lequel nous fusionnerons des corpus d'apprentissage traduits en une langue commune d'apprentissage (voir la figure 6.4 page 94). Nous espérons ainsi que les particularités propres à chaque langue ne seront plus retenues par les modèles estimés.

		LAT C.V.		LAT classifier applied on LM		LAT classifier applied on SDA		LM C.V.		SDA C.V.	
		3-NN	C4.5	3-NN	C4.5	3-NN	C4.5	3-NN	C4.5	3-NN	C4.5
c_1	ρ	100%	93%	100%	92%	100%	91%	100%	83%	100%	100%
	π	93%	96%	96%	100%	100%	84%	96%	87%	100%	88%
c_2	ρ	99%	97%	100%	94%	100%	97%	100%	93%	100%	95%
	π	98%	100%	100%	100%	100%	97%	100%	99%	100%	100%
c_3	ρ	100%	0%	96%	0%	97%	0%	96%	96%	100%	99%
	π	100%	0%	82%	0%	100%	0%	85%	88%	100%	99%
c_4	ρ	83%	100%	100%	84%	100%	81%	100%	11%	100%	95%
	π	100%	32%	100%	12%	100%	28%	100%	17%	100%	100%
c_5	ρ	96%	93%	100%	85%	100%	96%	97%	97%	100%	93%
	π	90%	81%	87%	78%	96%	100%	91%	100%	98%	100%
c_6	ρ	93%	70%	96%	0%	100%	17%	91%	30%	50%	0%
	π	93%	86%	88%	0%	50%	100%	88%	44%	75%	0%
c_7	ρ	72%	75%	90%	73%	100%	100%	92%	79%	96%	92%
	π	88%	80%	74%	51%	80%	41%	86%	70%	96%	88%
c_8	ρ	93%	93%	67%	84%	84%	5%	79%	63%	93%	98%
	π	84%	97%	76%	82%	95%	67%	89%	96%	98%	98%
c_9	ρ	80%	0%	95%	0%	100%	0%	95%	52%	100%	0%
	π	67%	0%	87%	0%	100%	0%	87%	44%	94%	0%
c_{10}	ρ	93%	79%	78%	29%	70%	17%	91%	80%	91%	96%
	π	98%	96%	97%	97%	100%	100%	98%	64%	84%	27%
μ	$\rho^\mu = \pi^\mu$	92%	85%	92%	54%	95%	50%	95%	76%	97%	85%
M	ρ^M	91%	70%	89%	52%	92%	62%	94%	68%	93%	77%
	π^M	91%	67%	90%	59%	96%	56%	92%	71%	95%	70%

TAB. 9.7.: Rappel et Précision sur le corpus anglais (LAT) décrit par 100 mots

		Appris et appliqué sur LAT, validation croisée		Appris sur LAT, appliqué à LM		Appris sur LAT, appliqué à SDA		Appris et appliqué sur LM, validation croisée		Appris et appliqué sur SDA, validation croisée	
		3-NN	C4.5	3-NN	C4.5	3-NN	C4.5	3-NN	C4.5	3-NN	C4.5
c_1	ρ	100%	96%	100%	96%	100%	100%	100%	100%	100%	96%
	π	100%	96%	92%	96%	100%	100%	100%	67%	100%	85%
c_2	ρ	99%	99%	100%	99%	100%	98%	100%	100%	100%	94%
	π	98%	99%	100%	98%	100%	97%	99%	100%	100%	98%
c_3	ρ	100%	0%	100%	0%	100%	0%	100%	96%	100%	99%
	π	100%	0%	92%	0%	100%	0%	92%	92%	100%	99%
c_4	ρ	94%	78%	100%	74%	100%	49%	100%	32%	100%	100%
	π	94%	34%	95%	16%	100%	25%	100%	35%	100%	100%
c_5	ρ	94%	94%	100%	100%	98%	96%	97%	100%	100%	96%
	π	88%	85%	89%	72%	95%	100%	97%	100%	100%	96%
c_6	ρ	89%	70%	96%	0%	100%	17%	100%	91%	67%	0%
	π	89%	86%	92%	0%	75%	100%	88%	88%	67%	0%
c_7	ρ	72%	69%	92%	67%	96%	88%	96%	85%	96%	71%
	π	82%	67%	72%	49%	79%	36%	82%	82%	88%	77%
c_8	ρ	90%	93%	65%	81%	79%	5%	77%	70%	86%	79%
	π	86%	93%	93%	74%	100%	50%	97%	88%	93%	72%
c_9	ρ	70%	0%	67%	0%	100%	0%	95%	29%	100%	65%
	π	88%	0%	88%	0%	89%	0%	91%	46%	100%	55%
c_{10}	ρ	94%	89%	91%	58%	91%	65%	93%	97%	91%	100%
	π	96%	95%	96%	90%	100%	25%	99%	94%	88%	96%
μ	$\rho^\mu = \pi^\mu$	93%	84%	91%	58%	96%	52%	95%	84%	98%	92%
M	ρ^M	90%	69%	91%	50%	94%	53%	96%	80%	94%	80%
	π^M	92%	66%	91%	67%	97%	54%	95%	79%	94%	78%

TAB. 9.8.: Rappel et Précision sur le corpus anglais (LAT) décrit par 200 4-grammes