Cadre pour la catégorisation de textes multilingues

Radwan Jalam, Jérémy Clech, Ricco Rakotomalala

Laboratoire ERIC – Université Lumière Lyon 2
5, av. Pierre Mendès-France – 69676 Bron, France
{jalam, jclech, rakotoma}@eric.univ-lyon2.fr

Abstract

In this paper, we propose an original framework for multilingual text categorization. The objective is to classify a set of texts, written in some language, using a predictive model learned from a set of texts written in a given language, called learning language. Contrary to the unilingual classical phase of text categorization, the classification phase contains two new steps : firstly identify the language of the text, and then automatically translate it into the learning language. As shown in this paper, first applications of multilingual text categorization on real data, that is over English, French and German newspapers, indicate that the approach is viable.

Résumé

Dans cet article, nous proposons un cadre pour la catégorisation de textes multilingues. L'objectif est de pouvoir, à partir d'un modèle de prédiction construit par apprentissage sur un corpus de textes rédigés dans un langue donnée, inférer sur une série de textes qui sont rédigés dans une langue quelconque. Cette phase d'inférence, par rapport à la généralisation classique, comprend deux étapes supplémentaires : la détection de la langue du texte, puis sa traduction automatique vers la langue de référence. Nos premiers résultats sur une application réelle : la catégorisation d'articles de journaux allemands, anglais et français, montrent la viabilité de l'approche.

Keywords: multilingual text categorization, language identification, machine learning, n-grams representation, CLEF collection, translation effects on text categorization.

1. Introduction

La catégorisation de texte consiste à chercher une liaison fonctionnelle entre *un ensemble de textes* et *un ensemble de catégories* (étiquettes, classes). Cette liaison fonctionnelle, que l'on appelle également *modèle de prédiction*, est estimée par un apprentissage automatique (traduction de *machine learning method*). Pour ce faire, il est nécessaire de disposer d'un ensemble de textes préalablement étiquetés, dit *ensemble d'apprentissage*, à partir duquel nous estimons les paramètres du modèle de prédiction le plus performant possible, c'est-à-dire le modèle qui produit le moins d'*erreurs* en prédiction (voir la figure 1).

Formellement, la catégorisation de texte consiste à associer une valeur booléenne à chaque paire $(d_j, c_i) \in \mathcal{D} \times \mathcal{C}$, où $\mathcal{D}$ est l'ensemble des textes et $\mathcal{C}$ est l'ensemble des catégories. La valeur V (Vrai) est alors associée au couple (d_j, c_i) si le texte d_j appartient à la classe c_i tandis que la valeur F (Faux) lui sera associée dans le cas contraire. Le but de la catégorisation de texte est de construire une procédure (modèle, classifieur) $\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{V, F\}$ qui associe une ou plusieurs catégories à un document d_j telle que la décision donnée par cette procédure « coïncide le plus possible » avec la fonction $\breve{\Phi} : \mathcal{D} \times \mathcal{C} \rightarrow \{V, F\}$, la vraie fonction qui retourne pour chaque vecteur d_j une valeur c_i (Sebastiani, 2002).

Dans ce travail, nous proposons des solutions pour étendre la catégorisation de textes aux *corpus*

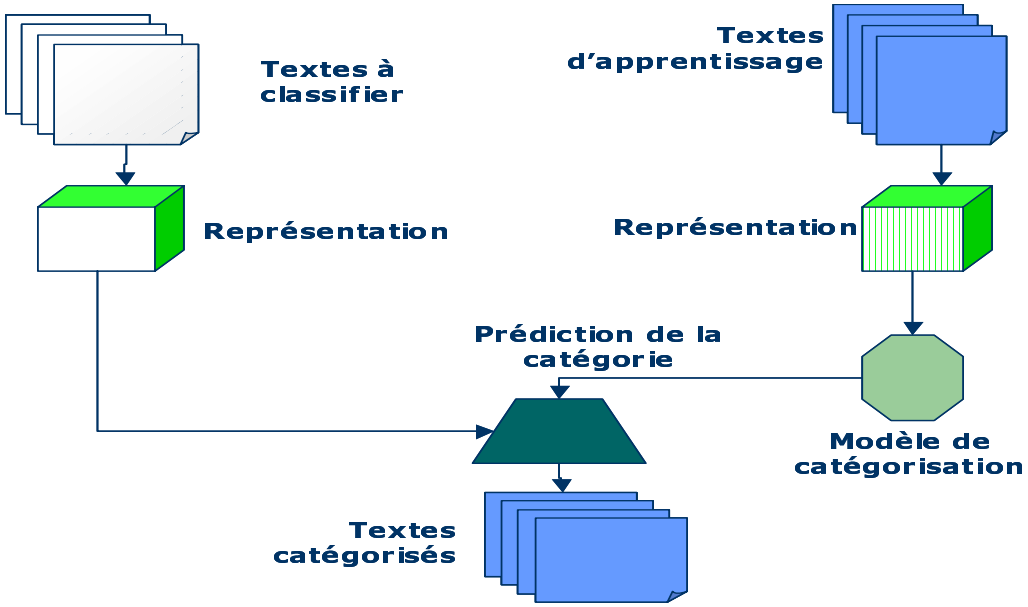

Figure 1. Schéma général pour la catégorisation de textes « monolingue ».

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 dans la catégorisation de texte « 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.

Cet article est organisé de la manière suivante : dans la section 2, nous exposerons l'approche que nous avons pour étendre la catégorisation de textes au cas multilingue. Dans la section 3, nous mettrons en oeuvre le cadre proposé sur un exemple réel de catégorisation de journaux. Dans la section 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 5.

2. 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.

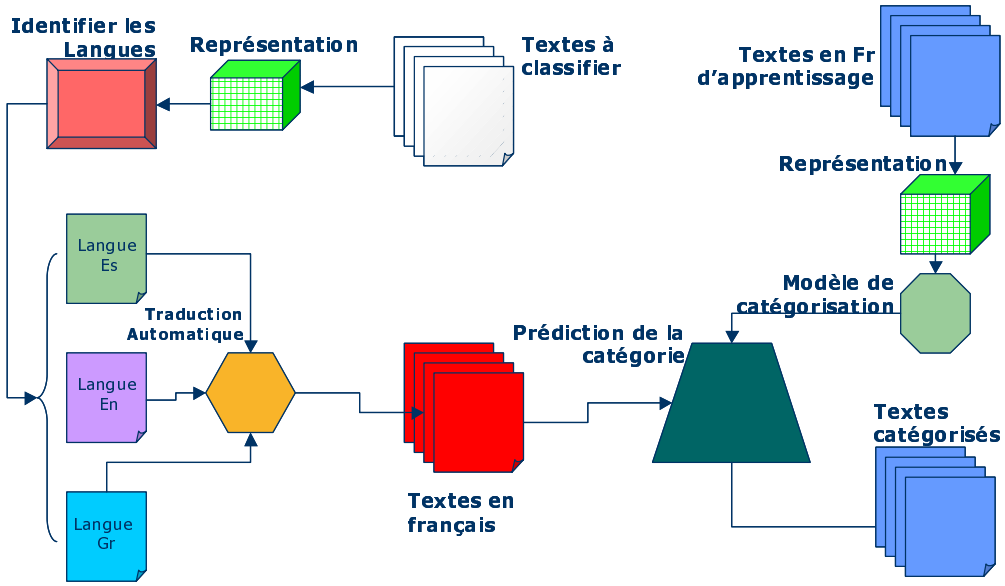

Figure 2. Catégorisation de textes multilingues : choix d'une langue d'apprentissage et utilisation d'un seul modèle, dans cette langue

Pour répondre à ces nouvelles contraintes, nous aménageons le processus de catégorisation exposé dans la figure 1. La phase d'apprentissage n'est pas modifiée, en revanche la phase de classement comporte deux étapes supplémentaires (figure 2) :

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 $(w_1, w_2, w_k, ..., w_{|\mathcal{T}|})$ dans $d_{j_{app}}$ afin de pouvoir appliquer le modèle prédictif et ainsi associer une catégorie c_i au texte à classer. Rappelons qu'en catégorisation de texte, chaque document d_j est transformé en un vecteur $\mathbf{d}_j = (w_{1j}, w_{2j}, w_{kj}, ..., w_{|\mathcal{T}|j})$, où $\mathcal{T}$ est l'ensemble de termes (descripteurs) qui apparaissent au moins une fois dans l'ensemble d'apprentissage. Le poids w_{kj} correspond à la contribution du termes t_k à la sémantique du texte d_j.

2.1. *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. Nous avons utilisé les 3-grammes qui sont des séquences de 3 caractères consécutifs extraits du texte à classer (Jalam et Chauchat, 2002). 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% (Jalam et Teytaud, 2001). Et pour des textes plus longs, la reconnaissance est parfaite.

2.2. *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.

3. Application sur les corpus CLEF

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'*ATS*, 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 monolingues, bilingues et multilingues. Les fichiers de ces corpus sont au format SGML. Par exemple, le fichier `le-monde_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 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 classes. Ces exemples serviront, dans le processus d'apprentissage, à apprendre un classifieur qui sera ensuite appliqué sur les textes à classer. Une première approche peut consister à utiliser les termes-index, extraits de la balise « *subjects* », 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.

<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. Etonnement 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>

Tableau 1. Exemple extrait de la collection CLEF et qui montre un article du journal Le Monde publié le 4 juin 1994.

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 16 200 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 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 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 2.

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},

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**

Tableau 2. Description des catégories utilisées.

présenté dans (Clech *et al.*, 2003), en sélectionnant 100 mots avec pour seuls prétraitements l'uniformisation de la casse et la suppression des mots-outils (*Stop Words*). 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 d'être traduit par le terme « Kuwait ».

3.3. *Algorithmes d'apprentissage*

Dans notre application, nous utilisons deux méthodes renvoyant à des paradigmes d'apprentissage très différents :

– 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).

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 (Jalamet et Teytaud, 2001) : pour des textes de taille égale ou supérieure à 100 caractères le taux de reconnaissance de la langue est de 100%.

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

	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%

Tableau 3. Taux d'erreur, en validation croisée (V.C.), pour les articles de journaux dans leur langue originelle.

	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%

Tableau 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).

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 le tableau 2), chaque document n'appartient qu'à une classe, et une seule. Il en résulte que les rappel et précision micro-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 (tableau 3), ils ont servi de référence pour juger de la qualité de ceux obtenus dans un contexte multilingue.

3.5.1. *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*) sont présentés dans le tableau 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.

(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%

Tableau 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 ».

3.5.2. *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 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 1) et évalué l'erreur en validation croisée (tableau 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 (tableau 4) comparées à ceux obtenus dans la langue d'origine (tableau 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 5 regroupe les résultats obtenus pour le corpus LM et le tableau 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 5 et 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 5) et un saut de plus de 70 points pour le corpus SDA (Tableau 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 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. 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 le tableau 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.

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. 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.

(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%

Tableau 6. Précision et Rappel (micro et macro-moyen) pour le corpus allemand SDA représenté par 100 mots (tableau a) et pour 200 4-grammes (tableau b). Les deux tableaux 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 ».

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.

5. Conclusion

L'objet de cet article 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. Nous espérons ainsi que les particularités propres à chaque langue ne seront plus retenues par les modèles estimés.

Références

Afnor (1993). Information et documentation. Principes généraux pour l'indexation des documents. NF Z 47-102.

Aha D.W., Kibler D. et Albert M.K. (1991). Instance-based learning algorithms. *Machine Learning*, vol. (6) : 37-66.

Clech J., Rakotomalala R. et Jalam R. (2003). Séléction multivariée de termes. In *Société Française de Statistique* : 933-936.

Jalam R. et Chauchat J.-H. (2002). Pourquoi les n-grammes permettent de classer des textes ? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques. In *Actes des JADT 2002*, vol. (1) : 381-390.

Jalam R. et Teytaud O. (2001). Kernel based text categorization. In *Proceedings of IJCNN-01, 12th International Joint Conference on Neural Networks* : 1891-1896.

Mitchell T.M. (1997). *Machine Learning*. Computer Science. McGraw-Hill.

Quinlan J. (1993). *C4.5 : Programs for Machine Learning*. San Mateo, CA.

Savoy J. (2002). Report on clef-2002 experiements : Combining multiple sources of evidence. Results of the CLEF-2002, cross-language evaluation forum. University of Neuchatel.

Sebastiani F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, vol. (34/1) : 1-47.