Impact de la reconnaissance de l'écriture en-ligne sur une tâche de catégorisation

Sebastián Peña Saldarriaga* — **Emmanuel Morin*** — **Christian Viard-Gaudin****

* *LINA - UMR CNRS 6241 - Université de Nantes*
2, rue de la Houssinière - BP 92208, 44322 NANTES Cedex 3
{sebastian.pena-saldarriaga, emmanuel.morin}@univ-nantes.fr

** *IRCCyN - UMR CNRS 6597 - École Polytechnique de l'Université de Nantes*
rue Christian Pauc - BP 50609, 44306 NANTES Cedex 3
christian.viard-gaudin@univ-nantes.fr

RÉSUMÉ. *Cet article s'intéresse à la problématique de la catégorisation automatique de documents manuscrits en-ligne et plus particulièrement à l'impact de la reconnaissance de l'écriture dans un processus de catégorisation utilisant des méthodes d'apprentissage automatique. Nous comparons les performances obtenues avec des documents issus d'un système de reconnaissance de l'écriture en-ligne et leur version originale électronique. Les résultats montrent qu'aucune perte significative des performances n'est à signaler lorsque 78 % des termes d'indexation sont correctement reconnus dans les documents à catégoriser. Nous montrons également que lorsque plus de la moitié de ces termes sont mal reconnus, l'utilisation d'une liste de candidats mots permet d'améliorer le taux de classification.*

ABSTRACT. *This paper deals with the automated categorization of on-line handwritten documents. We experimentally show the effects of word recognition errors on a categorization engine using machine learning algorithms. We compared the performances of a categorization system over the texts obtained through on-line handwriting recognition and the same texts available as ground truth. Results show that no significant accuracy loss is expected when about 78% percent of indexation terms are correctly recognized. Results also show that using the top n recognition-candidates increases categorization rates of texts where more than 50% of indexation terms are incorrectly recognized.*

MOTS-CLÉS : *Catégorisation de textes, documents en-ligne, reconnaissance de l'écriture*

KEYWORDS: *Text categorization, noisy text, on-line handwriting recognition*

1. Introduction

L'émergence de nouveaux dispositifs de saisie que sont les stylos numériques couplés à des supports papier, permet de produire de documents en-ligne de façon très efficace. De véritables documents peuvent être produits grâce à ces dispositifs, ils peuvent consister aussi bien en des prises de notes, des cours, des copies d'examens, des rédactions d'articles, etc. Cela élargit les champs d'application de la saisie d'écriture en-ligne cantonnés souvent à des terminaux de petites tailles (PDA, smartphone) où seule la reconnaissance des caractères se justifiait.

Les documents en-ligne constituent une nouvelle source d'information en langue naturelle pour laquelle peu d'applications de RI existent. La catégorisation permettrait d'apporter diverses fonctionnalités aux documents en-ligne telles l'organisation automatique, le routage ou la recherche par thème. De manière générale, la catégorisation peut servir de base à une recherche ou extraction d'information efficace.

Le travail que nous présentons ici, a pour but d'étendre la catégorisation à des données initiales qui ne sont pas des documents textuels. Dans le cas d'un document manuscrit en-ligne, il s'agit de la trajectoire échantillonnée de l'instrument d'écriture disponible sous la forme d'une séquence de points $(x(t), y(t))$ dans l'espace, ordonnés dans le temps. Il est donc possible de retracer un caractère trait par trait comme l'illustre la figure 1.

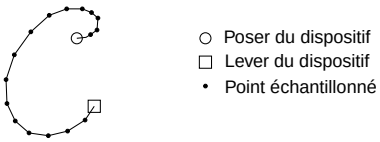


Figure 1. *Écriture en-ligne, appélée également encre numérique*

Une façon d'appréhender le problème de la catégorisation de ce type de documents est de se ramener à des données textuelles grâce à un système de reconnaissance de l'écriture. Dans ce travail nous explorons les effets produits par les erreurs de reconnaissance sur différents algorithmes de catégorisation que nous évaluons sur une base de documents reproduisant sous forme manuscrite les dépêches du corpus Reuters-21578 (Lewis, 1992) bien connu dans le domaine de la catégorisation de textes.

2. Travaux connexes

La collecte de grandes quantités de données manuscrites pour la catégorisation est une tâche difficile. À cause de cette difficulté, cette tâche n'a été explorée que très récemment (Vinciarelli, 2005, Koch, 2006, Peña Saldarriaga *et al.,* 2008, Milewski *et al.,* 2009). En revanche, des travaux sur la catégorisation de documents issus de la reconnaissance optique de caractères (OCR) existent depuis la dernière décennie (Ittner *et al.,* 1995, Junker *et al.,* 1998, Taghva *et al.,* 2000, Murata *et al.,* 2006).

Ces travaux utilisent différentes approches de catégorisation comme l'algorithme de Rocchio (Ittner *et al.,* 1995), des classifieurs bayésiens naïfs (Taghva *et al.,* 2000) ou des méthodes à base de n-grammes (Junker *et al.,* 1998). Cependant, la plupart s'appuient sur des données très spécifiques (Koch, 2006, Taghva *et al.,* 2000, Milewski *et al.,* 2009) ou des bases non standardisées (Ittner *et al.,* 1995, Junker *et al.,* 1998).

Les travaux les plus récents (Vinciarelli, 2005, Murata *et al.,* 2006, Peña Saldarriaga *et al.,* 2008) utilisent comme support un corpus bien connu dans le domaine de la catégorisation : le corpus Reuters-21578 (Lewis, 1992). Dans l'un des nos travaux précédents (Peña Saldarriaga *et al.,* 2008), nous avons évalué l'impact de la reconnaissance de l'écriture lorsque la catégorisation est effectué avec des modèles entraînés sur des documents électroniques. Ces expériences sont proches de celles décrites par (Vinciarelli, 2005).

La contribution présentée dans cet article se différencie des travaux antérieurs sur deux aspects. D'une part, ceci est à notre connaissance la première fois que la catégorisation est appliquée à des documents issus du domaine en-ligne en utilisant des documents manuscrits aussi bien pour l'entraînement que pour l'évaluation. Il faut également noter que les travaux en relation avec la RI dans le domaine se contentent souvent de rechercher des séquences de caractères dans l'encre numérique (Lopresti *et al.,* 1994, Jain *et al.,* 2003). D'autre part, nous utilisons la plus grande base de documents en-ligne existant à notre connaissance, elle est basée sur un corpus parfaitement standardisé. De plus nous donnons une analyse minutieuse des performances du système selon différentes perspectives applicatives.

3. Catégorisation automatique de textes

La catégorisation automatique de textes (CAT) est l'affectation d'une ou plusieurs étiquettes à un document en fonction de son contenu textuel. Un algorithme de catégorisation est un modèle mathématique dont l'objectif est de détecter le ou les thèmes abordés dans un texte. Plus formellement, la CAT peut être définie par une un fonction f telle que :

$$f : (d_i, c_j) \rightarrow \{vrai, faux\}, \qquad \forall (d_i, c_j) \in D \times C \qquad [1]$$

Avec d_i un document appartenant au domaine D et c_j une catégorie de l'ensemble $C = \{c_1, c_2, ..., c_{|C|}\}$.

L'apprentissage automatique permet d'approcher f par induction à partir d'un jeu d'entraînement (Sebastiani, 2002), c'est-à-dire, en utilisant un ensemble de textes dont la catégorie est connue au préalable.

Nous avons retenu deux méthodes de catégorisation parmi les nombreux algorithmes existants : la méthode des k-Plus Proches Voisins (kPPV) et les Séparateurs à Vaste Marge (SVM). Ces deux méthodes ont été choisies parce qu'elles figurent

parmi les approches les plus performantes développées durant la décennie (Yang *et al.,* 1999, Joachims, 2002, Debole *et al.,* 2005). De plus, elles nous permettent d'évaluer les effets de la reconnaissance sur deux méthodes de nature très différente.

Ces deux méthodes sont basées sur une représentation vectorielle des données (Salton *et al.,* 1975). Cela veut dire que la donnée de base, des textes bruts en langue naturelle, doit subir un certain nombre de transformations afin de se conformer au formalisme vectoriel. Dans ce formalisme, chaque dimension de l'espace vectoriel correspond à une entité représentative du sens, appelée communément terme, préalablement extraite du jeu d'apprentissage (*cf.* figure 2). La sélection des termes de l'espace vectoriel a été effetuée en utilisant l'algorithme de (Forman, 2004) sur les scores donnés par le test du χ^2 (Yang *et al.,* 1997) pour chacun des termes et chacune des catégories.

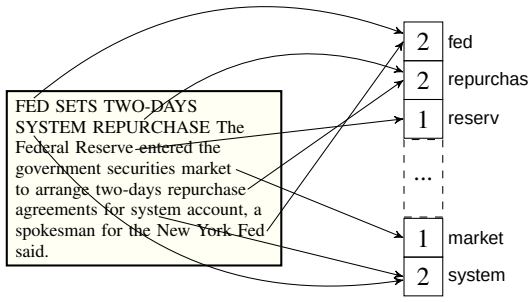


Figure 2. *Représentation vectorielle d'un texte*

L'ensemble des pré-traitements effectués pour transformer un texte brut en une liste de *termes*, est donné par la figure 3. Chacune de ces étapes est décrite ci-dessous.

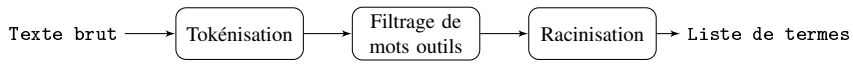


Figure 3. *Pré-traitements linguistiques*

La **tokénisation** permet de segmenter un texte en occurrences de formes (*tokens*). La liste de tokens obtenue après l'étape de tokénisation est filtrée en utilisant une liste de **mots outils** prédéfinie, celle-ci est constituée en particulier de prépositions, conjonctions, déterminants et auxiliaires. Enfin, la **racinisation** est utilisée afin de réduire les variations morphologiques d'un mot. Elle consiste à supprimer tous les affixes d'un mot, même si l'algorithme que nous utilisons n'effectue que la désuffixation des mots (Porter, 1980).

Vient naturellement ensuite une étape de pondération qui permet de déterminer de manière quantitative la représentativité de chacun des termes de la liste obtenue après l'étape de pré-traitements. La mesure $tf \times idf$ normalisée (Spärck Jones, 1979)

permet d'évaluer l'importance d'un terme en prenant en compte sa fréquence locale, c'est-à-dire relative à un document (*term frequency*) et sa fréquence globale, relative à un corpus (*inverse document frequency*). Le poids d'un terme i dans un vecteur, représentant un document, est donné par la formule suivante :

$$w_i = \frac{f_i \times \log \frac{N}{n_i}}{\sqrt{\sum_{j=1}^{M} \left(f_j \times \log \frac{N}{n_j} \right)^2}}, i = 1, \ldots, M \quad [2]$$

Avec f_i la fréquence du terme i dans un document, N le nombre de documents dans le corpus, M le nombre total de termes et n_i la fréquence du terme i dans le corpus.

3.1. *Les k-Plus Proches Voisins*

L'algorithme des kPPV est une méthode très connue dans le domaine de la catégorisation automatique. La catégorisation d'un document d, s'opère en comparant le vecteur du document à l'ensemble des vecteurs du jeu d'entraînement. Les k plus proches vecteurs sont sélectionnés et une probabilité d'appartenance à une catégorie c, pondérée par le cosinus des k documents les plus semblables, est calculée.

$$p(c|d) = \frac{\sum_{i=1}^{k} \cos(d, x_i) \times I(x_i, c)}{\sum_{i=1}^{k} \cos(d, x_i)} \quad [3]$$

avec

$$I(x_i, c) = \begin{cases} 1 & si\ categorie(x_i) = c \\ 0 & sinon \end{cases} \quad [4]$$

3.2. *Séparateurs à Vaste Marge*

Les Séparateurs à Vaste Marge (en Anglais *Support Vector Machines*) ont été introduits par Vapnik (1995). Ils sont devenus une méthode phare pour la classification supervisée et les travaux de Joachims, (2002) montrent qu'ils sont de par leur nature adaptés pour la CAT.

Les vecteurs des documents d'entraînement sont projetés dans un espace à grande dimension où il est possible de définir par apprentissage une surface de séparation entre des exemples positifs et négatifs appelée hyperplan. L'hyperplan optimal est

choisi de façon à minimiser les erreurs de catégorisation et à maximiser la marge de séparation entre les exemples.

La catégorisation utilisant l'approche par SVM a été réalisée grâce au package SVM^{light} V6.0 developpée par Joachims (2002)[1].

3.3. *Évaluation des méthodes de catégorisation*

Puisque nous disposons de plusieurs méthodes de catégorisation, nous devons mesurer la qualité des réponses données par le catégoriseur. Pour cela, nous disposons de deux mesures classiques : la précision (P) et le rappel (R).

Pour une catégorie c, la précision évalue la qualité du classifieur à ne pas introduire de documents d'une autre catégorie dans c. Il s'agit du nombre de documents bien classés sur le nombre de documents classés dans c.

$$P(c) = \frac{Documents\ bien\ classés\ dans\ c}{Documents\ classés\ dans\ c} \quad [5]$$

Le rappel, quant à lui, évalue le degré de complétude, c'est-à-dire le nombre de documents bien classés sur le nombre total de documents de la classe c.

$$R(c) = \frac{Documents\ bien\ classés\ dans\ c}{Documents\ de\ c} \quad [6]$$

Ces deux mesures prises l'une sans l'autre ne permettent d'évaluer qu'une facette du système de catégorisation : la qualité ou la quantité. Les courbes de précision *vs* rappel (Baeza-Yates *et al.,* 1999, chap. 3) permettent de mieux comprendre le comportement du classifieur, et de visualiser l'évolution de la précision en fonction du rappel pour les 11 niveaux standard $[0, 0, 1, \ldots, 1, 0]$. Comme le système est évalué sur un ensemble de catégories, nous utilisons les deux méthodes classiques pour moyenner la précision et le rappel : micro-moyenne et macro-moyenne (Sebastiani, 2002).

Les courbes de précision *vs* rappel sont utiles pour évaluer un système, lorsqu'il s'agit de retrouver un ensemble de documents étant donnée une catégorie (catégorisation centrée-catégorie). *A contrario*, lorsque le but est de retrouver un ensemble de catégories étant donné un document (catégorisation centrée-document), la mesure la plus appropriée est le taux de classification, il s'agit du nombre de documents bien classés sur le nombre de documents du corpus. Lorsque chaque document appartient à une catégorie et à une seule, le taux de classification correspond à la micro-moyenne de la précision ou le rappel (Beney, 2008).

1. L'application est librement téléchargeable à l'adresse suivante : `http ://svmlight.joachims.org`

4. Reconnaissance de l'écriture en-ligne

L'objectif de la reconnaissance de l'écriture en-ligne est de déterminer la suite de caractères la plus vraisemblable étant donné un signal correspondant au tracé manuscrit et les informations fournies par un ensemble de connaissances *a priori* sur la langue (*cf.* figure 4). L'objet de ce travail n'étant pas directement la reconnaissance de l'écriture, nous nous contenterons de décrire ci-dessous le moteur de reconnaissance en tant qu'outil plutôt qu'en tant que système.

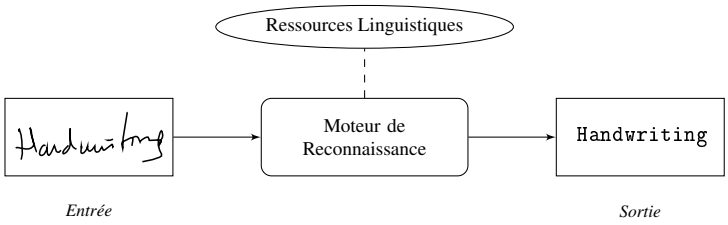


Figure 4. *Reconnaissance de l'écriture*

4.1. *Moteur de reconnaissance*

Le moteur de reconnaissance utilisé dans nos expériences est celui de MyScript Builder[2]. Il s'agit d'un ensemble de bibliothèques tournées vers la reconnaissance de l'écriture en-ligne.

MyScript Builder SDK est un outil stable, paramétrable et documenté. Il permet d'associer différentes ressources linguistiques afin de guider et d'optimiser la reconnaissance. Il est possible de définir des ressources spécifiques, soit sous forme de lexiques ou encore d'automates lexicaux, ou bien d'utiliser les deux ressources standard livrées avec le SDK :

– **lk-text** est une ressource contrainte par un modèle statistique du langage au niveau mot et un lexique standard. Le premier permet de favoriser la reconnaissance des séquences de mots les plus probables. Ainsi, *'je tue'* sera prioritaire par rapport à *'je tu'*. Cette ressource permet également de reconnaître des éléments hors-lexique comme les dates, les nombres, les codes postaux, etc.

– **lk-free** apporte peu de contraintes sur ce que l'on veut reconnaître. Il n'y pas de lexique mais seulement un modèle de langage au niveau caractère. Cette ressource permet principalement de reconnaître la séquence de caractères la plus vraisemblable. Ainsi, '$MATIN$' sera prioritaire par rapport à '$MAT1N$'.

2. `http ://www.visionobjects.com/products/software-development-kits/myscript-builder/`

Il faut noter que les deux ressources décrites ci-dessus sont des ressources livrées avec un système commercial, il nous est impossible de dire combien de mots sont présents dans le lexique, ni sur combien de documents ou mots ont été entraînés les modèles de langage. Il s'agit d'un système de reconnaissance totalement générique qui n'intègre aucune connaissance *a priori* sur le corpus de catégorisation.

4.2. *Évaluation de la reconnaissance*

Les mesures d'évaluation que nous décrivons ci-dessous vont nous permettre de vérifier le comportement du reconnaisseur lorsque telle ou telle ressource est utilisée.

Le *bruit* induit par la reconnaissance (insertion, suppression et remplacement de mots) est souvent mesuré au niveau des mots. Le taux d'erreur au niveau mot (Word Error Rate, WER) correspond au pourcentage de mots mal reconnus sur la totalité de mots à reconnaître pour une séquence donnée :

$$WER = 1 - \frac{\sum_{i=1}^{N} \min\left(wf(i), wf'(i)\right)}{\sum_{k=1}^{N} wf(k)} \quad [7]$$

Avec $wf(i)$ and $wf'(i)$ les fréquences du mot i dans le texte d'origine et le texte reconnu respectivement, et N le nombre de mots à reconnaître.

Une autre façon de mesurer le bruit, est de travailler au niveau terme. Le taux d'erreur au niveau terme (Term Error Rate, TER) est plus adapté à la catégorisation car il tient compte des pré-traitements qui normalisent les textes. Puisque '*rêvas*' et '*rêves*' ont la même racine, reconnaître l'un à la place de l'autre ne modifie pas la liste de termes reconnus. Reconnaître '*pour*' à la place de '*par*' ne la modifie pas non plus, car quelque soit le mot reconnu, il sera filtré puisque c'est un mot outil.

Le TER est calculé grâce à la formule suivante (Vinciarelli, 2005) :

$$TER = 1 - \frac{\sum_{i=1}^{N} \min\left(tf(i), tf'(i)\right)}{\sum_{k=1}^{N} tf(k)} \quad [8]$$

Avec $tf(i)$ and $tf'(i)$ les fréquences du terme i dans le texte d'origine et le texte reconnu respectivement, et N le nombre de termes de référence.

5. Corpus manuscrit

Le corpus utilisé dans nos expériences est un sous-ensemble du corpus Reuters-21578 (Lewis, 1992) reproduit sous forme manuscrite. Le corpus Reuters-21578 est l'un des plus utilisés dans la littérature scientifique pour l'évaluation de méthodes de catégorisation (Debole *et al.,* 2005). Ses documents sont distribuées en 135 catégories,

dont seulement 90 sont représentées dans l'ensemble d'entraînement et de test. Les 10 catégories ayant le plus d'effectifs comptent pour 90 % des documents du corpus.

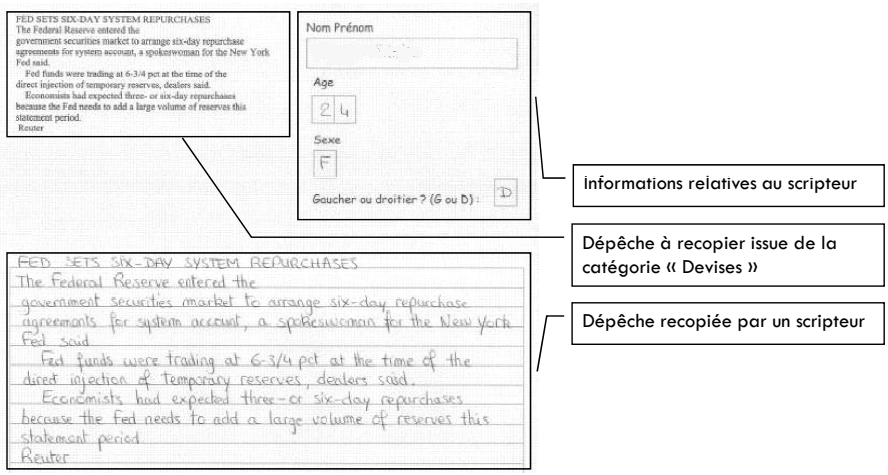


Figure 5. *Formulaire pour la collecte de données*

La collecte a été effectuée à l'aide de formulaires (*cf.* figure 5) et de stylos numériques. Nous avons mobilisé plus de 1 500 scripteurs pendant une période de 4 mois.

Pour cette collecte, nous avons dû choisir un nombre réduit de documents. Comme 90 % des documents appartiennent aux 10 premières catégories, nous avons choisi au hasard 2 000 documents d'entraînement et 500 documents de test parmi les documents de celles-ci. Les documents choisis n'appartiennent qu'à une seule catégorie et comportent au maximum 120 mots pour faciliter la tâche de saisie.

Une fois la collecte terminée, nous avons dû trier, anonymiser et associer le texte original à chacun des formulaires. À la fin du tri des documents, seuls 2 029 documents étaient exploitables. La distribution de ces documents par catégorie selon les ensembles d'entraînement et de test et donnée par le tableau 1.

Les documents du corpus abordent divers sujets comme les fusions-acquistions d'entreprises, les marchés de matières premières agricoles (céréales, sucre, etc.), le cours du pétrole et ses dérivés, les marchés de changes et du taux d'intérêt, etc.

6. Expériences

Afin de valider les méthodes de catégorisation sur le corpus que nous venons de présenter, nous avons mené plusieurs expériences. En premier lieu, nous avons effectué la reconnaissance des documents manuscrits (*cf.* section 6.1). Ensuite, nous avons catégorisé les documents tels qu'ils étaient donnés par le moteur de reconnaissance (*cf.*

Catégorie	Entraînement	Test
Dividendes (*earn*)	642	108
Fusion-Acquisition (*acq*)	349	72
Céréales (*grain*)	125	51
Devises (*money-fx*)	177	49
Pétrole (*crude*)	81	40
Taux d'intérêt (*interest*)	76	30
Import/Export (*trade*)	59	21
Transport Maritime (*ship*)	54	13
Sucre (*sugar*)	32	10
Café (*coffee*)	30	10
Total	1 625	404

Tableau 1. *Documents par catégorie*

section 6.2) ou alors en exploitant des informations supplémentaires en provenance du reconnaisseur, à savoir la liste de candidats mots à la reconnaissance (*cf.* section 6.3).

6.1. *Reconnaissance des documents*

Le moteur de reconnaissance de MyScript Builder a été utilisé pour effectuer la reconnaissance des documents en utilisant les deux ressources disponibles *lk-free* et *lk-text*. Le tableau 2 montre les performances de la reconnaissance en fonction de la ressource et du jeu de documents.

Ressource	WER	TER
lk-free	52,47 %	55,75 %
lk-text	22,30 %	23,01 %

(a) Ensemble d'entraînement

Resource	WER	TER
lk-free	52,48 %	55,85 %
lk-text	22,08 %	21,90 %

(b) Ensemble de test

Tableau 2. *Taux d'erreur à la reconnaissance*

Les textes reconnus avec la ressource *lk-free* sont fortement degradés, en effet plus d'un mot sur deux est perdu en moyenne. Le nombre de termes perdu est encore plus important. Les textes issus de la reconnaissance avec *lk-text* sont beaucoup moins bruités : 77 % des mots et autant de termes présents dans les textes sont correctement reconnus. Cette différence entre les deux ressources s'explique par la stratégie de reconnaissance de *lk-text* qui consiste à cherche la séquence de mots la plus vraisemblable.

Si nous comparons les performances sur le jeu d'entraînement et de test, nous remarquons également que MyScript Builder se comporte de manière cohérente sur les deux ensembles. De plus les taux de reconnaissance de *lk-text* sont bons si on consi-

dère que la couverture lexicale de la ressource est de 70,64 % et le WER théorique de 17,88 % (TER ≈19,45 %) par rapport au corpus électronique.

6.2. *Catégorisation des documents*

Notre première expérience a consisté à confronter notre système à d'autres systèmes de l'état de l'art. Pour cela nous avons utilisé le sous-ensemble $R(90)$ (Debole *et al.,* 2005) du corpus Reuters-21578. Cela permet de valider notre système de catégorisation en comparant nos résultats avec ceux disponibles pour le même jeu de données avec les mêmes méthodes de catégorisation et des paramètres expérimentaux aussi proches que possible (Yang *et al.,* 1999, Joachims, 2002).

Afin de nous conformer avec les résultats de ces travaux, nous présentons la micro-moyenne de la F_1-mesure (Manning *et al.,* 1999, p. 269) dans le tableau 3.

	(Yang *et al.,* 1999)	(Joachims, 2002)	*Cet article*
$\text{kNN}_{k=30}$	0,857	0,826	0,840
SVM	0,859	0,875	0,889

Tableau 3. *Résultats pour le corpus original - Micro-moyenne de la* F_1

Les performances rapportés ci-dessus montrent que nos résultats sont comparables aux résultats de l'état de l'art pour les mêmes algorithmes de classification. La prochaine étape est d'appliquer notre système de catégorisation aux données issues du système de reconnaissance.

Eu égard à la taille réduite du corpus ayant servi pour la collecte de données manuscrites, les différents paramètres de classifieurs ont été ajustés. Ceci a été fait par validation sur un sous-ensemble du jeu d'entraînement électronique de 1 625 documents. Les SVM sont utilisés avec 1 000 termes et les kPPV avec 15 plus proches voisins et 300 termes. Ces paramètres sont conservés pour la catégorisation des documents manuscrits. Il faut noter que ces paramètres peuvent ne pas être optimaux pour les données manuscrites, cependant notre but était de tester notre système sur la même tâche de catégorisation et non pas d'optimiser minutieusement les paramètres.

Les résultats de la catégorisation selon le type de documents et de classifieur sont indiqués dans le tableau 4. Nous avons utilisé les documents électroniques pour calculer des performances de référence. Lorsque les documents manuscrits sont utilisés, ils le sont aussi bien pour la sélection de termes et l'apprentissage que pour le test.

Ces résultats ont été obtenus en cherchant pour un document donné une liste de catégories potentielles et en assignant la catégorie ayant le score le plus important.

Le premier constat que nous faisons est une baisse d'environ 10 % du taux de classification avec *lk-free* par rapport à la référence, et ce quel que soit le type de classifieur. Cette ressource ne semble pas adaptée à une reconnaissance orientée catégorisation. Cependant, les performances ne sont pas aussi mauvaises, si nous considé-

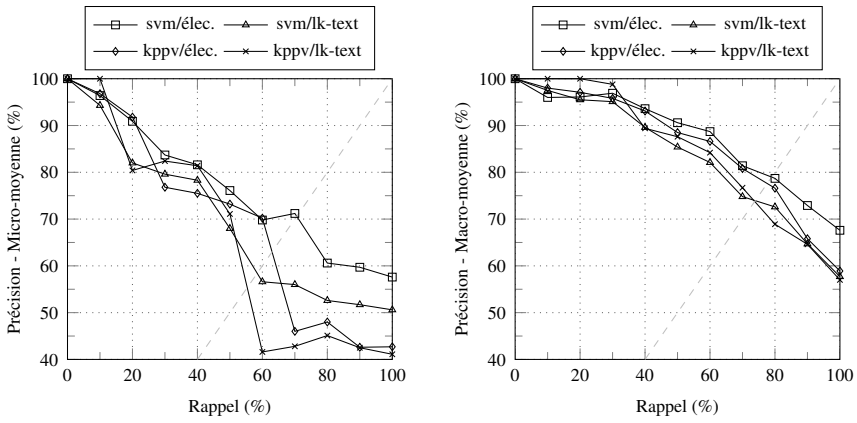
Documents	kPPV	SVM
électroniques	90,10 %	93,56 %
lk-text	89,36 %	91,83 %
lk-free	79,46 %	84,65 %

Tableau 4. *Taux de classification pour les différents documents*

rons que plus de la moitié de l'information textuelle contenue dans les documents est perdue à cause de la reconnaissance.

De son côté, la ressource *lk-text* obtient des résultats convenables avec les deux méthodes. Une baisse de 1 % est enregistrée avec kPPV tandis qu'avec les SVM cette baisse est d'environ 2 %.

Une vue détaillée des performances du système, lorsque les documents de *lk-text* sont utilisés, est donnée par les courbes de précision vs rappel (*cf.* figure 6). Ces courbes sont le résultat d'une catégorisation centrée-catégorie. L'intérêt de calculer les courbes sur le rang des documents est double. D'une part cela évite d'avoir à définir un seuil (Yang, 2001) dépendant foncièrement de l'application finale (avoir plus ou moins de précision ou de rappel). D'autre part ces courbes constituent une borne supérieure des performances d'un système utilisant un seuil[3].



(a) Micro-moyenne (b) Macro-moyenne

Figure 6. *Précision vs Rappel*

La figure 6(a) montre que jusqu'à 50 % de rappel, les courbes restent proches les unes par rapport aux autres. Au-delà, une nette différence apparaît. Les courbes des kPPV chutent de façon importante, mais restent proches pour les taux de rappel supé-

3. Cela découle du procédé d'interpolation (voir (Baeza-Yates *et al.,* 1999, p. 78))

rieurs à 70 %. La courbe des SVM avec les documents électroniques dépasse de façon visible toutes les autres à partir de 60 % de rappel.

En macro-moyenne (*cf.* figure 6(b)), aucune différence significative entre les courbes n'est observée avant 80 % de rappel. Au-delà de ce taux, seuls les SVM avec les documents électroniques surpassent les autres séries de manière visible.

Bien que les SVM donnent des meilleurs résultats que les kPPV dans tous les cas de figure, la baisse des performances qu'ils enregistrent est plus importante. Ceci est dû au choix des paramétres optimaux. En effet, les approches basées sur kPPV doivent leurs performances à l'identification d'un espace vectoriel discriminant dans le jeu d'entraînement (Dasarathy, 1991). Puisque les 300 termes utilisés avec kPPV sont les 300 plus discriminants d'un point de vue statistique, ils ont plus de chances d'être correctement reconnus dans l'ensemble de test.

6.3. *Utilisation des n-best*

En plus de donner le mot le plus probable, le moteur de reconnaissance peut donner une liste ordonnée des meilleurs candidats mots à la reconnaissance, où à chaque candidat est associé une probabilité. Ainsi, il est possible de garder l'information correspondant à un terme de l'espace vectoriel qui ne serait pas arrivé en première position.

Cependant cela introduit du bruit supplémentaire dans les textes, plus la liste est grande plus le bruit est important. Afin de réduire l'impact de ce bruit, nous utilisons la probabilité d'un terme pour la pondération plutôt que sa fréquence dans le document.

La figure 7 montre l'évolution du taux de classification en fonction de la taille de la liste de n-best. Dans cette expérience nous avons utilisé les documents de *lk-free* et *lk-text*, et les catégoriseurs tels qu'ils ont été paramétrés précédemment.

Lorsque nous utilisons la liste des n-best avec *lk-text*, le taux de classification baisse avec l'augmentation de n. En revanche, lorsque les documents de *lk-free* sont utilisés, pour tout $n > 1$ le taux de classification est supérieur à celui obtenu en ne prenant que le premier candidat, et ce quel que soit le classifieur utilisé. L'augmentation moyenne est de 2,1 % avec un écart-type de 1,2. Cependant cette augmentation n'est pas régulière et ne semble pas être correlée avec n.

7. Conclusion et perspectives

L'augmentation de la production de documents manuscrits en-ligne, intégrant de l'écriture dynamique, nécessite le développement de nouveaux outils de gestion adaptés à la nature même des documents. La catégorisation peut permettre d'organiser les documents de façon à pouvoir effectuer, ultérieurement, une extraction ou une recherche d'information efficace.

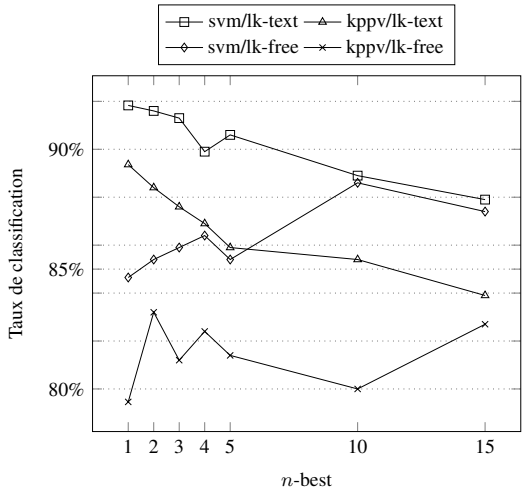


Figure 7. *Taux de classification en fonction de* n

Mais l'information textuelle contenue dans ces documents n'est accessible que grâce à un processus de reconnaissance. Ce processus induit des erreurs dans le texte résultant. Ce travail montre expérimentalement l'influence que peuvent avoir ces erreurs lorsque nous voulons catégoriser des textes en aval de la reconnaissance.

Deux algorithmes d'apprentissage automatique ont été évalués et comparés dans cette étude. L'évaluation des performances a été effectuée sur un sous-ensemble du corpus Reuters-21578 d'environ 2 000 documents manuscrits. La qualité du classifieur a été mesurée sur deux versions reconnues du corpus manuscrit et sa version électronique.

Lorsque nous comparons les performances obtenues avec les documents électroniques et les documents manuscrits, nous constatons qu'il n'y a pas de baisse significative des performances lorsque les documents reconnus avec *lk-text* sont utilisés : entre 1 % et 2 % pour un TER d'environ 56 % . Bien que peu significative, nous avons tenter de pallier cette baisse en utilisant la liste des n-best mots donnée par le moteur de reconnaissance, mais aucun effet positif sur le taux de catégorisation n'a été observé.

En revanche, l'utilisation de la liste des n-best mots s'est révélée bénéfique pour les documents issus de la reconnaissance avec *lk-free*. En effet, aussi bien pour les kPPV que pour les SVM, une augmentation du taux de classification allant de 1 % à 4 % a été observée. Bien qu'étant inférieur aux résultats de *lk-text*, les résultats obtenus avec *lk-free* en utilisant les n-best suggèrent que des niveaux convenables de catégorisation peuvent être atteints même dans des conditions extrêmes de dégradation des documents. Mais cette amélioration du taux de classification n'est pas regulière ou

correlée avec la taille de la liste des n-best. Cela montre que les interactions entre le niveau et le type de bruit présent dans un document reconnu, et le processus de catégorisation n'ont pas encore été comprises et explorées en détail. Puisque nous disposons aujourd'hui d'une base considérable de documents manuscrits, d'autres expériences doivent être effectuées dans ce sens.

Remerciements

Ces travaux ont ete soutenus par la Région Pays de la Loire à travers le projet MILES et par l'Agence Nationale de la Recherche à travers le programme Technologies Logicielles (ANR-06-TLOG-009).

8. Bibliographie

Baeza-Yates R., Ribeiro-Neto B., *Modern Information Retrieval*, Addison-Wesley, 1999.

Beney J., *Classification Supervisée de Documents*, Hermès Science / Lavoisier, 2008.

Dasarathy B. V., *Nearest Neighbor (NN) Norms - NN Pattern Classification Techniques*, IEEE Computer Society Press, 1991.

Debole F., Sebastiani F., « An Analysis of the Relative Hardness of Reuters-21578 Subsets », *Journal of the American Society for Information Science and Technology*, vol. 56, n^o 6, p. 584-596, 2005.

Forman G., « A Pitfall and Solution in Multi-class Feature Selection for Text Classification », *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, p. 38-46, 2004.

Ittner D. J., Lewis D. D., Ahn D. D., « Text Categorization of Low Quality Images », *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR '95)*, p. 301-315, 1995.

Jain A. K., Namboodiri A. M., « Indexing and Retrieval of On-line Handwritten Documents », *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR '03)*, vol. 2, p. 655-659, 2003.

Joachims T., *Learning to Classify Text using Support Vector Machines*, Kluwer Academic Publishers, 2002.

Junker M., Hoch R., « An Experimental Evaluation of OCR Text Representations for Learning Document Classifiers », *International Journal on Document Analysis and Recognition*, vol. 1, n^o 2, p. 116-122, 1998.

Koch G., Catégorisation Automatique de Documents Manuscrits : Application aux Courriers Entrants, PhD thesis, Université de Rouen, 2006.

Lewis D. D., « An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task », *Proceedings of the 15th Annual International ACM SIGIR Conference (SIGIR '92)*, p. 37-50, 1992.

Lopresti D., Tomkins A., « On the Searchability of Electronic Ink », *Proceedings of the 4th International Workshop on Frontiers in Handwriting Recognition (IWFHR '94)*, p. 156-165, 1994.

Manning C., Schütze H., *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

Milewski R. J., Govindaraju V., Bhardwaj A., « Automatic Recognition of Handwritten Medical Forms for Search Engines », *International Journal on Document Analysis and Recognition*, vol. 11, n^o 4, p. 203-218, 2009.

Murata M., Busagala L. S. P., Ohyama W., Wakabayashi T., Kimura F., « The Impact of OCR Accuracy and Feature Transformation on Automatic Text Classification », *Proceedings of the 7th IAPR Workshop on Document Analysis Systems (DAS '06)*, p. 506-517, 2006.

Peña Saldarriaga S., Morin E., Viard-Gaudin C., « Categorization of On-Line Handwritten Documents », *Proceedings of the 8th IAPR International Workshop on Document Analysis Systems (DAS '08)*, p. 95-102, 2008.

Porter M. F., « An Algorithm for Suffix Stripping », *Program*, vol. 14, n^o 3, p. 130-137, 1980.

Salton G., Wong A., Wang C. S., « A Vector Space Model for Automatic Indexing », *Communications of the ACM*, vol. 18, n^o 11, p. 613-620, 1975.

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

Spärck Jones K., « Experiments in Relevance Weighting of Search Terms », *Information Processing & Management*, vol. 15, p. 133-144, 1979.

Taghva K., Nartker T. A., Borsack J., Lumos S., Condit A., Young R., « Evaluating Text Categorization in the Presence of OCR Errors », *Proceedings of Document Recognition and Retrieval VII, IS&T/SPIE Int Symposium on Electronic Imaging (DRR '00)*, vol. 4307, p. 68-74, 2000.

Vapnik V., *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

Vinciarelli A., « Noisy Text Categorization », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, n^o 12, p. 1882-1895, 2005.

Yang Y., « A Study of Thresholding Strategies for Text Categorization », *Proceedings of the 24th Annual International ACM SIGIR Conference (SIGIR '01)*, p. 137-145, 2001.

Yang Y., Liu X., « A Re-examination of Text Categorization Methods », *Proceedings of the 22nd Annual International ACM SIGIR Conference (SIGIR '99)*, p. 42-49, 1999.

Yang Y., Pedersen J. O., « A Comparative Study on Feature Selection in Text Categorization », *Proceedings of the 14th International Conference on Machine Learning (ICML '97)*, p. 412-420, 1997.