Catégorisation automatique de pages web chinoises

Documents spécialisés *vs* grand public sur le tabagisme

Guiyao Ke, Pierre Zweigenbaum

LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
pz@limsi.fr

RÉSUMÉ. *La catégorisation (ou classification supervisée) de textes concerne généralement le thème traité ou le type de document. Nous nous intéressons ici à une dimension particulière, le public visé, en distinguant deux grandes catégories : textes destinés au grand public, et textes destinés à des spécialistes du domaine traité. Nous testons la catégorisation, selon cette opposition, de pages web en langue chinoise sur le thème du tabagisme. Dans ce contexte, nous obtenons les conclusions suivantes : une segmentation des textes chinois en mots plutôt qu'en sinogrammes n'améliore pas la catégorisation mais facilite son interprétation ; des attributs supplémentaires relevés à la lecture humaine du corpus n'améliorent pas la catégorisation ; un arbre de décision ou un SVM sont plus performants sur un corpus de test proche du corpus d'entraînement ($F1 = 98,5$ %) que Naïve Bayes ou Kppv ; les Kppv ou un arbre de décision sont plus performants sur un corpus de test plus éloigné ($F1 = 93,4$ %) ; une sélection automatique d'un sous-ensemble des attributs est performante sauf pour les SVM.*

ABSTRACT. *Text categorization (or supervised classification) generally addresses the topic or the type of a text. We tackle here a different dimension, the intended audience, contrasting two broad categories: texts intended for the general public, or texts intended for specialists. We test the categorization, according to this contrast, of Chinese Web pages about smoking. In this context, we obtain the following conclusions: segmenting Chinese texts into words instead of into sinograms does not improve categorization but facilitates the interpretation of results; additional attributes elicited after human reading of the corpus do not improve categorization; decision trees or SVM outperform ($F1 = 98.5\%$) Naïve Bayes or kNN on a test corpus close to the training corpus; kNN or decision trees outperform ($F1 = 93.4\%$) Naïve Bayes or SVM on a more distant test corpus; automatic feature selection improves results except for SVM.*

MOTS-CLÉS : *Catégorisation de textes ; chinois ; public visé.*

KEYWORDS: *Text categorization; Chinese; Intended audience.*

1. Introduction

La recherche d'information sur l'internet ramène des pages web de types variés ; plusieurs moteurs de recherche ont proposé de regrouper les pages ramenées par classes homogènes, généralement thématiques. Un courant de travaux s'est intéressé à la catégorisation [1] des genres de documents (Karlgren 2000), et plus particulièrement de ceux trouvés sur le web. Par exemple, (Rosso 2005) cherche à caractériser une typologie de genres pour améliorer la recherche d'information sur le web ; et (Santini 2007) s'intéresse à sept genres du web : blogs, boutiques en lignes, FAQ, pages d'accueil, listes (annuaires), pages personnelles, pages de recherche.

Le public visé par un document est l'une des composantes qui entrent en jeu dans la définition d'un genre (Sinclair & Ball 1996, Lee 2001). De nombreux travaux, dont ceux cités plus haut, le mentionnent comme attribut à prendre en compte pour la catégorisation de documents. Nous n'avons cependant pas trouvé de travaux qui proposent de déterminer automatiquement le public visé par un document à partir du contenu de ce document, par exemple de distinguer des pages web destinées au grand public de celles destinées à des spécialistes. C'est l'un des aspects abordés par le projet C-Mantic [2], qui cherche à mettre en œuvre des méthodes de sémantique textuelle pour caractériser certains constrastes dans des documents du web. Le projet s'est fixé un thème, il s'agit du tabagisme, et différents contrastes à explorer : opposition grand public / spécialistes, pro- / anti-tabac, etc. Il a prévu de travailler sur trois langues : le français, l'anglais et le chinois. Nous nous intéressons ici à l'opposition grand public / spécialistes dans des textes chinois. La catégorisation de pages web chinoises a elle aussi fait l'objet de travaux depuis plusieurs années (par exemple, (He *et al.* 2000)), mais de nouveau, nous n'avons pas trouvé d'étude de la catégorisation du public visé.

Pour préparer un point de comparaison pour les méthodes ultérieures de sémantique textuelle, nous avons cherché à mettre en œuvre des méthodes de catégorisation automatique de textes (Sebastiani 2002) fondées sur une approche « sac de mots ». Les questions posées par ce travail sont les suivantes : une catégorisation du public visé par un texte peut-elle utiliser les méthodes classiques de la catégorisation en thèmes ; quels attributs employer ; quelles méthodes de catégorisation automatique sont les plus appropriées ? D'autre part, le chinois pose un problème spécifique de segmentation en mots, que nous présentons plus bas : une segmentation en mots est-elle nécessaire, ou une simple segmentation en sinogrammes (caractères chinois) unitaires suffit-elle ?

Dans le reste de cet article, nous commençons par présenter la préparation des corpus sur lesquels nous avons travaillé (section 2), puis les attributs retenus et les

1. Nous employons le terme « catégorisation » pour désigner la répartition d'objets dans un ensemble de classes prédéterminées (classification supervisée), et « catégoriseur » pour désigner un algorithme ou un programme qui réalise cette opération.

2. Ces travaux ont été partiellement financés par le projet C-Mantic (ANR-07-MDCO-002) coordonné par l'ERTIM, Inalco, et auquel collaborent le LINA, le LIMSI et l'Inserm UMR_S 872.

méthodes de catégorisation mises en œuvre (section 3). Nous commentons les résultats obtenus (section 4) et les discutons (section 5) puis concluons (section 6).

2. Préparation du corpus

2.1. *Constitution d'un corpus de textes chinois sur le tabagisme*

Un corpus sur le tabagisme a été collecté en partant de requêtes au moteur de recherche chinois Baidu (百度, `http://www.baidu.com/`), réputé pour mieux traiter les pages en langue chinoise que Google. Le terme de départ était 烟草, qui désigne le tabac. Ce terme a été affiné progressivement au vu des pages et des sites renvoyés, la sélection finale des pages étant faite manuellement. Le tableau 1 donne un extrait de la liste de mots clés employés. Cela a permis de repérer 20 sites particulièrement riches sur le tabagisme. Ces sites incluent des sites spécialisés dans le tabagisme ainsi que des sites généraux contenant des pages sur ce thème. Les pages HTML de ces 20 sites ont été téléchargées [3]. Un échantillon des fichiers résultants a été sélectionné aléatoirement et catégorisé manuellement en « spécialistes » ou « grand public ».

TAB. 1. *Mots clés chinois employés pour collecter des pages sur le tabagisme. *Le mot* 文化 *a le sens de « culture », « civilisation » (et pas celui d'« agriculture).*

mot clé	traduction
烟草文化	culture* du tabac
烟草医学	médecine et tabac
烟草科学	science du tabac
烟草健康	tabac et santé
烟草研究	recherche sur le tabac
烟草新闻	nouvelles sur le tabac
吸烟	fumer
戒烟	cesser de fumer

Nous avons pris pour cela les définitions suivantes de « spécialiste » et « grand public ». Nous considérons comme spécialistes les personnes qui connaissent bien le domaine du fait qu'elles y font de la recherche ou plus largement y exercent leur métier. Cela recouvre différentes disciplines (médecine, psychologie, pharmacie, sciences de la vie, écologie...) ou organismes (Ministère de la santé, organismes anti-tabac, administrations qui prennent en charge le contrôle, l'exportation et l'importation du tabac, fabricants de cigarettes, culture du tabac...). Ces personnes possèdent ainsi des

3. Le programme wget a été utilisé pour cela. Il a été appelé récursivement sur l'URL racine de chaque site, en le forçant à rester à l'intérieur du site. Même s'il reste difficile, du fait de l'existence de liens brisés ou de pannes de serveur ou de communication, de télécharger un site entier de façon réellement exhaustive, on obtient de cette manière une partie très importante des pages du site.

connaissances professionnelles sur le tabac. Par opposition aux spécialistes, le grand public comprend toutes les personnes qui ne possèdent pas de connaissances professionnelles sur le tabac.

Certaines pages sont cependant difficiles à classer. C'est le cas par exemple de pages de sites gouvernementaux donnant des informations administratives en relation avec le tabac (réglementations, lois), qui peuvent concerner aussi bien le grand public que des professionnels. Ou encore de chiffres de résultats du commerce du tabac, dont on n'est pas sûr qu'ils s'adressent uniquement aux professionnels. La distinction n'a pas toujours été simple à faire, et il se peut donc que notre choix ait été arbitraire dans quelques cas. Pour des raisons de temps, nous avons cependant conservé ces deux catégories. Idéalement, il faudrait créer des catégories intermédiaires ; ou encore, comme l'a suggéré l'un des relecteurs, avoir recours à plusieurs personnes pour effectuer la catégorisation (avec une part de spécialistes et une part de personnes du grand public), et considérer comme intermédiaires les pages sur lesquelles l'accord entre annotateurs est trop faible.

2.2. *Découpage en corpus d'entraînement et de test*

Cet échantillon a ensuite été découpé aléatoirement en corpus d'entraînement et corpus de test. Comme de coutume, le corpus d'entraînement servira de base d'exemples pour des systèmes de catégorisation automatique présentés plus bas ; le corpus de test servira à évaluer la performance de ces systèmes une fois entraînés. Le fait que les pages du corpus d'entraînement et de test proviennent des mêmes sites web rend le corpus de test homogène au corpus d'entraînement. Nous l'appellerons « corpus de test proche ». Pour tester sur un corpus plus éloigné la catégorisation qui sera apprise, nous avons collecté ensuite les pages de 16 sites supplémentaires. Nous appellerons le corpus correspondant « corpus de test éloigné ». Le tableau 2 donne les tailles respectives des corpus collectés. Chaque page des corpus d'entraînement et de test (proche et éloigné) a été catégorisée à la main en « spécialiste » ou « grand public ». La figure 1 montre une page « grand public » et une page « spécialiste ».

TAB. 2. *Constitution et taille du corpus collecté. Les tailles sont données en méga-octets et en nombre de caractères chinois (sinogrammes). La taille moyenne des textes est aussi indiquée.*

Corpus	URL	Mo	sinogrammes : total	sinogrammes : moyenne
corpus original (20 sites)	22646	115	64 297 265	2862
échantillon sélectionné	3178	20	11 232 248	3534
entraînement	2547	16	8 978 821	3525
test proche	631	4	2 295 505	3638
corpus de test éloigné	420	2	1 182 228	2815

(a) Page à destination du grand public

(b) Page à destination de spécialistes

FIG. 1. *Deux pages web chinoises sur le tabagisme.*

2.3. *Conversion en texte brut*

Les pages téléchargées étaient au format HTML. Le codage des caractères utilisé en Chine continentale est généralement GB2312. Lorsque c'était le cas, les documents ont été convertis en Unicode UTF-8 à l'aide du programme iconv. La conversion en texte brut a ensuite été réalisée en deux étapes : conversion en XML (plus précisément, XHTML, à l'aide du programme tidy du W3C) puis extraction des zones de texte par un programme XSLT. Enfin, chaque texte a été segmenté en phrases selon les points, points d'interrogation et d'exclamation chinois (donc en « pleine chasse » : 。？！).

2.4. *Faut-il segmenter en caractères ou en mots*

Le traitement du chinois pose le problème de la segmentation en mots [4]. En effet, les mots en chinois ne sont pas séparés par des espaces. Une phrase chinoise se compose d'une séquence de caractères, avec éventuellement des signes de ponctuation. Un mot se compose d'un ou plusieurs caractères, et aucune marque graphique n'indique où sont les frontières. Ce problème a motivé de nombreux travaux sur la segmentation en mots de textes chinois (Jin & Chen 2008), et des segmenteurs automatiques sont disponibles au téléchargement. Nous avons utilisé l'un de ces segmenteurs sur les textes de notre corpus. Le tableau 3 montre un exemple de passage d'une phrase et la segmentation en mots qui est produite par ce segmenteur.

TAB. 3. *Exemple de segmentation en mots (réalisée automatiquement).*

Original	中华人民共和国烟草专卖法
Segmentation en mots	中华人民共和国 / 烟草 / 专卖 / 法
Glose	*République Populaire de Chine / tabac / monopole / loi*
Traduction	*loi sur le monopole du tabac en RPC*

Une question qui se pose est de savoir si la segmentation en mots est utile pour notre objectif de catégorisation. De la même façon que des connaissances linguistiques n'apportent pas toujours une contribution positive en recherche d'information, il n'est pas certain qu'un texte segmenté en mots soit plus facile à catégoriser qu'un texte simplement segmenté en caractères. (Foo & Li 2004) ont observé un effet positif de la segmentation en mots en recherche d'information, mais citent de nombreuses autres expériences dont les résultats sont divergents. (He *et al.* 2000) effectuent une segmentation en mots avant la catégorisation de pages web chinoises, mais ne testent pas son apport à cette tâche.

4. Ce problème est commun à d'autres langues comme par exemple le japonais, le vietnamien, le thaï, etc. Par exemple, (Nguyen *et al.* 2006) prennent pour cadre la segmentation du chinois pour présenter leurs travaux sur celle du vietnamien.

Nous avons effectué pour notre part deux séries d'expériences, les unes avec segmentation en caractères, les autres avec segmentation en mots, afin de tester l'apport de la segmentation en mots.

3. Expériences de catégorisation de pages web en chinois

De nombreux algorithmes de catégorisation automatique (classification supervisée) existent (Sebastiani 2002), et de nombreuses implémentations de ces algorithmes sont disponibles au téléchargement. Plusieurs boîtes à outils d'apprentissage regroupent de telles implémentations, ce qui en fait des outils idéaux pour lancer des expériences systématiques. Nous avons fait le choix de la plate-forme Weka (Witten & Frank 2005) (`http://www.cs.waikato.ac.nz/ml/weka/`), dont l'interface graphique est pratique pour le test interactif et dont les fonctionnalités peuvent également être appelées en ligne de commande.

Une expérience de catégorisation suit le protocole habituel suivant :

1) Représentation de chaque texte du corpus d'entraînement et du test par un vecteur d'attributs.

2) Éventuellement, sélection d'un sous-ensemble d'attributs qui semblent suffire à véhiculer l'information nécessaire pour chaque texte.

3) Entraînement d'un catégoriseur d'un type donné sur la partie *entraînement* du corpus.

4) Application de ce catégoriseur entraîné sur la partie *test* du corpus.

Différents choix sont possibles à chaque étape ; le reste de cette section décrit ces choix.

3.1. *Attributs initiaux*

L'approche classique en recherche d'information ou en catégorisation de textes consiste à représenter un texte par un sac de mots. Nous suivons cette approche : chaque texte est décrit par les unités qui le composent. Ces unités sont ici soit les sinogrammes (caractères) soit les mots obtenus après segmentation. Techniquement, nous insérons des espaces entre les sinogrammes ou les mots. Cela nous permet d'utiliser ensuite la fonction de Weka qui convertit une chaîne de caractères en vecteur de mots (StringToWordVector) en gardant au maximum 3000 unités par texte. Chaque « mot » est accompagné d'une valeur indiquant s'il est présent ou pas dans le texte (nous n'avons pas cherché à prendre en compte le nombre d'occurrences de chaque unité dans les textes).

3.2. *Attributs supplémentaires*

Les unités élémentaires du texte sont-elles suffisantes à décrire le texte pour la tâche visée ? La lecture des textes fait apparaître certains indices qui semblent révélateurs d'un texte destiné aux spécialistes. Nous avons relevé les indices suivants :

– les marques d'énumérations ou d'alinéas : par exemple, « premièrement » peut être exprimé par "[1", "（一）", "1.", "1,", "1、", "(1)", "一，", "一：", "一、", "一.", "一要", "一是", "一方面", "首先" (le sinogramme "一" signifie « 1 ») ;

– des termes ou marques discursives typiques du discours scientifique : par exemple, "其次" (ensuite), "最后" (enfin), "总之" (en conclusion), "论文" (mémoire ou rapport), "(图"(« (figure »), "(表" (« (tableau »), "《" (guillemet ouvrant) ;

– certains symboles mathématiques ou physiques ;

– la taille du texte en nombre de sinogrammes ;

– la taille moyenne des phrases en sinogrammes ;

– le fait que le texte contienne un nombre de chiffres supérieur à un seuil fixé ;

– le fait que le texte contienne un nombre de mots anglais supérieur à un seuil fixé ;

– etc.

Au total, 32 nouveaux attributs ont été définis et calculés sur chaque texte. Leur introduction ou pas dans les catégoriseurs est un paramètre de nos expériences. Cet ensemble d'attributs pourrait être étendu davantage, comme dans (Goeuriot *et al.* 2009), où sont recensés de façon plus systématique, pour un corpus comparable français et japonais, 13 critères structurels (en particulier obtenus à partir du marquage HTML), 18 critères modaux (actes allocutifs et élocutifs) et 11 critères lexicaux (vocabulaire spécialisé, caractères numériques, symboles, etc.).

Notons que ces attributs supplémentaires ont été définis manuellement. Des méthodes existent pour déterminer automatiquement des attributs supplémentaires, par exemple par recherche de termes associés (Qian *et al.* 2007), ce qui pourrait fournir des pistes intéressantes.

3.3. *Sélection d'attributs*

Certains attributs présents sont souvent en tout ou partie redondants ; de plus, un trop grand nombre d'attributs peut rendre plus difficile ou plus long l'entraînement. La sélection d'un sous-ensemble d'attributs cherche à réduire ces défauts potentiels. Différentes méthodes sont disponibles pour cela. Nous avons utilisé la fonction CfsSubsetEval de Weka, qui sélectionne des attributs fortement corrélés à la classe mais peu corrélés entre eux, alliée à BestFirst, qui effectue une recherche en meilleur d'abord avec possibilité de retour arrière. Ici aussi, cette possibilité de sélectionner un sous-ensemble d'attributs ou pas est un paramètre de notre jeu d'expériences.

3.4. *Catégoriseurs*

Parmi les divers algorithmes de catégorisation, nous avons choisi quatre catégoriseurs appartenant à des familles différentes :

– Arbre de décision : J48

– Bayésien naïf (Naïve Bayes) : NaiveBayes (binomial)

– K plus proches voisins (Kppv) : IBk

– Machine à points de support (SVM) : SMO

Nous avons utilisé ces catégoriseurs avec leurs paramètres par défaut.

3.5. *Expériences réalisées*

Au final, nous avons donc testé :

– quatre jeux d'attributs, comme indiqué dans le tableau 4 ;

– deux types d'unités, sinogrammes ou mots (section 2.4) ;

– trois corpus : ceux appelés *entraînement*, *test proche* et *test éloigné* dans le tableau 2 ;

– les quatre catégoriseurs ci-dessus (section 3.4) ;

ce qui fait un total de 96 expériences. Pour chacune, nous avons obtenu à travers Weka la précision P, le rappel R et la F-mesure de la catégorisation « grand public » (ou de la catégorisation « spécialiste »). Nous en rappelons ci-dessous les définitions (corrects = ensemble des documents « grand public » du corpus ; ramenés = documents que le catégoriseur a considérés comme « grand public ») :

$$P = \frac{\text{corrects} \cap \text{ramenés}}{\text{ramenés}} \; ; \; R = \frac{\text{corrects} \cap \text{ramenés}}{\text{corrects}} \; ; \; F = \frac{2PR}{P+R}$$

TAB. 4. *Les quatre jeux d'attributs employés. Six séries de telles expériences ont été menées, selon les « unités brutes » obtenues par la segmentation : soit des sinogrammes, soit des mots, selon la segmentation appliquée au préalable ; et selon le corpus de test : développement, test proche ou test éloigné. Soit au total 24 expériences par catégoriseur.*

	s/s attributs supplémentaires	avec attributs supplémentaires
s/s sélection	unités brutes	unités brutes + attributs ajoutés
sélection	sélection(unités brutes)	sélection(u brutes + attributs ajoutés)

4. Résultats

Un script lançant les expériences décrites ci-dessus a été exécuté. Nous présentons leurs résultats ci-dessous.

Le tableau 5 montre les résultats obtenus pour une catégorisation par arbre de décision, en utilisant tous les attributs initiaux, et en faisant varier le corpus de test et la segmentation. Comme on pouvait s'y attendre, la catégorisation est moins bonne sur les corpus de test que sur le corpus d'entraînement, et moins bonne sur le corpus de test éloigné que sur le corpus de test proche. La F-mesure sur le corpus de test éloigné est la plus pertinente pour évaluer l'application d'un tel catégoriseur sur de nouvelles pages.

TAB. 5. *Arbre de décision, attributs complets : catégorisation « grand public » ou « spécialiste ». P= précision, R= rappel, F= F-mesure, PICC = proportion des instances catégorisées correctement. S = segmentation en sinogrammes, M = segmentation en mots. 16 séries d'expériences similaires ont été réalisées pour les 4 jeux d'attributs et les 4 catégoriseurs. Le gras indique le meilleur résultat entre sinogrammes et mots.*

corpus	seg.	grand public			spécialiste			PICC
		P	*R*	*F*	*P*	*R*	*F*	
entraînement	S	0,982	**0,999**	0,990	**0,998**	0,979	0,989	0,989
	M	**0,987**	**0,999**	**0,993**	**0,998**	**0,986**	**0,992**	**0,992**
test proche	S	**0,971**	**0,985**	**0,978**	**0,983**	**0,966**	**0,974**	**0,976**
	M	0,968	**0,985**	0,976	**0,983**	0,963	0,973	0,975
test éloigné	S	0,926	**0,910**	**0,918**	**0,901**	0,920	**0,910**	**0,914**
	M	**0,930**	0,900	0,915	0,893	**0,925**	0,909	0,912

4.1. *Segmentation en sinogrammes ou en mots*

On observe de plus sur le tableau 5 que la segmentation en mots (M) amène des résultats sensiblements équivalents à ceux obtenus avec une simple segmentation en sinogrammes (S), et ce quel que soit le corpus de test : la différence observée est toujours largement inférieure à 1 point. Les quinze autres séries d'expériences confirment cette tendance.

4.2. *Ajout d'attributs supplémentaires, sélection d'attributs*

Le tableau 6 indique les attributs conservés après application de la sélection à l'ensemble d'attributs initial (sans ajout des attributs supplémentaires indiqués à la section 3.2). On y trouve aussi bien des termes plutôt associés à des documents à destination de spécialistes (分类 classification, 本文 cet article) que des termes associés à des documents destinés au grand public (下午 après-midi). Même si la segmentation en mots n'apporte pas d'avantage particulier pour la catégorisation, son application

permet de travailler sur des unités plus facilement interprétables humainement que des sinogrammes simples, qui sont souvent très ambigus hors contexte.

TAB. 6. *Attributs conservés après sélection, sans ajout d'attributs supplémentaires. La traduction en français a été ajoutée à la main pour en faciliter la lecture.*

terme	traduction	terme	traduction
2008年	année 2008	效果	effet
1998	(année ?)	数据	donnée
3	(mois ?)	是	être, oui, d'accord, correct
下午	après-midi	本文	cet article
分布	distribution	机械	mécanique
分类	classification	来源	source
北京	Pékin	栏目	rubrique
医院	hôpital	百科	encyclopédie
品种	espèce	综合	synthèse
品质	qualité, moralité	质量	qualité
差异	différence	近日	récemment
应用	application	量	mesurer, quantité, capacité
性	nature, sexe, genre, performance...	:	(ponctuation)
情况	situation		

Le tableau 7 montre l'influence de l'ajout d'attributs supplémentaires et de l'application d'une méthode de sélection d'attributs dans le cas de l'usage d'arbre de décision avec une segmentation en mots. Dans ce cas, la sélection d'attributs, bien que n'améliorant pas les résultats sur le corpus d'entraînement, donne de meilleurs résultats sur les corpus de test (chiffres en gras). L'ajout d'attributs supplémentaires avant une éventuelle sélection donne des résultats variables (chiffres en italiques) selon le corpus et l'application ou non de la sélection. Sur le corpus de test proche, sans sélection, l'ajout d'attributs supplémentaires améliore un peu les résultats ; mais si l'on applique la sélection d'attributs, les attributs supplémentaires font baisser la F-mesure. En revanche, sur le corpus de test éloigné, les attributs supplémentaires diminuent les performances avant sélection, mais les augmentent si l'on effectue une sélection. Nous revenons sur ces résultats ci-dessous dans la comparaison des catégoriseurs.

4.3. *Comparaison des catégoriseurs*

Les figures 2 et 3 synthétisent les F-mesures de la catégorisation « spécialiste » (Fs) des quatre catégoriseurs pour les différents jeux d'attributs, toujours avec une segmentation en mots, respectivement pour le corpus de test proche et pour le corpus de test éloigné. On y confirme que *la sélection d'attributs améliore généralement les résultats*, que ce soit sur les attributs originaux (« complets ») ou avec ajout d'attributs supplémentaires (« complets + ajoutés »), que ce soit sur le corpus de test proche

TAB. 7. *Arbre de décision, segmentation en mots. Influence de l'ajout d'attributs supplémentaires et de la sélection d'attributs. Fs = F-mesure de la catégorisation en « spécialiste » ; Fg = F-mesure de la catégorisation en « grand public ». Le gras indique le meilleur résultat entre sélection et non sélection, l'italique le meilleur résultat entre l'ajout ou pas d'attributs supplémentaires.*

Corpus	Cat.	originaux		originaux+supplémentaires	
			sélection		sélection
entraînement	Fs	**0,992**	0,987	**0,991**	0,984
	Fg	**0,993**	0,989	**0,992**	0,986
test proche	Fs	0,973	***0,985***	*0,976*	**0,981**
	Fg	0,976	***0,987***	*0,979*	**0,984**
test éloigné	Fs	*0,909*	**0,919**	0,905	***0,934***
	Fg	*0,915*	**0,924**	0,909	***0,941***

ou éloigné, *sauf* dans les deux cas suivants. D'une part, le catégoriseur à SVM fonctionne mieux sur des attributs sans sélection sur le corpus proche (ou sur le corpus éloigné si l'on n'ajoute pas les attributs supplémentaires) ; d'autre part, sur le corpus de test éloigné, le catégoriseur bayésien naïf est insensible à la sélection ou à l'ajout d'attributs supplémentaires.

Par ailleurs, *l'ajout de nos attributs supplémentaires n'améliore en général pas les résultats*, sauf dans de rares cas comme celui de l'arbre de décision avec sélection d'attributs sur le corpus de test éloigné, ou encore celui de SVM sans sélection d'attributs sur le corpus de test éloigné.

Globalement, sur le corpus de test éloigné (figure 3), k plus proches voisins ou arbre de décision combinés à la sélection d'attributs et pour les kPPV à l'ajout de nos attributs supplémentaires obtiennent les meilleures F-mesures (0,934). Sur le corpus de test proche (figure 2), SVM (sans sélection d'attributs, Fs=0,986) ou arbre de décision (plutôt sans nos attributs supplémentaires, avec sélection d'attributs, Fs=0,985) sont les plus performants. *Les arbres de décision semblent donc plus stables* sur notre type de tâche et nos corpus.

Nous avons réalisé des tests de significativité (test de Student unilatéral) sur les résultats obtenus avec les attributs « originaux + supplémentaires » avec sélection (groupes de droite « Sél(Orig+Sup) » sur les figures 2 et 3). Les différences entre corpus de test proche et éloigné pour un même catégoriseur sont toutes très significatives ($p < 0,001$). Pour le corpus de test proche, Naïve Bayes < {kPPV, AD} ($p = 0,02$) et SVM est moins distinct de {kPPV, AD} ($p = 0,07$), par contre Naïve Bayes n'a pas de différence significative avec SVM (et kPPV $\sim$AD). Pour le corpus de test éloigné, SVM < {Naïve Bayes, kPPV, AD} ($p < 0,05$) et Naïve Bayes n'a pas de différence significative avec {kPPV, AD} (et de nouveau kPPV$\sim$AD).

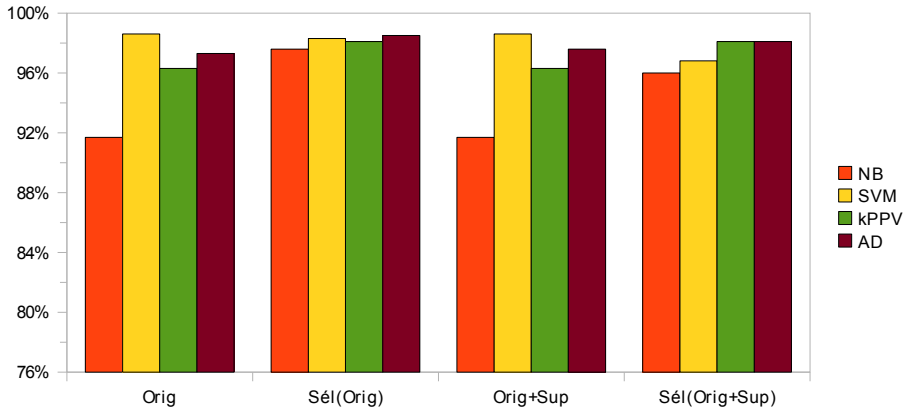

FIG. 2. *Comparaison des quatre catégoriseurs (NB = Naïve Bayes, SVM, kPPV = k plus proches voisins, AD = arbre de décision) pour les différents jeux d'attributs : corpus de test « **proche** », F-mesure de la catégorisation « spécialiste » (vs non spécialiste) (appelée Fs dans le tableau 7). Attributs complets = originaux ; ajoutés = supplémentaires.*

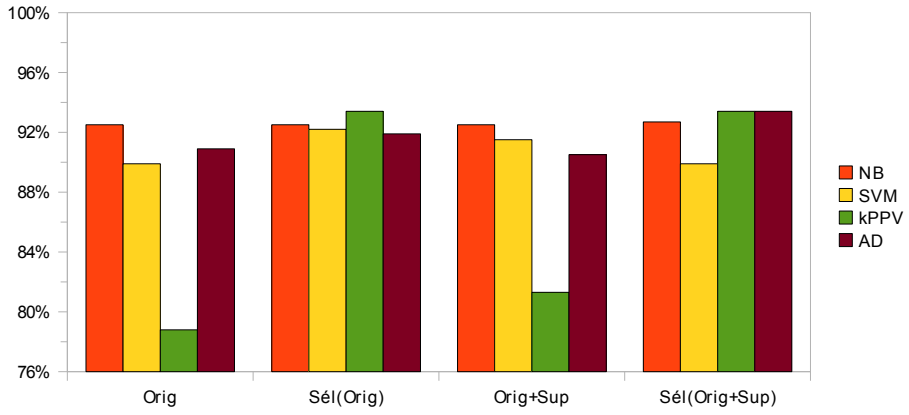

FIG. 3. *Comparaison des quatre catégoriseurs (NB = Naïve Bayes, SVM, kPPV = k plus proches voisins, AD = arbre de décision) pour les différents jeux d'attributs : corpus de test « **éloigné** », F-mesure de la catégorisation « spécialiste » (vs non spécialiste) (appelée Fs dans le tableau 7). Attributs complets = originaux ; ajoutés = supplémentaires.*

5. Discussion

Les différents catégoriseurs testés ont des résultats bons à excellents sur les deux corpus de test, avec des F-mesures en catégorisation « spécialiste » allant pour le corpus de test éloigné (avec attributs supplémentaires et sélection) de 0,899 (SVM) à 0,934 (kPPV, arbre de décision), et pour le corpus de test proche de 0,960 (Naïve Bayes) à 0,981 (kPPV, arbre de décision).

En considérant ces résultats positifs, on peut se demander si la tâche n'est pas trop facile, avec un corpus d'entraînement de plusieurs milliers de documents pour une catégorisation à deux classes. Il faut cependant s'attacher surtout aux résultats sur le corpus de test « éloigné », qui donne une meilleure idée des performances de catégorisation auxquelles on peut s'attendre sur de nouvelles pages web. Ceux-ci plafonnent à 0,934, ce que l'on pourrait vouloir encore améliorer.

D'après des sondages réalisés parmi les documents mal catégorisés, l'une des raisons qui peuvent expliquer les erreurs effectuées est que notre découpage en deux classes disjointes, « spécialiste » et « grand public », est une approximation trop abrupte de la réalité. Certaines pages web ont un caractère intermédiaire et s'adressent à un public sensibilisé au sujet sans être pour autant professionnel du domaine. (Sinclair & Ball 1996) distinguent par exemple les catégories suivantes de types de public : grand public, public sensibilisé, professionnels, spécialistes, étudiants [5]. De façon générale, un nombre de classes plus grand pourrait mieux refléter la réalité et aboutir à des classes plus homogènes, donc potentiellement mieux séparables. Il faut à l'inverse tenir compte du fait que la difficulté d'une catégorisation à N classes augmente généralement avec N ; et qu'il peut aussi exister un recouvrement entre les classes plus fines proposées.

Le corpus « éloigné » se compose de pages provenant de 16 sites web distincts de ceux du corpus initial. Ce nombre de sites est cependant encore trop faible pour que le test réalisé puisse être considéré comme général. Une nouvelle évaluation sur des pages web portant sur le tabagisme et tirées au hasard sur le web serait plus probante de ce point de vue.

6. Conclusion

Nous avons présenté ici des expériences de catégorisation selon le public visé de pages web chinoises sur le thème du tabagisme. Ces expériences obtiennent une catégorisation de bonne qualité sur le corpus de test préparé pour l'occasion (F-mesure jusqu'à 0,934 sur le corpus « éloigné »). Nous avons observé l'influence des

5. Notre traduction de : *members of the general public, informed lay people, professional people, specialists, students / trainees*.

différentes méthodes de catégorisation et du choix des attributs sur ces résultats, et noté qu'un corpus d'évaluation complémentaire plus divers serait nécessaire pour les confirmer.

Dans le cadre du projet C-Mantic, deux autres langues sont traitées : le français et l'anglais. Nous comptons tester les mêmes méthodes sur ces deux autres langues dès qu'un corpus suffisant sera disponible. Cela permettra de comparer les attributs et méthodes de catégorisation qui fonctionnent le mieux sur ces différentes langues, et d'étudier si certaines combinaisons sont stables à travers les langues. L'abord d'une autre opposition est également planifié : pages pro- et anti-tabac. Cette opposition, pour laquelle un corpus est en cours de constitution, sera sans doute à rapprocher de celle entre sites racistes et anti-racistes examinée lors du projet PRINCIP (Valette & Grabar 2004).

Remerciements

Nous remercions les relecteurs anonymes pour leurs commentaires constructifs. Les défauts restants sont notre fait.

7. Bibliographie

Foo S., Li H., « Chinese Word Segmentation and Its Effect on Information Retrieval », *Information Processing & Management*, vol. 40, p. 161-190, 2004.

Goeuriot L., Morin E., Daille B., « Reconnaissance du type de discours dans des corpus comparables spécialisés », *Actes CORIA 2009*, ARIA, 2009. Ce volume.

He J., Tan A.-H., Tan C.-L., « Machine Learning Methods for Chinese Web Page Categorization », *Actes ACL2000 International Workshop on Chinese Language Processing*, ACL, Hong Kong, 2000.

Jin G., Chen X., « The Fourth International Chinese Language Processing Bakeoff : Chinese Word Segmentation, Named Entity Recognition and Chinese POS Tagging », *Actes Sixth SIGHAN Workshop on Chinese Language Processing, IJCNLP 2008*, Asian Federation of Natural Language Processing, Hyderabad, India, 2008.

Karlgren J., Stylistic Experiments for Information Retrieval, PhD thesis, Stockholm University, 2000.

Lee D. Y., « Genres, Registers, Text Types, Domains, and Styles : Clarifying the Concepts and Navigating a Path Through the BNC Jungle », *Language, Learning, and Technology*, vol. 5, n° 3, p. 37-72, 2001.

Nguyen T. V., Tran H. K., Nguyen T. T., Nguyen H., « Word Segmentation for Vietnamese Text Categorization : An online corpus approach », *Actes IEEE RIFV 2006, Research, Innovation and Vision of the Future — The 4rd IEEE International Conference in Computer Science*, Hochiminh, Vietnam, 2006.

Qian T., Xiong H., Wang Y., Chen E., « On the strength of hyperclique patterns for text categorization », *Information Sciences*, vol. 177, n° 19, p. 4040-4058, 2007.

Rosso M. A., Using Genre to Improve Web Search, PhD thesis, University of North Carolina at Chapel Hill, 2005.

Santini M., « Automatic Genre Identification : Towards a Flexible Classification Scheme », *BCS IRSG Symposium : Future Directions in Information Access 2007 (FDIA 2007)*, Glasgow, Scotland, 2007. Held in conjunction with the European Summer School on IR (ESSIR 2007).

Sebastiani F., « Machine Learning in Automated Text Categorization », *ACM Computing Surveys*, vol. 34, n° 1, p. 1-47, 2002.

Sinclair J. M., Ball J., Preliminary recommendations on Text Typology, page WWW n° `http://www.ilc.cnr.it/EAGLES/pub/eagles/corpora/texttyp.ps.gz`, EAGLES (Expert Advisory Group on Language Engineering Standards), juin, 1996.

Valette M., Grabar N., « Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? l'exemple du projet PRINCIP », *in* G. Purnelle, C. Fairon, A. Dister (eds), *Le poids des mots, Actes 7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, UCL — Presses Universitaires de Louvain, Louvain-la-Neuve, Belgique, p. 1106-1116, 2004.

Witten I. H., Frank E., *Data Mining : Practical machine learning tools and techniques*, 2nd edition edn, Morgan Kaufmann, San Francisco, Ca, 2005.