Approches textuelles pour la catégorisation et la recherche de documents manuscrits en-ligne

Sebastián Peña Saldarriaga

Université de Nantes

École Doctorale

« Sciences et Technologies de l'Information et de Mathématiques »

Année : 2010

Thèse

de

Doctorat

de l'Université de Nantes

Spécialité : Informatique

Présentée et soutenue publiquement par

Sebastián Peña Saldarriaga

le 24 mars 2010

à la faculté de Sciences et Techniques de Nantes

Titre

Approches textuelles pour la catégorisation et la recherche de documents manuscrits en-ligne

Jury

Président :	Pascale Sébillot, Professeur	IRISA
Rapporteurs :	Laurence Likforman-Sulem, MC-HDR	Télécom ParisTech
	Pascale Sébillot, Professeur	IRISA
Examinateurs :	Alessandro Vinciarelli, Senior researcher	Université de Glasgow
	Pierre-Michel Lallican, Responsable R&D	Vision Objects

Directeur de thèse :	**Christian Viard-Gaudin**, Professeur
Laboratoire :	IRCCyN - UMR CNRS 6597
Co-encadrant:	Emmanuel Morin, Professeur
Laboratoire :	LINA - UMR CNRS 6241

Composante de rattachement du directeur de thèse : IUT - Université de Nantes

N° bu _____ N° ed 503-085

Approches textuelles pour la catégorisation et la recherche de documents manuscrits en-ligne

Text-based Approaches to On-Line Handwritten Document Categorization & Retrieval

Sebastián Peña Saldarriaga

Université de Nantes

Sebastián Peña Saldarriaga

Approches textuelles pour la catégorisation et la recherche de documents manuscrits en-ligne

xviii+145 p.

Ce document a été préparé avec LaTeX, logiciel libre, et la classe `memoir`. Toutes les images originales ont été créées avec le paquetage `PGF/TikZ`, les graphiques ont été créés avec `pgfplots`.

L'auteur a tenté autant que possible de privilégier la langue française aux anglicismes récurrents. Ainsi, il privilégie notamment *et collab.* au latinisme *et al.* importé de l'usage anglais, ainsi que *c-à-d* et *p. ex.* aux latinismes *i.e.* et *e.g.* également importés de l'usage anglais.

Fichier : phd-thesis.tex Modifié le : 2010-05-05 14:16:10.

A la memoria de nuestros muertos

La memoria

Los viejos amores que no están,
la ilusión de los que perdieron,
todas las promesas que se van,
y los que en cualquier guerra se cayeron

El engaño y la complicidad
de los genocidas que están sueltos,
el indulto y el punto final
a las bestias de aquel infierno

Todo está guardado en la memoria,
sueño de la vida y de la historia

La memoria despierta para herir
a los pueblos dormidos
que no la dejan vivir
libre como el viento

León Gieco

Remerciements

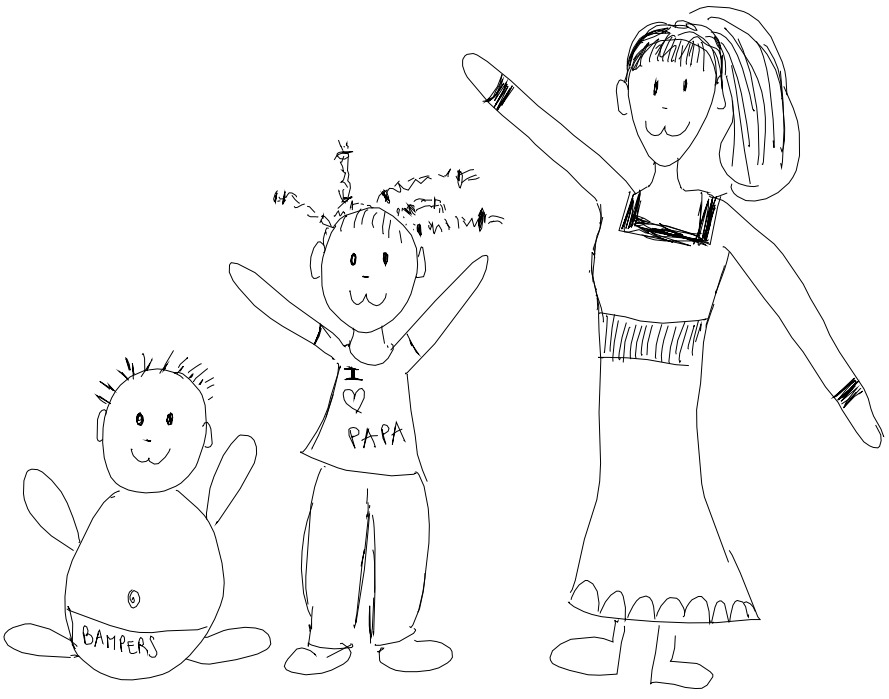
J'ai fait une thèse pour essentiellement deux raisons. La première c'est pour pouvoir faire des remerciements. La deuxième c'est pour les ellipses...

❧ ❧ ❧

Comme ça vous savez qu'un certain temps s'est écoulé entre le moment où j'ai écrit les premières lignes et celles-ci. J'ai toujours trouvé ça génial les effets typographiques comme dans les romans des éditions Harlequin.

❧ ❧ ❧

Les gens ignorent sans doute que pendant la rédaction de ma thèse je suis devenu alcoolique, caféinomane et dépressif. Alors, j'avais trois traitements d'appoint dans les manifestations anxieuses ou dépressives sévères pouvant survenir.



Malgré les nombreux effets secondaires et excipients à effet notoire (bave radioactive, troubles du sommeil ; énurésie, questionnite aiguë ; rédaction forcée jusqu'à tard la nuit, troubles du *farniente* en rond), cette thèse a été entièrement motivée par, et d'une certaine façon elle est aussi due à, ces traitements d'appoint. Leur soutien a été capital dans les moments les plus difficiles de la rédaction. Merci à eux !

L'accomplissement de quelque chose à un moment donné est dû à un ensemble de facteurs. Par exemple, sans mes parents cette thèse n'aurait jamais vu le jour (à la limite une contrefaçon aurait été possible), mais sans mes grands-parents il n'y aurait pas eu de parents et sans arrière-grands-parents il n'y aurait pas eu de grands-parents...

❧ ❧ ❧

... et puis le Big Bang. Enfin, parmi toutes ces entités-là, il y a tout de même les partenaires du projet CIEL. Je tiens à remercier l'ANR d'avoir financé le projet. Je tiens à remercier aussi mon directeur de thèse Christian V.-G et mon encadrant préféré Emmanuel M., qui ont toujours su trouver les mots positifs pour encourager mon travail, et ce malgré des débuts un peu difficiles. Il a été pour moi un plaisir d'avoir pu travailler avec eux, ainsi qu'avec l'équipe du partenaire industriel : Pédro et Freddy, qui m'ont permis de travailler en toute liberté (il ne faut pas que j'oublie Dorothée pour la logistique de la collecte aussi !). Cela a grandement contribué à la réussite de ce travail.

Je tiens également à remercier également les rapporteuses de cette thèse Laurence Likforman-Sulem et Pascale Sébillot pour l'intérêt qu'elles ont porté à mon travail. Je remercie également Alessandro Vinciarelli d'avoir accepté de faire partie de mon jury ainsi que pour les différentes discussions que nous avons eues et qui ont influencé d'une manière ou d'une autre mon travail.

❧ ❧ ❧

Il paraît que quelqu'un qui te casse les pieds en France peut déclencher une tornade aux Texas, alors vous imaginez pas ce que ça ferait comme monde à remercier ! En réalité je ne sais pas si pour tous les gens qui vont être cités ici, comme à la fin d'un film, il s'agit de remerciements, finalement je ne leur dois rien, mais d'une façon ou d'une autre ils ont été là à un moment de ma vie.

Par exemple, je peux commencer par les douaniers qui m'ont accueilli à mon arrivée en France avec un bocal pour faire pipi et un vidage en règle de mes trésors adolescents soigneusement emballés dans une valise pleine à craquer.

À Paris

Je n'oublie pas M. Gourdain, artisan de mon apprentissage de la langue française. Je n'oublie pas non plus mes amis du lycée Dorian à l'époque dorée du BTS Info Indus' (Ad Mérieux, Benji, Stéphon, Cyrille, Berat) ni mes enseignants (M^{mes} Cardona et Santembien, M. Alegria) en particulier M. Bouré (pour son application à la prononciation sans faut de mon nom de famille). Un grand salut à mes camarades d'errances parisiennes Tom Chekler et Mingo Jamet, ainsi qu'à Pierre Berger (à la revoyure camarade !) et Jody et Mary Neil (porte de Clignancourt).

Mes salutations vont également à tous les potes des années MLV, en particulier Sus' Vallée et Nath Persin, les Juliens, L. Garcia, Baranzini, Phil, les rouquins, les

malgachos, les colocs et M. Petidant entre autres. Je me rappelle avec une certaine gratitude d'une partie du corps enseignant (G. Roussell, R. Forax, P. Peterlongo, ...). Á François et tous les membres du (feu le) service logistique-ventes (Jungheinrich) ; Fabienne et Aurélien (chez DayByDay).

À Lannion

Je remercie également Dom, Karl, Lisa, Joseph et Fréd B. pour tous les moments agréables passés à l'époque de France Télécom, sans oublier bien sûr Virginie (la mawaine) qui depuis le temps a toujours été là pour soutenir la Peña Saldarriaga Family.

À Rennes

Cyrille et Laurent, Julie, Manu, Stéphane, Fusco et ses soirées au cercle Paul Bert, Jingqiang et Yifan, les Arguel, les Perrin, Jeannot le brûlot anar et Sophie, Abdallah et Alice, les Chaplain-Delanoë, Jenny Trump *spellcheckeuse*, Hyam et les autres Syriaques, Ludo de l'armée rouge ; Vincent J., Arnaud R. et Pacôme du master chez les toubibs ; Jürgen M., sociologue du volant ; le Dr. Louapre pour avoir favorisé la survie de la famille. Je remercie également les anciens de Fréville : Daria et Choupi, mais aussi ceux qui n'en étaient pas mais qu'on aime quand même : Sandrine et Jeanne, Cécile et Hamid, Arnaud Lamarcq.

Entre Rennes et Nantes

Un grand merci à tous les passagers du TER 58397 à destination de Nantes, départ 7h07, arrivée 8h26, qui ont partagé avec moi pendant presque deux ans la joie des transports en commun (et même des fois un peu de saucisson et un coup de pif). Ils poussaient même le vice jusqu'à prendre avec moi le 17h58 au départ de Nantes et à destination de Rennes, ce train sera sans arrêt jusqu'à Rennes, terminus du train.

À Nantes

Je tiens à remercier également toutes les personnes rencontrées durant ces trois dernières années au LINA : Fabien Poulard pour avoir et pour continuer à assurer la logistique de la Peña Saldarriaga Family, Benoît et Manu pour l'opération toit contre nourriture, Jim pour m'avoir bourré la gueule à Barcelone, Lorraine et Anthony parce que Soazic a fait pipi chez eux, Anne-Françoise et les Annies pour la logistique de mes nombreux voyages et les réponses à mes nombreuses questions.

À tous les doctorants qui ont un jour ou l'autre partagé une table du RU avec moi : Nico Berger, Olivier L., Matthieu, Momo J., Momo M., Manal (merci pour le lit de Soazic), Guillaume P., Olivier B., Marie P., Thomas V., Amir H. À certains membres permanents du labo, dont une partie de ceux siégeant au conseil : Alexandre G., Charlotte T., Christian A., Philippe L., sans oublier tous les membres de l'équipe TALN (y compris sa directrice), Sébastien F., Jean-François R., Sylvie B., Jean G. et les gentils étudiants de l'IUT (mais seulement les gentils).

Je remercie également les collègues de l'IRCCyN (IVC) : Patrick, Sylvain, Florent, Tan, Jin, Montaser, Marcus, Frank Ciaramello et tous les habitués de la Californie au mois de Janvier, ses incomparables soirées Ramen cinq minutes avant la fermeture du restaurant, ses billards, son Ramada San José, ses restos routiers et son magnifique cru du Clos LaChance (not to oaky) pour arroser le tout. Je profite également pour

remercier Thierry P. et Kamel A.M. pour le tour guidé de San Francisco et pour la visite du SFMoMA ; et Jean-Luc Bloechle pour nos enrichissantes rencontres à diverses conférences et ses chemises hawaïennes.

Merci également au Dr. Albert d'avoir pris le relais du Dr. Louapre, j'espère qu'un jour on aura le droit à un tarif groupé ; à la Famille Estéguy Sanchez et Alcides ; à Céline et ses filles.

❧ ❧ ❧

Ma plus sincère affection à Don Jesús y Doña Leonila, à tous les membres, extrêmement nombreux, de la famille Peña restés en Colombie, ainsi qu'à ceux résidant dans la bonne vieille Europe : la tía Lucia, Natay, Tatiana, Ana Sofia et Catalina.

Ma plus sincère affection également à Don Gabriel, Vicky, las niñas, et tous les membres, moins nombreux, de la famille Saldarriaga restés en Colombie, ainsi qu'à ceux résidant en France pour leur soutien tout au long de mon séjour icite : Martin et les Saldarriaga Fuertes, Carolina et les Antiphon Saldarriaga.

Ma plus sincère affection également à Jeannine, Jean-Louis, Gilbert, Margot et tous les membres de la belle famille restés dans le Loir-et-Cher, à Saint-Malo ou en Normandie pour leur soutien, l'initiation culinaire, viticole et goutticole.

❧ ❧ ❧

« De façon générale je remercie toutes les personnes avec qui mes rapports furent aussi divers qu'enrichissants. »

Sommaire

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Bien que l'idée de capturer l'écriture existe depuis le XIX^{e} siècle, grâce au télégraphe manuscrit - « télé-autographe » - (*cf.* figure 1.1), les technologies de saisie gestuelle autorisant l'acquisition de l'écriture manuscrite, issues principalement des recherches sur l'ergonomie des interfaces homme-machine, n'ont véritablement émergé que dans les années 60 (Sutherland, 1963 ; Davis et Ellis, 1964).

Pendant très longtemps, l'utilisation principale de l'écriture en-ligne était de servir d'alternative aux interfaces de commande classiques des ordinateurs : le clavier et la souris. Le geste écrit n'était alors considéré que comme un moyen d'interaction ergonomique pour appareils mobiles disposant de terminaux de petite taille (PDA ou Smartphone).

Les évolutions technologiques des dernières décennies ont favorisé la démocratisation de dispositifs capables de produire de l'écriture en-ligne. Ces évolutions ont permis également d'élargir le spectre des applications de ce type de signaux manuscrits. Les stylos numériques permettent aujourd'hui d'élaborer efficacement des documents complexes. Les Tablet PC peuvent favoriser la prise de notes en situation de mobilité par exemple, où l'utilisation du clavier devient impossible. L'éventualité d'utiliser les stylets

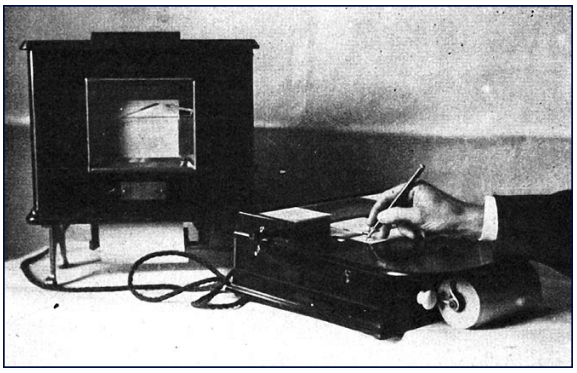


FIGURE 1.1 : Le télégraphe manuscrit inventé par Elisha Gray (US Patent 386,815, Juillet 1888).

(a) Smartphone (b) Tablet PC

(c) Stylo digital (d) Tableau blanc interactif

FIGURE 1.2 : Dispositifs permettant la saisie de documents manuscrits en-ligne.

pour la saisie de textos peut également être envisagée (Prochasson, Viard-Gaudin et Morin, 2007).

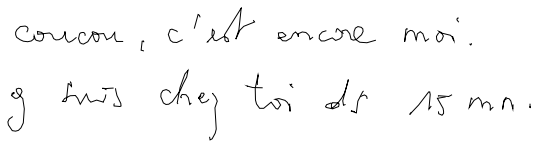
La figure 1.2 montre différents dispositifs de saisie manuscrite en-ligne. Ces dispositifs peuvent se différencier par leur mode de capture. La surface tactile est le mode le plus courant de capture présent dans les appareils mobiles. Dans ce cas, un stylet est utilisé pour exercer une pression sur la surface. D'autres dispositifs enregistrent le mouvement d'écriture dans l'espace grâce à des systèmes de tracking magnétique (Wacom Bamboo [1], CrossPad [2], etc.), optique (stylos électroniques couplés à des papiers à trame de type Anoto [3]) ou par triangulation ultrasonique (ZPen [4], eBeam [5], etc.).

Les documents qui peuvent être produits par ces différents dispositifs sont de nature très variée. La figure 1.3 montre des exemples d'échantillons d'écriture en-ligne. Ils peuvent consister aussi bien en de courtes phrases destinées à être envoyées grâce à un appareil mobile, des équations, des dessins, des prises de notes de cours, des poèmes ou des documents annotés. Des blogs manuscrits ont même fait leur apparition [6].

L'engouement autour de ces technologies se traduit par l'apparition de quantités de

1. http://www.wacom.eu/index2.asp?pid=294&lang=en&spid=1
2. http://www.research.ibm.com/electricInk/
3. http://www.anoto.com/
4. http://www.danedigital.com/6-Zpen/
5. http://www.e-beam.com/
6. http://livescribe.com/cgi-bin/WebObjects/LDApp.woa/wa/CommunityResultsPage?cidx=1

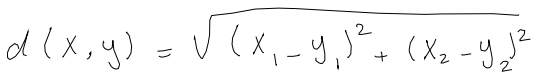
(a) Arabe



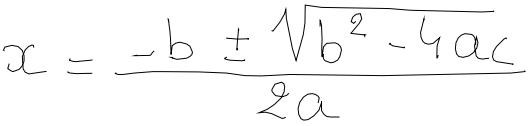
(b) Français

(c) Népalais

Mini-messages manuscrits



(d) Distance entre deux points

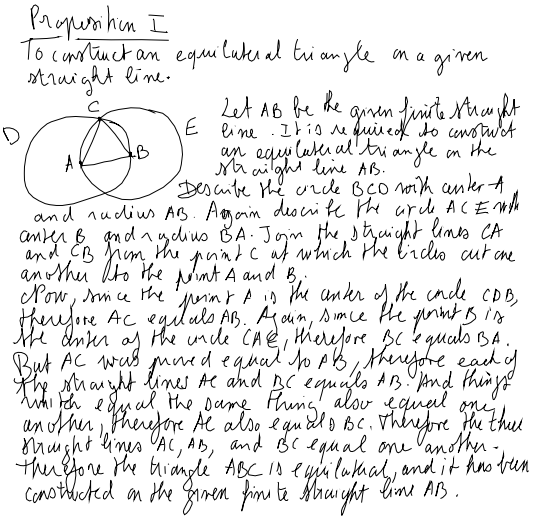


(e) Solution de l'équation du second degrée

Équations

(f) Portrait

FIGURE 1.3 : Échantillons d'écriture manuscrite. Mini-messages manuscrits, équations et dessins.



Proposition I

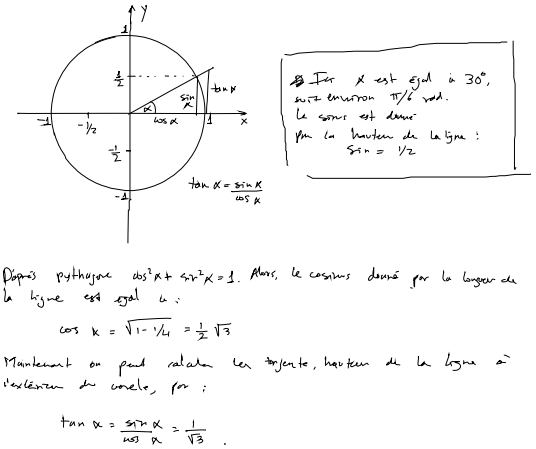
To construct an equilateral triangle on a given straight line.

Let AB be the given finite straight line. It is required to construct an equilateral triangle on the straight line AB.

Describe the circle BCD with center A and radius AB. Again describe the circle ACE with center B and radius BA. Join the straight lines CA and CB from the point C at which the circles cut one another to the point A and B.

Now, since the point A is the center of the circle CDB, therefore AC equals AB. Again, since the point B is the center of the circle CAE, therefore BC equals BA. But AC was proved equal to AB, therefore each of the straight lines AC and BC equals AB. And things which equal the same thing also equal one another, therefore AC also equals BC. Therefore the three straight lines AC, AB, and BC equal one another. Therefore the triangle ABC is equilateral, and it has been constructed on the given finite straight line AB.

(g) Géométrie euclidienne



Pour x est égal à 30°, soit environ $\pi/6$ rad. Le sinus est donné par la hauteur de la ligne : $\sin = 1/2$

Depuis pythagore $\cos^2 x + \sin^2 x = 1$. Alors, le cosinus donné par la longueur de la ligne est égal à :

$$\cos x = \sqrt{1 - 1/4} = \frac{1}{2}\sqrt{3}$$

Maintenant on peut calculer la tangente, hauteur de la ligne à l'extérieur du cercle, par :

$$\tan x = \frac{\sin x}{\cos x} = \frac{1}{\sqrt{3}}.$$

(h) Trigonométrie

Notes de cours

Il règne sur la ville une mort négatrice
L'Arlequin blanc et noir noir et blanc devenu
N'y voit rien de changé
  sinon que les actrices
Accrochant au moyen ~~d'épingles~~
  d'épingles à nourrice
L'ombre des rayons X à leur
  épaule nue
Equation fantôme aux belles inconnues
Ces jours-ci s'est ~~ouvert~~
  ouvert le Carnaval
  à Nice
Personne excepté moi ne s'en est
  souvenu

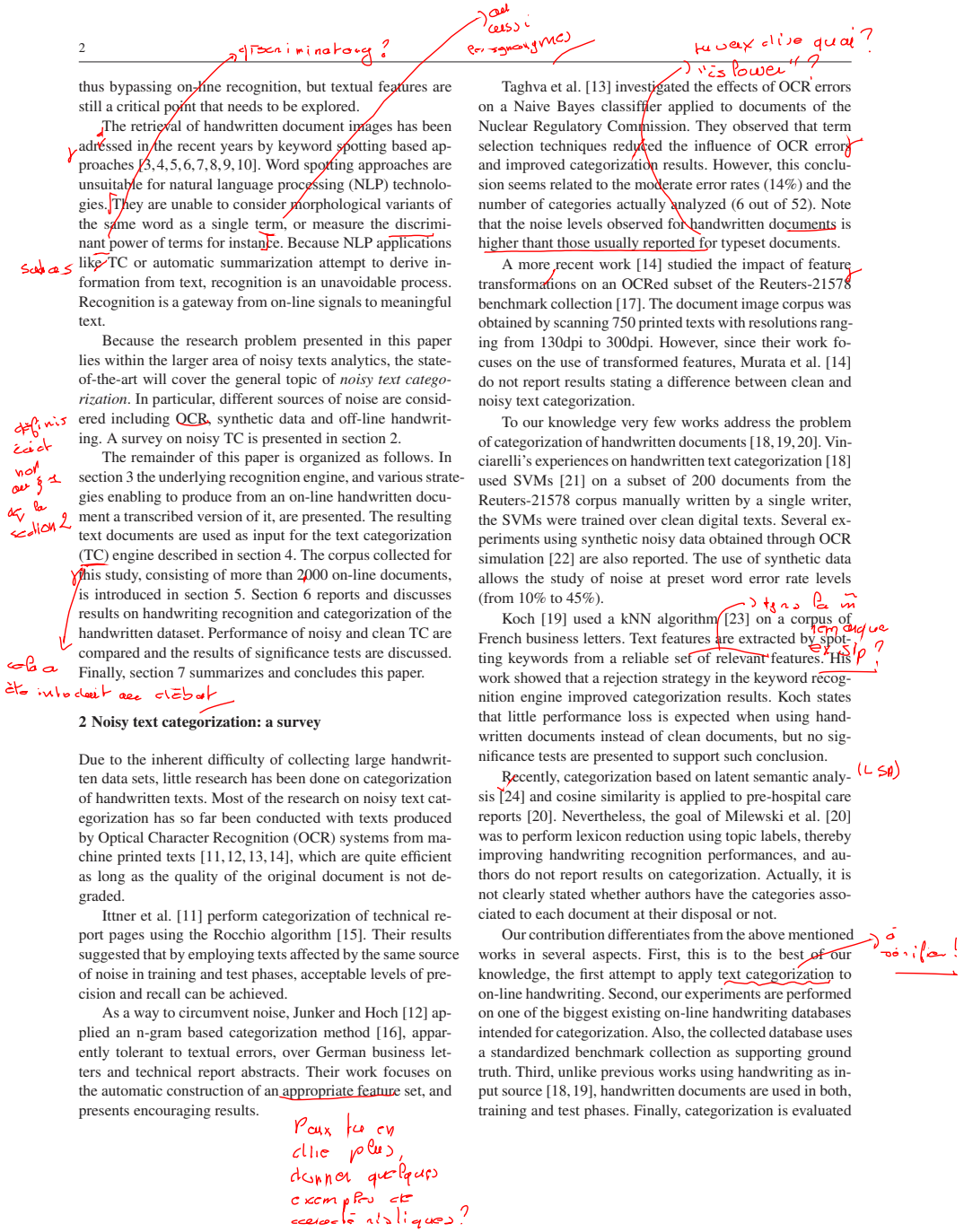
(i) « La nuit en plein midi », Louis Aragon

Tan sólo es necesario vestirnos
color de poesía,
impregnarnos la frente de fragancia
verso libre,
ser prototipos del estilo
canto sin barreras,
caminar del lado de la vida
duro contra el viento
para que seamos declarados
elementos fuera de orden.

(j) « Señales », Chucho Peña.

Poèmes

FIGURE 1.3 : Échantillons d'écriture manuscrite (suite). Notes manuscrites.



2

thus bypassing on-line recognition, but textual features are still a critical point that needs to be explored.

The retrieval of handwritten document images has been adressed in the recent years by keyword spotting based approaches [3,4,5,6,7,8,9,10]. Word spotting approaches are unsuitable for natural language processing (NLP) technologies. They are unable to consider morphological variants of the same word as a single term, or measure the discriminant power of terms for instance. Because NLP applications like TC or automatic summarization attempt to derive information from text, recognition is an unavoidable process. Recognition is a gateway from on-line signals to meaningful text.

Because the research problem presented in this paper lies within the larger area of noisy texts analytics, the state-of-the-art will cover the general topic of *noisy text categorization*. In particular, different sources of noise are considered including OCR, synthetic data and off-line handwriting. A survey on noisy TC is presented in section 2.

The remainder of this paper is organized as follows. In section 3 the underlying recognition engine, and various strategies enabling to produce from an on-line handwritten document a transcribed version of it, are presented. The resulting text documents are used as input for the text categorization (TC) engine described in section 4. The corpus collected for this study, consisting of more than 2000 on-line documents, is introduced in section 5. Section 6 reports and discusses results on handwriting recognition and categorization of the handwritten dataset. Performance of noisy and clean TC are compared and the results of significance tests are discussed. Finally, section 7 summarizes and concludes this paper.

2 Noisy text categorization: a survey

Due to the inherent difficulty of collecting large handwritten data sets, little research has been done on categorization of handwritten texts. Most of the research on noisy text categorization has so far been conducted with texts produced by Optical Character Recognition (OCR) systems from machine printed texts [11,12,13,14], which are quite efficient as long as the quality of the original document is not degraded.

Ittner et al. [11] perform categorization of technical report pages using the Rocchio algorithm [15]. Their results suggested that by employing texts affected by the same source of noise in training and test phases, acceptable levels of precision and recall can be achieved.

As a way to circumvent noise, Junker and Hoch [12] applied an n-gram based categorization method [16], apparently tolerant to textual errors, over German business letters and technical report abstracts. Their work focuses on the automatic construction of an appropriate feature set, and presents encouraging results.

Taghva et al. [13] investigated the effects of OCR errors on a Naive Bayes classifier applied to documents of the Nuclear Regulatory Commission. They observed that term selection techniques reduced the influence of OCR errors and improved categorization results. However, this conclusion seems related to the moderate error rates (14%) and the number of categories actually analyzed (6 out of 52). Note that the noise levels observed for handwritten documents is higher thant those usually reported for typeset documents.

A more recent work [14] studied the impact of feature transformations on an OCRed subset of the Reuters-21578 benchmark collection [17]. The document image corpus was obtained by scanning 750 printed texts with resolutions ranging from 130dpi to 300dpi. However, since their work focuses on the use of transformed features, Murata et al. [14] do not report results stating a difference between clean and noisy text categorization.

To our knowledge very few works address the problem of categorization of handwritten documents [18,19,20]. Vinciarelli's experiences on handwritten text categorization [18] used SVMs [21] on a subset of 200 documents from the Reuters-21578 corpus manually written by a single writer, the SVMs were trained over clean digital texts. Several experiments using synthetic noisy data obtained through OCR simulation [22] are also reported. The use of synthetic data allows the study of noise at preset word error rate levels (from 10% to 45%).

Koch [19] used a kNN algorithm [23] on a corpus of French business letters. Text features are extracted by spotting keywords from a reliable set of relevant features. His work showed that a rejection strategy in the keyword recognition engine improved categorization results. Koch states that little performance loss is expected when using handwritten documents instead of clean documents, but no significance tests are presented to support such conclusion.

Recently, categorization based on latent semantic analysis [24] and cosine similarity is applied to pre-hospital care reports [20]. Nevertheless, the goal of Milewski et al. [20] was to perform lexicon reduction using topic labels, thereby improving handwriting recognition performances, and authors do not report results on categorization. Actually, it is not clearly stated whether authors have the categories associated to each document at their disposal or not.

Our contribution differentiates from the above mentioned works in several aspects. First, this is to the best of our knowledge, the first attempt to apply text categorization to on-line handwriting. Second, our experiments are performed on one of the biggest existing on-line handwriting databases intended for categorization. Also, the collected database uses a standardized benchmark collection as supporting ground truth. Third, unlike previous works using handwriting as input source [18,19], handwritten documents are used in both, training and test phases. Finally, categorization is evaluated

(k) Version préliminaire d'un article (Peña Saldarriaga et collab., 2010d)

FIGURE 1.3 : Échantillons d'écriture manuscrite (suite et fin). Document imprimé annoté avec de l'encre numérique.

plus en plus importantes de documents manuscrits en-ligne. Cela pose la question de l'accès à l'information contenue dans ces données. En effet, contrairement aux documents textuels, directement interprétables par l'ordinateur, les traits d'écriture n'ont pas de sens immédiat pour lui. De plus, en considérant la diversité de contenus que peut contenir un même document (*cf.* figure 1.3(h), par exemple), le challenge s'accentue.

Ce travail s'intéresse à l'accès à l'information textuelle contenue dans les documents manuscrits en-ligne. Il est donc considéré ici que les blocs textuels ont été préalablement extraits des documents. L'un des avantages du texte par rapport aux croquis, équations, tableaux, etc., est qu'il s'agit déjà d'un ensemble cohérent de symboles qui véhiculent un message. Le parti pris de notre étude est celui de la « textualité ». La textualité est ce qui est lié au texte, ce qui peut englober les phénomènes de production et identification de mots, de production et de compréhension de textes, indépendamment de leur mode de représentation informatique.

1.1 Catégorisation et recherche de documents manuscrits en-ligne

Alors que les travaux sur l'utilisation des moyens informatiques pour la recherche d'information (RI) ont débuté dans les années 50, les premières études concernant la recherche de documents en-ligne ne sont apparues que dans les années 90. Ceci est dû au fait que jusqu'à cette date, le souci principal dans le domaine du *pen computing* était le perfectionnement des dispositifs de saisie. La figure 1.4 montre l'évolution du domaine de la RI parallèlement à celui de l'écriture en-ligne à travers divers évènements clés survenus depuis les années 50. L'axe central de la frise chronologique correspond aux travaux à la croisée de ces deux domaines de recherche.

Comparativement au domaine du texte électronique, peu d'applications de recherche d'information existent pour les documents manuscrits. De plus, ces approches correspondent à une approche booléenne de la RI aujourd'hui dépassée (Vinciarelli, 2006). Ce travail a pour but d'apporter des fonctionnalités de catégorisation et de recherche pour documents en-ligne. Nos contributions se situent dans la ligne temporelle centrale (*cf.* figure 1.4). Il s'agit d'enrichir les travaux en RI dans le domaine des documents en-ligne en s'appuyant sur un panel de méthodes ayant fait leurs preuves en RI classique.

La catégorisation est un processus d'organisation automatique des données dans le but de faciliter leur exploitation ultérieure. L'objectif est de regrouper des documents en classes à partir d'exemples. Parmi les applications de la catégorisation nous pouvons citer, entre autres, le routage et l'indexation thématique de documents.

En ce qui concerne la recherche de documents, nous cherchons à développer des techniques permettant, en fonction de critères de recherche propres à un utilisateur, de trouver des documents particuliers dans des fonds de documents manuscrits en-ligne.

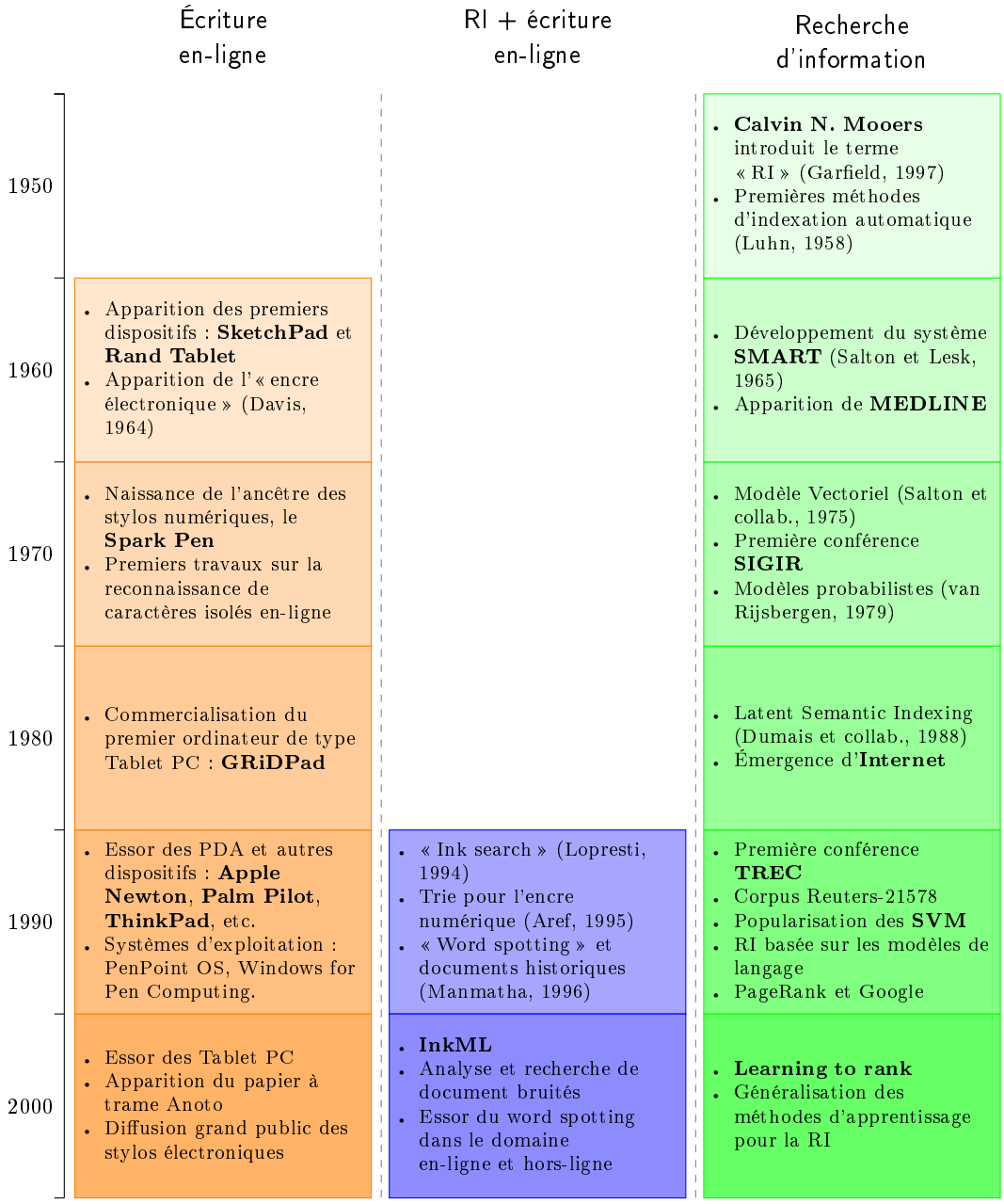


FIGURE 1.4 : Frise chronologique de divers évènements clés du développement des deux disciplines concourantes abordées par ce travail : la recherche d'information et l'écriture en-ligne. L'axe central correspond au domaine résultant du croisement de ces dernières.

1.2 Reconnaissance de l'écriture en-ligne

Une façon d'appréhender le problème de la recherche de documents en-ligne est de s'intéresser aux transcriptions textuelles issues d'un système de reconnaissance de l'écriture.

La reconnaissance de l'écriture manuscrite peut être envisagée à différents niveaux

de difficulté croissante : caractères isolés, mots ou phrases. Il s'agit d'une application typique de la reconnaissance de formes mais qui soulève des problématiques très complexes, liées, par exemple, à la variabilité interindividuelle, voire intra-individuelle, très importante de l'écriture. Cette variabilité est encore plus importante dans le cas de l'écriture en-ligne de par la prise en compte de la temporalité (voir chapitre 2, §2.1).

Force est de constater que malgré des avancées significatives dans le domaine de la reconnaissance (Perraud, 2005), les textes qui en résultent contiennent des erreurs, *du bruit*. Selon la quantité d'erreurs, les transcriptions qui en résultent peuvent être illisibles pour un humain. Cependant, elles contiendraient quand même suffisamment d'information pour que l'application des technologies de RI soit possible (Vinciarelli, 2006).

Par ailleurs, tout un domaine de recherche s'est constitué autour de la question de l'exploitation de ces textes « bruités », appelé *Noisy Text Analytics* (Knoblock, Lopresti, Roy et Subramaniam, 2007 ; Lopresti, Roy, Schulz et Subramaniam, 2008, 2009). L'utilisation des transcriptions issues d'un processus de reconnaissance se rapproche inéluctablement de travaux sur des textes aux origines diverses : documents manuscrits hors-ligne, images de documents dactylographies, facsimilés, enregistrements audio, etc. Notre travail se situe pleinement dans ce nouveau domaine de recherche puisque cette thèse se concentre en partie sur l'analyse du comportement des approches proposées face aux erreurs de reconnaissance.

1.3 Contributions et plan du document

La plupart des approches proposées dans ce travail sont appliquées pour la première fois au domaine manuscrit. Il s'agit de méthodes issues du domaine de la RI. Du point de vue de la RI, l'élément novateur de ce travail est lié à l'utilisation d'un type de données émergeant.

L'originalité de ce travail est liée à son positionnement au carrefour de plusieurs de disciplines. Ce caractère interdisciplinaire nous a conduit à consacrer une partie de ce travail à identifier des méthodes pertinentes pour notre problématique tout en motivant leur utilisation.

Cadre de la thèse

Ces travaux de thèse ont été effectués dans le cadre du projet CIEL soutenu par l'Agence Nationale pour la Recherche (ANR) sous le numéro ANR-06-TLOG-009. Le projet ANR CIEL (Conversion Indexation de l'Écriture en-Ligne) associe les laboratoires IRCCyN (Institut de Recherche en Communications et Cybernétique de Nantes - UMR CNRS 6597) et LINA (Laboratoire d'Informatique de Nantes-Atlantique - UMR CNRS 6241) avec la société Vision Objects.

La société Vision Objects est actuellement le principal fournisseur mondial de technologies de reconnaissance de l'écriture manuscrite en-ligne. Sa vocation n'est pas le développement d'applications dédiées à des utilisateurs finaux, mais le développement de briques logicielles intégrables. Sa technologie est proposée à travers un SDK (Software

Development Kit). Ce dernier, nommé MyScript® Builder, permet à des intégrateurs, clients de la société, d'introduire des fonctionnalités de reconnaissance de l'écriture dans leurs produits logiciels. Ce SDK a été intégré dans les prototypes logiciels issus de ce travail et devient de fait un support important de cette étude.

Un des axes du projet, RIDEL (Recherche d'Information dans les Documents En-Ligne), vise à développer des modèles de recherche d'information dans des bases de données constituées de documents manuscrits en-ligne, aussi bien à partir de requêtes exprimées en langage naturel ou sous la forme d'échantillons d'écriture. Il s'agit de l'axe dans lequel se situe ce travail.

Contributions

Les principales contributions de cette thèse sont :

1. Une analyse détaillée du comportement de deux méthodes de catégorisation, basées sur l'apprentissage automatique, face aux erreurs de reconnaissance. De plus, une nouvelle perspective d'analyse est ouverte par l'intérêt porté aux perturbations induites par le processus de reconnaissance dans la structure des données. Cette perspective se situe au plus prés de l'élément essentiel des deux méthodes de classification étudiées, à savoir la fonction noyau.
2. Une étude systématique de la combinaison des différentes approches existantes pour la recherche de documents manuscrits en-ligne. En pratique, il est difficile de déterminer laquelle de ces approches est la meilleure. Plutôt que de les opposer nous proposons de les faire coopérer. Nous montrons qu'il est possible de tirer parti de la diversité des résultats restitués par des algorithmes de recherche différents, pour construire des systèmes combinés plus robustes que les différents systèmes considérés individuellement.
3. Un algorithme souple de reclassement de documents et son application à la recherche de documents manuscrits. Cet algorithme trouve sa source au croisement de la *cluster hypothesis* (Jardine et van Rijsbergen, 1971) et de l'apprentissage semi-supervisé. L'algorithme exploite les relations de similarité entre les documents. Bien que nous l'ayons motivé pour sa capacité à pallier les limitations des algorithmes de *word spotting*, il s'agit d'une méthode générique pouvant être appliquée à des résultats de recherche tout venant et extensible à des critères de recherche non textuels.

Organisation du manuscrit

Hormis l'introduction et la conclusion, ce document est structuré autour de trois parties logiques. Une première partie introduit les deux objets transversaux à l'ensemble de l'étude :

Chapitre 2 : Reconnaissance de documents manuscrits en-ligne. Ce chapitre décrit les particularités liées à l'écriture en-ligne. Il fait également une revue des principales approches de reconnaissance de l'écriture avant d'aborder le système de reconnaissance utilisé.

Chapitre 3 : Collection de données manuscrites. Puisque l'étude de la catégorisation et de la RI requièrent des corpus particuliers, une première étape importante de cette étude a consisté à collecter un corpus manuscrit de référence pour sa validation expérimentale. Nous décrivons dans ce chapitre les modalités et résultats de la collecte d'un corpus de données manuscrites adapté à la recherche d'information.

La deuxième partie de ce manuscrit aborde le problème de la catégorisation de textes issus d'un processus de reconnaissance de l'écriture :

Chapitre 4 : Catégorisation automatique de textes. Ce chapitre est dédié à l'introduction des notions et des méthodes nécessaires à la compréhension des chapitresS de ce mémoire liées à la catégorisation automatique de documents manuscrits.

Chapitre 5 : Impact des erreurs de reconnaissance dans la catégorisation. Ce chapitre s'intéresse à l'application des méthodes d'apprentissage à notre problématique. Plus particulièrement, nous nous intéressons aux problèmes qui surgissent lorsqu'elles opèrent sur des textes contenant de nombreuses erreurs de reconnaissance.

Enfin, la troisième et dernière partie s'attaque à la recherche d'information dans des bases documentaires manuscrites :

Chapitre 6 : À la recherche du document manuscrit en-ligne. Ce chapitre a pour vocation de définir la recherche d'information. Il s'attache également à faire la distinction entre RI et *word spotting*. Nous présentons quelques modèles de recherche de documents manuscrits et livrons des données expérimentales sur l'application de ces modèles à notre corpus.

Chapitre 7 : Fusion de résultats en RI appliquée aux documents en-ligne. L'objectif de ce chapitre est de montrer l'intérêt des méthodes de métarecherche appliquées à notre problématique. Les résultats montrent que, de manière générale, la fusion permet d'obtenir un métasystème plus performant que les systèmes individuels.

Chapitre 8 : Régularisation de résultats de word spotting. Nous décrivons un algorithme permettant d'améliorer les résultats de recherche d'un algorithme de word spotting tout-venant. Le bilan expérimental atteste de son efficacité.

La synthèse, le bilan et les perspectives de cette thèse feront l'objet du chapitre 9 qui clôture le manuscrit.

Chapitre 2

Reconnaissance de documents manuscrits en-ligne

La reconnaissance de l'écriture est le processus de transformation d'un texte manuscrit en un texte électronique. Dans le domaine de la reconnaissance, les textes manuscrits se distinguent selon leur mode d'acquisition. La première partie de ce chapitre introduit les deux types de signaux manuscrits existants, à savoir, les signaux en-ligne et hors-ligne (§2.1), ainsi que les principales approches de reconnaissance de ces signaux (§2.2). Dans une seconde partie, ce chapitre présente le système de reconnaissance utilisé dans le cadre de nos expérimentations (§2.3) ainsi que les mesures d'évaluation classiques de la reconnaissance de l'écriture (§2.4).

2.1 Signaux en-ligne et hors-ligne

Selon les modes d'acquisition du signal, une distinction est souvent effectuée entre reconnaissance en-ligne et reconnaissance hors-ligne. Dans le cas de la reconnaissance hors-ligne, l'acquisition de l'écriture se fait généralement à l'aide d'un scanner. L'acquisition est donc réalisée après l'opération d'écriture et les données se présentent sous forme d'images numériques à deux dimensions, c'est-à-dire des matrices de pixels (*cf.* figure 2.1).

Dans le cas de l'écriture en-ligne, l'acquisition du signal peut se faire, par exemple, grâce à un stylo électronique ou à l'aide d'un stylet associé à une surface tactile. L'acquisition est, cette fois-ci, réalisée durant l'opération d'écriture et les données capturées se présentent alors comme une séquence de points ordonnés dans le temps.

Les approches de reconnaissance de ces deux types de signaux se basent souvent sur les mêmes techniques d'apprentissage et de modélisation (Lorette et Paquet, 2006). Cependant, le signal en-ligne possède des caractéristiques différentes du signal hors-ligne : il est plus pur et plus riche. Plus pur par l'absence de fond, ce qui évite les pré-traitements d'extraction des pixels d'écriture du reste de l'image. Plus riche, car il peut capturer des métadonnées sur l'écriture elle-même, notamment l'ordonnancement

FIGURE 2.1 : Exemple d'un tracé hors-ligne pour le mot *stock*.

temporel du tracé.

Lorsque seules les caractéristiques liées à la forme du tracé sont utilisées pour la reconnaissance, l'approche en-ligne ne se distingue guère de l'approche hors-ligne (Lorette et Paquet, 2006). Cependant, lorsque la temporalité du tracé est prise en compte, la différence entre les signaux devient évidente (*cf.* figure 2.2).

Les deux tracés pour la lettre `p` donnés en figure 2.2(a) paraissent identiques. Cependant, lorsque les points sont tracés en fonction de l'axe temporel t, la figure 2.2(b) montre que dans le cas du premier `p`, la boucle a été réalisée après le trait descendant, alors que pour le deuxième la boucle a été tracée en premier.

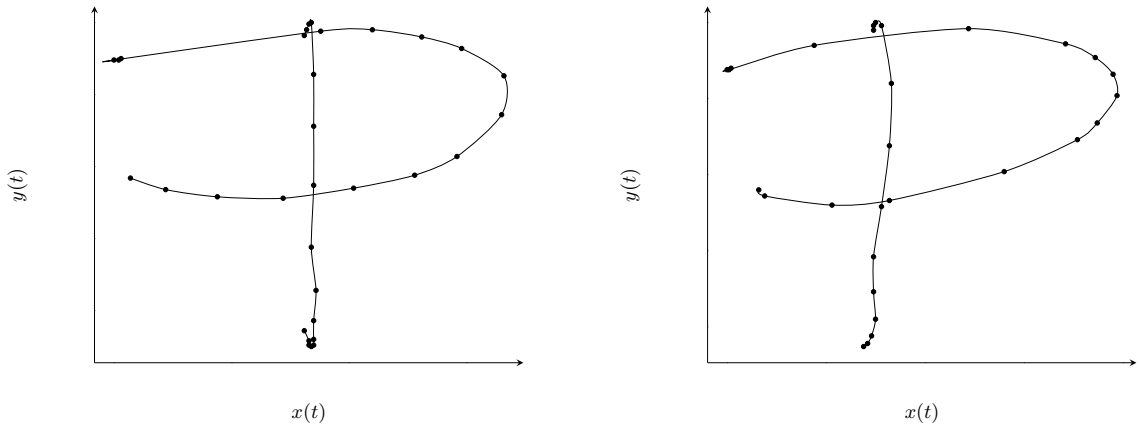
2.2 Reconnaissance de l'écriture

De manière générale, la reconnaissance a pour objectif la retranscription de séquences de mots : phrases, paragraphes ou textes complets. Cela suppose une segmentation préalable du tracé en-ligne en caractères et en mots. Cette section ne s'intéresse pas directement au processus de segmentation, mais plutôt aux résultats qu'il engendre. En effet, la segmentation peut être vue comme une transformation du signal brut en une représentation de plus haut niveau (§2.2.1). En elle-même, la reconnaissance peut être définie comme un ensemble de transformations permettant de passer d'un signal manuscrit à un texte électronique. La dernière partie de cette section présente succinctement différentes approches de la reconnaissance de mots manuscrits (§2.2.2), en s'inspirant principalement de l'état de l'art dressé par Lorette et Paquet (2006).

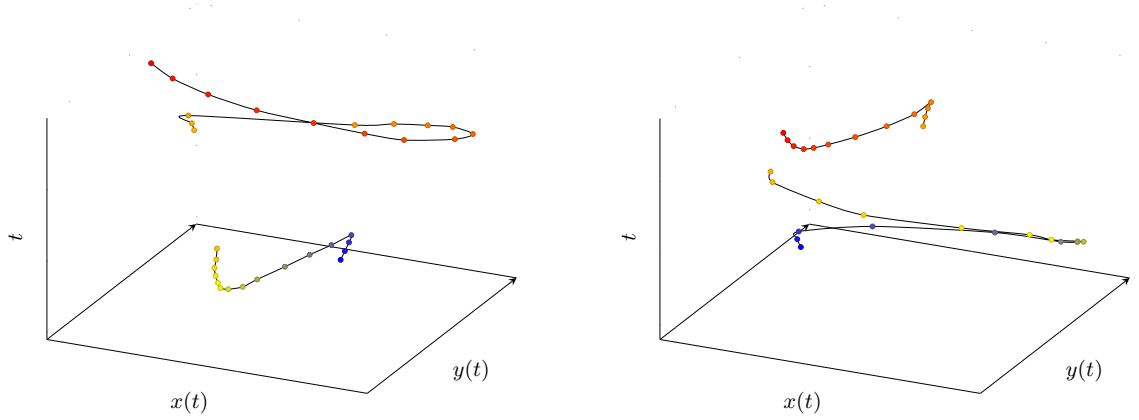
2.2.1 Niveaux de représentation

Selon Lopresti et Tomkins (1994), l'encre numérique peut être modélisée à travers divers niveaux d'abstraction. Au niveau le plus bas, il s'agit des points bruts, au niveau le plus haut, il s'agit des textes électroniques correspondant aux mots (*cf.* figure 2.3).

La suite de points $\mathcal{P} = \{p(0), p(1), \ldots, p(n)\}$ est la représentation la plus immédiate du signal acquis correspondant à un tracé manuscrit. S'en suit un ensemble de prétraitements visant à réduire la variabilité de $\mathcal{P}$. Schématiquement, il faut distinguer le ré-échantillonnage de la normalisation. Le ré-échantillonnage vise à rendre les points



(a) Tracés manuscrits



(b) Tracés en fonction du temps

FIGURE 2.2 : Exemples de tracés manuscrits pour la lettre p avec ou sans prise en compte de l'ordonnancement temporel. L'ordonnancement temporel est illustré par l'axe t et la couleur des points. Les points en bleu ont été échantillonnés en premier et les rouges en dernier.

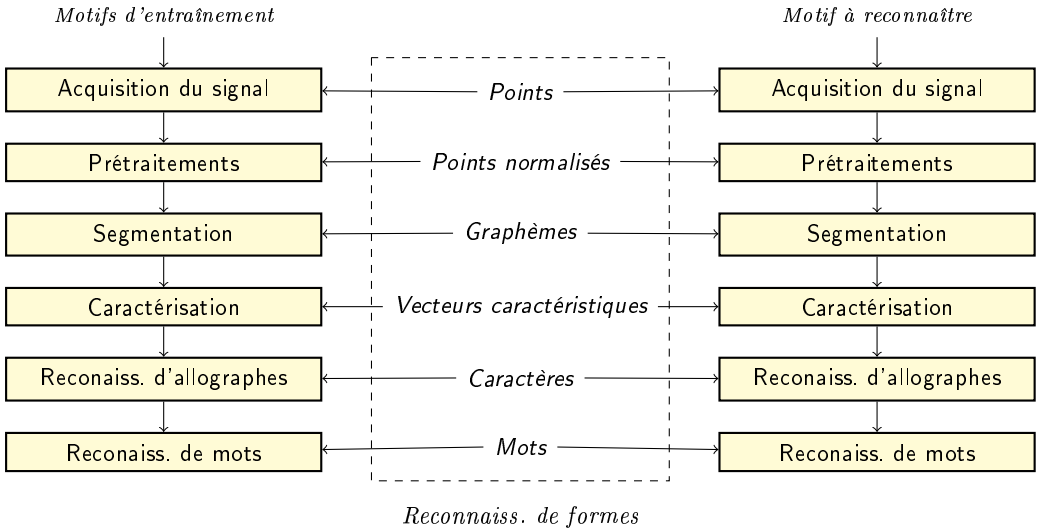


FIGURE 2.3 : Niveaux de représentation de l'encre.

du signal équidistants dans l'espace. Cela permet de s'affranchir de la vitesse d'écriture du scripteur. La normalisation du signal, quant à elle, permet de réduire la variabilité de l'écriture, principalement au niveau de la taille et de l'inclinaison. Le résultat de ces prétraitements est un ensemble de points normalisés $\widetilde{\mathcal{P}}$.

Le second niveau d'analyse du signal, après les prétraitements, vise à décomposer le mot manuscrit selon les éléments qui le composent. Ces éléments sont appelés graphèmes. Il peut s'agir de caractères ou bien de fragments de caractères. La figure 2.4 montre deux possibles segmentations pour le mot manuscrit `indication` et le résultat de la reconnaissance associé à chacune de ces possibilités.

(a) inolication (b) indication

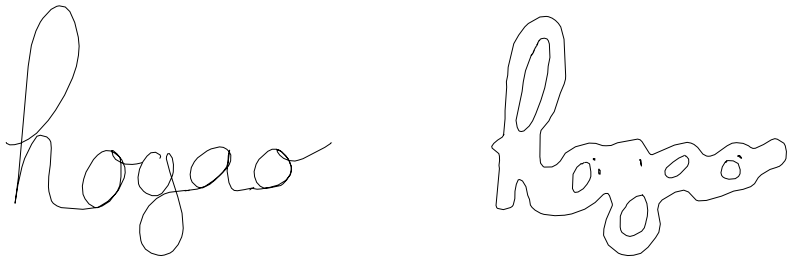
FIGURE 2.4 : Segmentations du mot `indication` en graphèmes.

La troisième étape consiste à extraire les caractéristiques, principalement locales, des éléments du signal comme la hauteur, la courbure ou l'aspect (Jaeger, Manke, Reichert et Waibel, 2001). À partir de ces caractéristiques, il s'agira de produire des hypothèses de reconnaissance de caractères puis de mots. Chaque niveau d'analyse peut être vu canoniquement comme une étape de la reconnaissance de mots manuscrits. Bien entendu, ce canevas ne convient pas à toutes les approches de la reconnaissance de l'écriture, mais peut servir de structure de référence non seulement dans le contexte de la reconnaissance mais aussi dans celui du *word spotting* (§6.2.3).

2.2.2 Approches de la reconnaissance

Les approches de la reconnaissance se différencient selon qu'elles modélisent ou non les mots grâce à la structure alphabétique de la langue. Dans le premier cas, l'approche est dite *analytique*, dans le second elle est dite *globale*.

Dans les approches globales, les mots manuscrits sont considérés comme des entités, uniques et indivisibles, décrites par un ensemble de caractéristiques comme le contour, les boucles ou les traits ascendants et descendants (*cf.* figure 2.5). Lorsque l'approche est dirigée par un lexique de taille fixe, la reconnaissance revient à trouver l'entrée du lexique qui correspond au motif donné en entrée. La majorité des approches globales se basent sur des méthodes qui déterminent une distance entre l'ensemble des caractéristiques du signal en entrée et le modèle de chaque mot du lexique de référence (Madhvanath et Govindaraju, 2001).



(a) Mot manuscrit (b) Contour et boucles

FIGURE 2.5 : Caractéristiques globales (contour et boucles) d'une instance du mot manuscrit `hogao`.

Les approches analytiques s'appuient sur la structure alphabétique de la langue pour modéliser chaque mot. Ainsi, il suffit de spécifier un alphabet pour être capable de modéliser, en théorie, les mots de la langue. Il convient de distinguer deux types d'approches analytiques : les approches *dirigées par le lexique* et les approches *sans lexique*. La reconnaissance sans lexique s'appuie sur les décisions prises au niveau caractère pour proposer des hypothèses de mots. Elle ne fait intervenir les contraintes lexicales qu'éventuellement en post-traitement de la reconnaissance, en utilisant une distance d'édition. Ces approches sont moins performantes que celles dirigées par le lexique car toute erreur dans la reconnaissance d'un caractère se répercute au niveau mot. Cependant, elles peuvent correspondre à des problématiques particulières comme la reconnaissance de codes postaux (Shridhar, Houle et Fumitaka, 1997).

L'approche analytique par *segmentation/reconnaissance* se base sur une modélisation statique de l'alphabet de la langue ainsi que sur une sursegmentation du tracé en éléments de taille inférieure ou égale au plus petit caractère attendu. C'est lors de la reconnaissance que le système devra choisir la segmentation qui correspond le mieux au tracé à reconnaître. Cette approche adopte le paradigme selon lequel « il faut segmenter pour reconnaître, et reconnaître pour segmenter » (Sayre, 1973).

L'approche analytique par *modèles dynamiques de caractères* introduit des entités de modélisation d'un niveau inférieur au caractère. Ces approches ne nécessitent pas

de segmentation explicite, les mots sont souvent décrits par les caractéristiques locales des points qui les composent. Il s'agit de modèles hiérarchiques où les caractères sont modélisés par des séquences de longueur variable et les mots par la concaténation des modèles de caractères.

La reconnaissance s'effectue la plupart du temps grâce à des architectures neuro-markoviennes (Jaeger et collab., 2001 ; Caillault, Viard-Gaudin et Ahmad, 2005). La reconnaissance de caractères s'effectue à l'aide d'un réseau de neurones à décalage temporel (TDNN) s'appuyant sur une fenêtre d'observations mobile (Waibel, Toshiyuki, Hinton, Kiyohiro et Lang, 1989). Les modèles markoviens sont utilisés pour la reconnaissance au niveau mot. Celle-ci consiste en un alignement des séquences de caractères issues du TDNN avec les modèles des mots du lexique (Pittman, 2007).

Un aperçu des techniques existantes pour la reconnaissance de l'écriture vient d'être dressé. Cet aperçu est donné à titre informatif, car il faut rappeler que la reconnaissance de l'écriture n'est pas l'objet d'étude de cette thèse. La reconnaissance est considérée ici comme un outil, qui va permettre l'accès au contenu textuel des documents manuscrits, dont il est possible de faire abstraction. Dans le cadre de ce travail, le choix a été fait de déléguer la tâche de reconnaissance au système de reconnaissance du partenaire industriel. Les principales caractéristiques de ce système, appelé MyScript® Builder, sont décrites dans la section suivante.

2.3 MyScript® Builder

Le moteur de reconnaissance utilisé dans le cadre de ce travail est celui de MyScript® Builder. Il s'agit d'un moteur stable, paramétrable et documenté. Contrairement à la description des approches qui vient d'être faite, la description de MyScript® Builder ne s'intéressera pas à ses principes fondamentaux, mais plus pragmatiquement à la mise en œuvre du processus de reconnaissance. L'objectif sous-jacent est d'analyser l'influence des différentes stratégies de reconnaissance sur les transcriptions générées. Cette section est tirée pour partie de la documentation livrée avec le kit de développement logiciel (SDK).

Le kit de développement logiciel

Dans le cadre de cette étude, c'est la version 4.4 du SDK MyScript® Builder qui a été utilisée. Il s'agit d'une bibliothèque de programmes pour la reconnaissance de l'écriture en-ligne. Celle-ci permet de créer facilement des applications de gestion et de reconnaissance de l'encre électronique. Pour une collection de données, le point de départ de la reconnaissance est l'instanciation du moteur de reconnaissance. Les documents sont lus et représentés selon les structures propres à MyScript® Builder, les résultats de la reconnaissance sont enregistrés dans des fichiers. L'algorigramme d'une application type de reconnaissance par lots est donné en figure 2.6.

L'issue du processus de reconnaissance est déterminée principalement par les ressources linguistiques associées au moteur de reconnaissance lors de son instanciation.

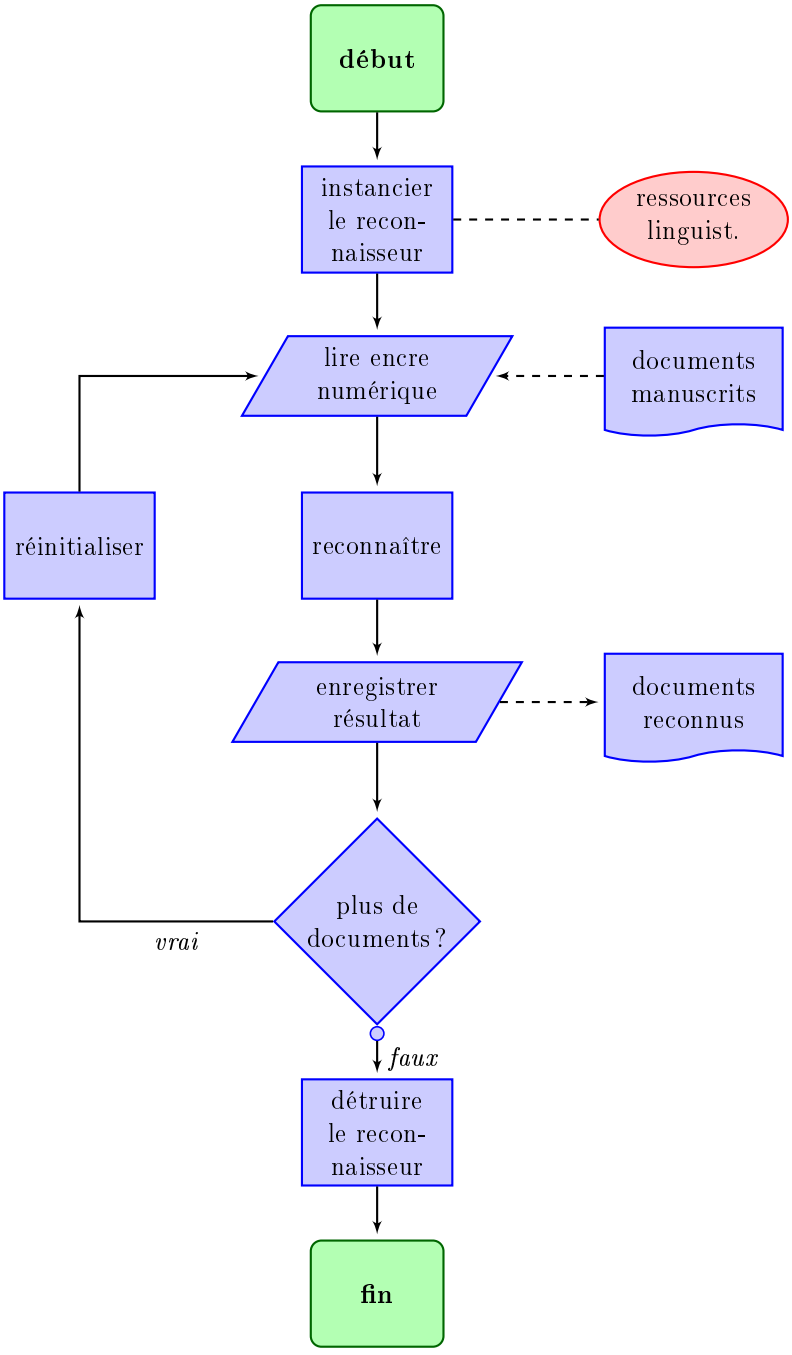


FIGURE 2.6 : Organigramme de programmation d'une application de reconnaissance par lots de documents manuscrits avec MyScript® Builder.

Un choix judicieux des ressources permet d'obtenir une reconnaissance optimale. Les ressources standard utilisées pour la reconnaissance sont énumérées ci-dessous.

Les ressources linguistiques

Les ressources linguistiques, attachables au moteur de reconnaissance, apportent un *a priori* sur la langue, c'est-à-dire sur ce que le moteur est censé reconnaître. Il est possible de définir des ressources spécifiques, sous forme de lexiques ou d'expressions régulières, ou bien d'utiliser les ressources standard livrées avec le SDK.

Reconnaissance dirigée par un lexique Une reconnaissance dirigée par un lexique standard est effectuée grâce à la ressource *lk-standard_ lexicon* (abrégé en *lk_ slex* dans la suite de ce document). Elle ajoute une contrainte lexicale sur le résultat de la reconnaissance. Elle est couplée à un ensemble d'expressions régulières permettant de reconnaître des éléments hors lexique comme les dates, les nombres ou les codes postaux.

Reconnaissance contrainte par un modèle de langage La ressource standard *lk-text* en plus d'être contrainte par le lexique standard, ajoute une contrainte supplémentaire au niveau de la séquence de mots. Cette ressource peut être assimilée à un modèle linguistique qui rend compte des règles d'enchaînement des mots. C'est l'équivalent des modèles n-grammes ou n-classes fréquemment utilisés pour la reconnaissance de phrases (Perraud, Viard-Gaudin, Morin et Lallican, 2003 ; Quiniou, Anquetil et Carbonnel, 2005 ; Quiniou et Anquetil, 2006).

Reconnaissance sans lexique L'utilisation de la ressource standard *lk-free* équivaut à effectuer une reconnaissance sans lexique. Elle inclut cependant un modèle de langage au niveau caractère. Il favorise la reconnaissance de lettres capitales, lettres minuscules et chiffres entre eux. Ainsi, il est possible de lever des ambigüités selon le contexte des caractères comme le montre la figure 2.7.

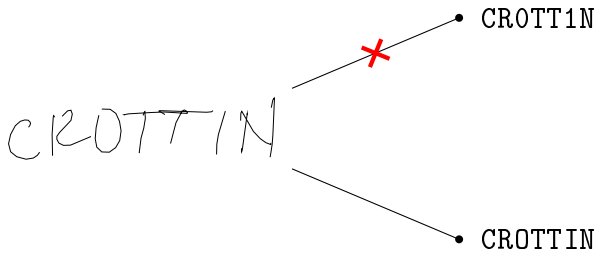


FIGURE 2.7 : Reconnaissance avec `lk-free`. Sans contrainte lexicale le système ne peut différencier entre `0` et `O` ainsi que `1` et `I`. La ressource peut aider à lever l'ambigüité.

2.4 Évaluation de la reconnaissance

Le choix d'une ressource est déterminé par la qualité de la reconnaissance qu'elle produit. Cette qualité est mesurée par un ensemble de métriques permettant de calculer l'écart entre un document de référence et le résultat de la reconnaissance. Ce document de référence est couramment appelé *vérité terrain*.

La comparaison entre les documents reconnus et la vérité terrain s'effectue grâce à des méthodes de programmation dynamique, et plus particulièrement la distance d'édition (Damerau, 1964 ; Levenshtein, 1966) ou la recherche des plus longues sous-séquences communes (Hirschberg, 1977). Ces méthodes procèdent par un alignement des chaînes de caractères en fonction des opérations élémentaires d'insertion (+), suppression (−) et substitution (•) de caractères (*cf.* figure 2.8). Une première mesure à laquelle il est possible de penser est le taux d'erreur au niveau caractère.

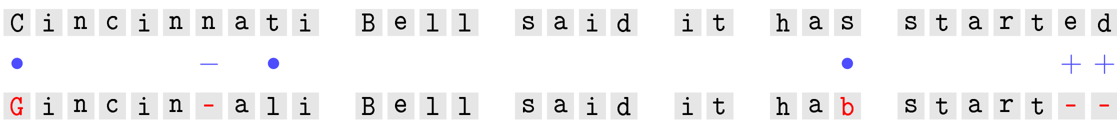


FIGURE 2.8 : Alignement de deux chaînes de caractères, les erreurs de reconnaissance sont signalées en rouge.

Définition 2.1. *Le taux d'erreur au niveau caractère (CER, Character Error Rate) représente le pourcentage de caractères mal reconnus dans le document. Il est obtenu en normalisant le nombre d'erreurs (c_e) par le nombre de caractères de la vérité terrain (c_{ref}).*

$$\mathsf{CER} = \frac{c_e}{c_{ref}} \tag{2.1}$$

À titre d'exemple, dans la figure 2.8, $c_e = 6$ et $c_{ref} = 30$. Le CER est alors égal à 0,2 (20 %).

L'objectif de la reconnaissance de l'écriture est la transcription de séquences de mots. C'est donc sa capacité à bien reconnaître les mots qui doit être évaluée. Or, le CER ne donne aucune idée de cette capacité. Par exemple, un CER = 0,1 peut sembler une très bonne performance. Toutefois, mal reconnaître un caractère sur dix lorsque les mots sont composés en moyenne de dix caractères, équivaut à mal reconnaître pratiquement tous les mots.

Définition 2.2. *Le taux d'erreur au niveau mot (WER, Word Error Rate) représente le pourcentage de mots mal reconnus dans le document. Il est obtenu en normalisant le nombre d'erreurs (w_e) par le nombre de mots du document de référence (w_{ref}).*

$$\mathsf{WER} = \frac{w_e}{w_{ref}} \tag{2.2}$$

Le WER est également calculé grâce à un alignement des séquences, mais cette fois-ci ce sont les mots qui sont alignés et non plus les caractères comme le montre la figure 2.9.

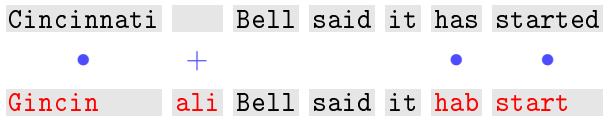


FIGURE 2.9 : Alignement de deux séquences de mots.

Cet exemple montre, paradoxalement, que la longueur de la séquence reconnue peut être supérieure à celle de la séquence attendue. Cela résulte principalement d'erreurs de segmentation du signal en mots. Ainsi, lorsque les insertions sont nombreuses, le taux d'erreur peut être supérieur à 1. Pour cette raison, les insertions ne sont pas comptabilisées dans le calcul du WER. Par exemple, le WER pour la figure 2.9 est égal à $\frac{3}{6}$ (50 %).

Cette façon de calculer le WER produit des incohérences lorsqu'un document reconnu ne contient que des insertions (*cf.* figure 2.10). Cependant, cette situation reste marginale car les erreurs de segmentation se traduisent généralement par une substitution suivie d'une ou plusieurs insertions.

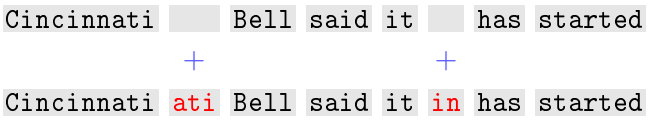


FIGURE 2.10 : Exemple d'alignement avec WER = 0 alors qu'il existe plusieurs erreurs de reconnaissance.

Dans un premier temps, le WER sera la mesure utilisée pour évaluer la reconnaissance du corpus de données manuscrites qui a été collecté pour supporter l'ensemble de cette étude. La présentation de ce corpus particulier fait l'objet du chapitre suivant.

Chapitre 3

Collection de données manuscrites

Il existe de nombreux corpus de RI, issus de multiples campagnes d'évaluation (Voorhees et Harman, 2005), des corpus pour la catégorisation automatique de textes (Hersh, Buckley, Leone et Hickam, 1994 ; Apté, Damerau et Weiss, 1994 ; Lang, 1995 ; Craven, DiPasquo, Freitag, McCallum, Mitchell, Nigam et Slattery, 1998 ; Lewis, Yang, Rose et Li, 2004), ainsi que des corpus pour la reconnaissance de l'écriture en-ligne (Guyon, Schomaker, Plamondon, Liberman et Janet, 1994 ; Viard-Gaudin, Lallican, Knerr et Binter, 1999 ; Liwicki et Bunke, 2005). Toutefois, il n'existe pas à notre connaissance de corpus qui soient à la fois adaptés à la recherche d'information et à la reconnaissance de l'écriture en-ligne.

Or, la problématique de cette étude se situe au carrefour de ces deux disciplines : la collecte de données manuscrites est donc la première étape de ce travail. Le corpus manuscrit utilisé dans toutes les expériences décrites ici est un objet polyvalent qui mérite une description propre. Ce chapitre procède en trois parties. Il commence par expliquer le choix d'un corpus de référence pour la catégorisation de textes (§3.1) ; suivra la description des modalités et résultats de la collecte de données manuscrites (§3.2). Enfin, il s'agira de montrer comment ce corpus, conçu pour la catégorisation de textes, peut être exploité dans des expériences liées à la recherche d'information ad hoc (§3.3).

3.1 Choix du corpus de référence

Le choix du corpus de référence a été déterminé par 3 critères : (1) sa standardisation ; (2) la possibilité d'être utilisé en catégorisation mono-étiquette (voir §4.1) ; (3) une prédisposition naturelle des documents à être reproduits sous forme manuscrite. Ce dernier critère est de loin le plus important.

Les corpus couramment utilisés pour la recherche en catégorisation automatique de textes sont le corpus OHSUMED (Hersh et collab., 1994), WebKB (Craven et collab., 1998), 20 Newsgroups (Lang, 1995), Reuters-21578 (Apté et collab., 1994) et Reuters RCV1-v2 (Lewis et collab., 2004).

Le corpus OHSUMED est un corpus médical extrait de PubMed. Il couvre toutes les références de 270 journaux indexés entre 1987-1991. Il s'agit d'un corpus très spécialisé dont les documents ne se prêtent pas pour une reproduction manuscrite (*cf.* figure 3.1(a)).

Les données du corpus WebKB sont un ensemble de pages web de grandes universités américaines (*cf.* figure 3.1(c)). Les pages sont organisées en six catégories, chaque catégorie est divisée à son tour en un ensemble de sous-catégories correspondant à chaque institution. La nature même du corpus est rédhibitoire pour une collecte de données manuscrites.

Lang (1995) a collecté un corpus issu de 20 groupes de discussion (*cf.* figure 3.1(b)). Les différentes études faisant référence à ce corpus ne l'utilisent pas de manière homogène. De plus, les forums de discussion ne correspondent pas à une situation où l'utilisation de l'écriture se justifie.

Les deux corpus Reuters sont les seuls de nature et taille aptes à être reproduits sous forme manuscrite. Ils peuvent correspondre à des situations réelles comme la prise de notes ou la rédaction en situation de mobilité. Les caractéristiques du corpus RCV1-v2, récemment proposé par Lewis et collab. (2004), le rendent apte avant tout pour la catégorisation hiérarchique. Un exemple tiré de ce corpus est donné par la figure 3.1(d). La nature hiérarchique des catégories peut être observée grâce aux préfixes des codes des catégories. Il faut également noter que la licence d'utilisation très contraignante de ce corpus a empêché sa large diffusion (Debole et Sebastiani, 2005).

Suite à l'étude des caractéristiques des différents corpus, notre choix s'est naturellement porté sur le corpus Reuters-21578. Un aperçu de son histoire et la description de ses caractéristiques principales sont donnés ci-dessous.

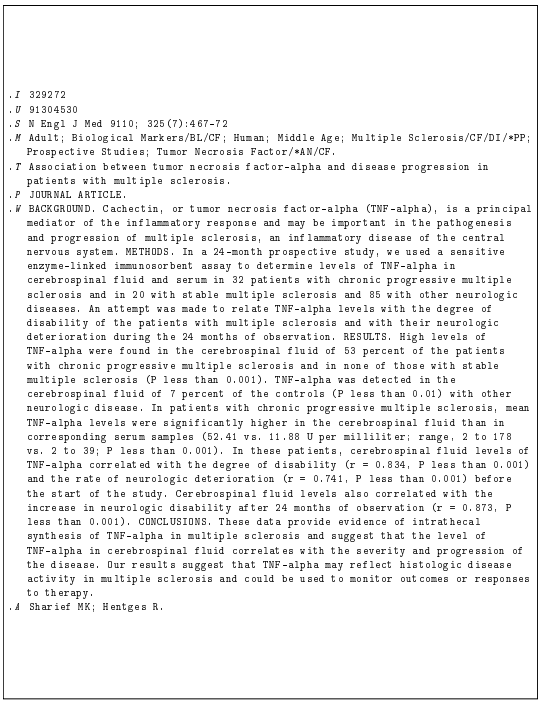
3.1.1 Historique

Le corpus Reuters est l'un des corpus les plus utilisés dans la littérature scientifique pour l'évaluation de méthodes de catégorisation. Il contient plus de 20 000 dépêches publiées à la fin des années 80. La première version du corpus (Reuters-22173) a été proposée par Hayes et Weinstein (1990) pour l'évaluation du système de catégorisation Construe. Il contenait 21 450 documents pour l'entraînement et 723 pour le test.

Lewis et Ringuette (1994) proposent une deuxième version (Reuters-21450) où les documents sont répartis en deux ensembles chronologiquement cohérents : les dépêches les plus anciennes sont utilisées pour l'entraînement et les récentes pour le test. Le corpus Reuters-21450 contenait beaucoup de documents sans catégorie qui avaient un impact négatif sur les performances des méthodes de catégorisation.

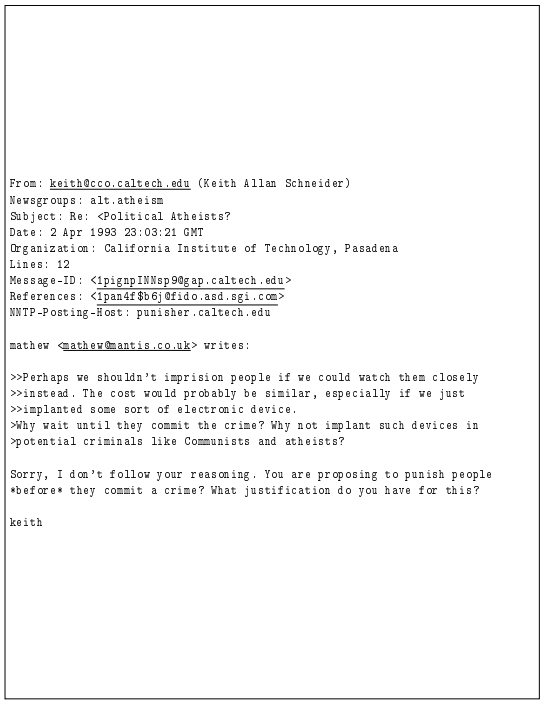
Une dernière version a été mise au point pour l'évaluation du système de catégorisation SWAP-1 (Apté et collab., 1994). Tous les documents ne possédant pas de catégorie ont été supprimés des ensembles d'entraînement et de test. De même, les catégories possédant moins de deux documents d'entraînement ont été écartées.

Afin de standardiser le corpus et de rendre les études comparables, David Lewis et Steve Finch ont entrepris une vérification générale du corpus Reuters-22173 qui a permis de corriger diverses erreurs typographiques et d'étiquetage ainsi que d'éliminer les doublons.



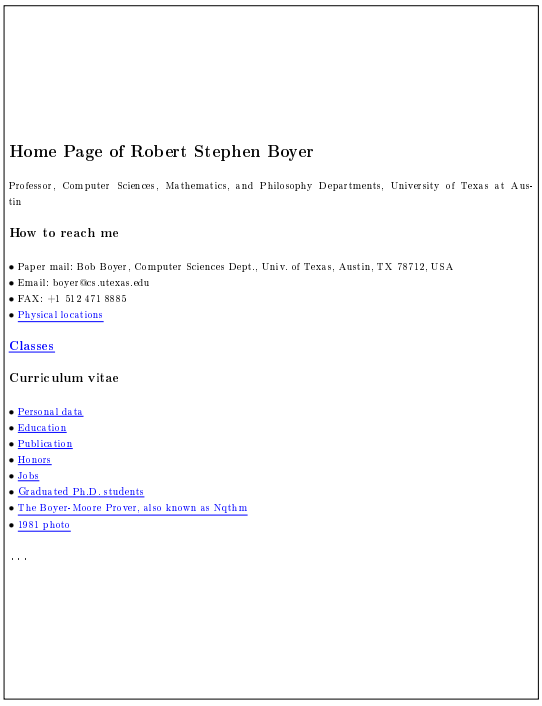
```
.I 329272
.U 91304530
.S N Engl J Med 9110; 325(7):467-72
.M Adult; Biological Markers/BL/CF; Human; Middle Age; Multiple Sclerosis/CF/DI/*PP;
   Prospective Studies; Tumor Necrosis Factor/*AN/CF.
.T Association between tumor necrosis factor-alpha and disease progression in
   patients with multiple sclerosis.
.P JOURNAL ARTICLE.
.W BACKGROUND. Cachectin, or tumor necrosis factor-alpha (TNF-alpha), is a principal
   mediator of the inflammatory response and may be important in the pathogenesis
   and progression of multiple sclerosis, an inflammatory disease of the central
   nervous system. METHODS. In a 24-month prospective study, we used a sensitive
   enzyme-linked immunosorbent assay to determine levels of TNF-alpha in
   cerebrospinal fluid and serum in 32 patients with chronic progressive multiple
   sclerosis and in 20 with stable multiple sclerosis and 85 with other neurologic
   diseases. An attempt was made to relate TNF-alpha levels with the degree of
   disability of the patients with multiple sclerosis and with their neurologic
   deterioration during the 24 months of observation. RESULTS. High levels of
   TNF-alpha were found in the cerebrospinal fluid of 53 percent of the patients
   with chronic progressive multiple sclerosis and in none of those with stable
   multiple sclerosis (P less than 0.001). TNF-alpha was detected in the
   cerebrospinal fluid of 7 percent of the controls (P less than 0.01) with other
   neurologic disease. In patients with chronic progressive multiple sclerosis, mean
   TNF-alpha levels were significantly higher in the cerebrospinal fluid than in
   corresponding serum samples (52.41 vs. 11.88 U per milliliter; range, 2 to 178
   vs. 2 to 39; P less than 0.001). In these patients, cerebrospinal fluid levels of
   TNF-alpha correlated with the degree of disability (r = 0.834, P less than 0.001)
   and the rate of neurologic deterioration (r = 0.741, P less than 0.001) before
   the start of the study. Cerebrospinal fluid levels also correlated with the
   increase in neurologic disability after 24 months of observation (r = 0.873, P
   less than 0.001). CONCLUSIONS. These data provide evidence of intrathecal
   synthesis of TNF-alpha in multiple sclerosis and suggest that the level of
   TNF-alpha in cerebrospinal fluid correlates with the severity and progression of
   the disease. Our results suggest that TNF-alpha may reflect histologic disease
   activity in multiple sclerosis and could be used to monitor outcomes or responses
   to therapy.
.A Sharief MK; Hentges R.
```

(a) Ohsumed (Adult, etc.)



```
From: keith@cco.caltech.edu (Keith Allan Schneider)
Newsgroups: alt.atheism
Subject: Re: <Political Atheists?
Date: 2 Apr 1993 23:03:21 GMT
Organization: California Institute of Technology, Pasadena
Lines: 12
Message-ID: <1pignpINNsp9@gap.caltech.edu>
References: <1pan4f$b6j@fido.asd.sgi.com>
NNTP-Posting-Host: punisher.caltech.edu

mathew <mathew@mantis.co.uk> writes:

>>Perhaps we shouldn't imprision people if we could watch them closely
>>instead. The cost would probably be similar, especially if we just
>>implanted some sort of electronic device.
>Why wait until they commit the crime? Why not implant such devices in
>potential criminals like Communists and atheists?

Sorry, I don't follow your reasoning. You are proposing to punish people
*before* they commit a crime? What justification do you have for this?

keith
```

(b) 20 NG (alt.atheism)



Home Page of Robert Stephen Boyer

Professor, Computer Sciences, Mathematics, and Philosophy Departments, University of Texas at Austin

How to reach me

- Paper mail: Bob Boyer, Computer Sciences Dept., Univ. of Texas, Austin, TX 78712, USA
- Email: boyer@cs.utexas.edu
- FAX: +1 512 471 8885
- Physical locations

Classes

Curriculum vitae

- Personal data
- Education
- Publication
- Honors
- Jobs
- Graduated Ph.D. students
- The Boyer-Moore Prover, also known as Nqthm
- 1981 photo

...

(c) WebKB (faculty/texas)

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<newsitem itemid="2330" id="root" date="1996-08-20" xml:lang="en">
<title>USA: Tylan stock jumps; weighs sale of company.</title>
<headline>Tylan stock jumps; weighs sale of company.</headline>
<dateline>SAN DIEGO</dateline>
<text>
<p>The stock of Tylan General Inc. jumped Tuesday after the maker of
process-management equipment said it is exploring the sale of the
company and added that it has already received some inquiries from
potential buyers.</p>
<p>Tylan was up $2.50 to $12.75 in early trading on the Nasdaq market.</p>
<p>The company said it has set up a committee of directors to oversee
the sale and that Goldman, Sachs &amp;amp; Co. has been retained as its
financial adviser.</p>
</text>
<copyright>(c) Reuters Limited 1996</copyright>
<metadata>
<codes class="bip:countries:1.0">
<code code="USA"> </code>
</codes>
<codes class="bip:industries:1.0">
<code code="I3442O"> </code>
</codes>
<codes class="bip:topics:1.0">
<code code="C15"> </code>
<code code="C152"> </code>
<code code="C18"> </code>
<code code="C181"> </code>
<code code="CCAT"> </code>
</codes>
</metadata>
</newsitem>
```

(d) RCV1-v2 (C15, etc.)

FIGURE 3.1 : Exemples de documents issus de différents corpus utilisés pour la recherche en catégorisation automatique de textes. Les catégories sont signalées entre parenthèses.

3.1.2 Caractéristiques

Le corpus Reuters-21578 est un ensemble de 21 578 dépêches de l'agence Reuters® parues en 1987. Parmi les 21 578 dépêches, seulement 12 902 sont utilisables, car les documents restant (8 676 documents) n'ont pas fait l'objet d'une catégorisation par les employés de l'agence.

Les dépêches sont distribuées en 135 catégories thématiques relatives à l'économie. Les documents du corpus abordent divers sujets comme les fusions-acquisitions d'entreprises, les marchés de matières premières (céréales, sucre, etc.), le cours du pétrole et ses dérivés, les marchés de changes et du taux d'intérêt, etc. Ces documents se caractérisent également par une forte présence d'abréviations propres au domaine économique ainsi que d'entités nommées : lieux, personnalités et noms d'entreprises ou organisations gouvernementales. Les documents de certaines catégories, en particulier la classe *earn*, sont rédigés dans un style télégraphique avec surabondance d'entités numériques.

Bien que le corpus se prête à une utilisation multi-étiquette (voir §4.1), environ 64 % des documents ne sont affectés qu'à une seule catégorie.

Il s'agit d'une collection déséquilibrée, dans le sens où certaines catégories ont très peu de documents d'entraînement, alors que d'autres en ont des milliers (Debole et Sebastiani, 2005). Par ailleurs, les 10 catégories les plus représentées comptent pour 90% des effectifs du corpus. Le sous-ensemble du corpus basé sur ces 10 catégories est couramment appelé $R(10)$ (Debole et Sebastiani, 2005).

Parmi les 135 catégories, seulement 90 sont représentées aussi bien dans l'ensemble d'entraînement que dans celui de test. Le sous-ensemble du corpus basé sur ces 90 catégories est couramment appelé $R(90)$ (Debole et Sebastiani, 2005). La communauté scientifique a adopté unanimement le partitionnement en ensembles d'entraînement et test proposé par Apté et collab. (1994).

3.2 Collecte de données

Étant donné le nombre important de textes dans le corpus, un nombre réduit de documents a dû être choisi. Par conséquent, la collecte de données manuscrites se base sur la version $R(10)$ du corpus ce qui réduit considérablement le nombre de documents.

Les documents à reproduire ont été choisis pour une utilisation mono-étiquette du corpus (voir §4.1). De plus, les documents contenant plus de 120 mots ont été écartés afin de rendre plus aisée la recopie de la dépêche pour les contributeurs. Les caractéristiques de la collecte et ses résultats sont présentés dans les sous-sections suivantes.

3.2.1 Méthodologie et matériel

L'objectif affiché de la collecte était 2 000 documents d'entraînement et 500 documents de test. La collecte a été effectuée à l'aide de formulaires sur papier Anoto® (*cf.* figure 3.2) et de stylos numériques Nokia® SU-1B. La résolution spatiale obtenue avec ces dispositifs est de 677 points par pouce (dpi) et la fréquence d'échantillonnage est de 100 points par seconde (100Hz).

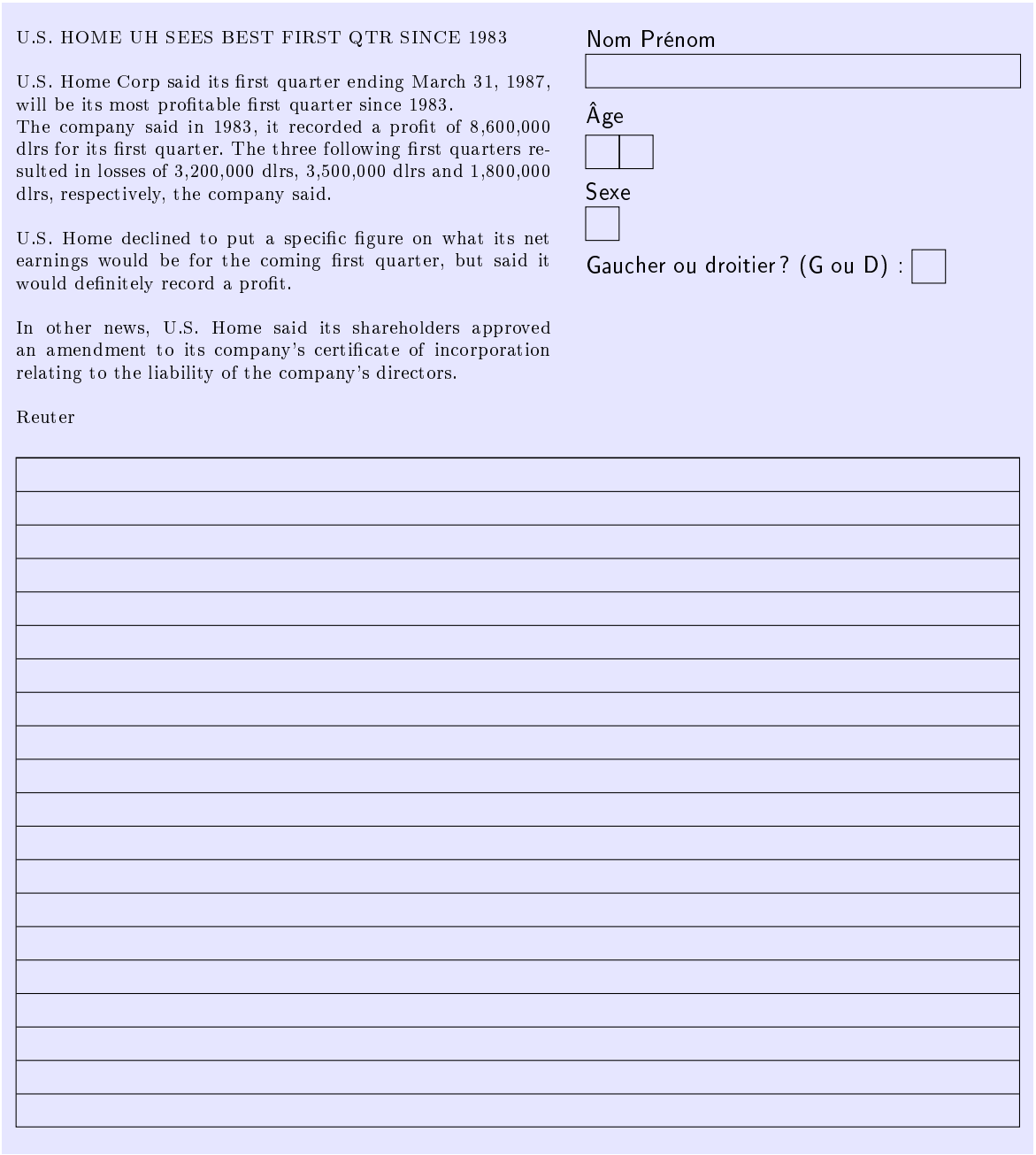


FIGURE 3.2 : Formulaire de collecte. En haut à gauche la dépêche à recopier. À droite les informations relatives au scripteur. En bas l'espace pour le texte manuscrit.

En plus des informations liées aux documents, le formulaire propose aux volontaires de renseigner leur nom, prénom, âge, sexe et la main utilisée pour écrire (*cf.* figure 3.2). Ces informations ont servi à des études sur la reconnaissance du scripteur (Tan, Viard-Gaudin et Kot, 2009b,a).

3.2.2 Résultats de la collecte

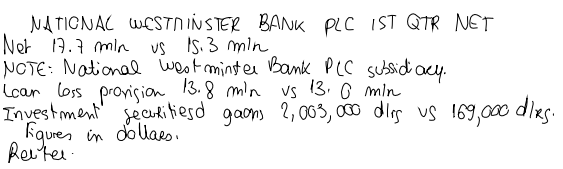
La collecte de données a mobilisé environ 1 500 scripteurs distincts pour un peu plus de 2 400 formulaires collectés. Après la phase de tri et d'anonymisation des formulaires, le texte électronique original ainsi que la catégorie ont été associés aux documents manuscrits. Le tri de documents a révélé un peu plus de 90 formulaires inutilisables, cela peut être dû à une mauvaise utilisation (inclinaison, pression) ou à une défaillance du stylo lors de l'écriture. Au total, 2 310 documents ont été utilisés dans toutes les expériences rapportées ici.

La distribution de ces 2 310 documents par catégorie et selon les ensembles d'entraînement et de test est donnée par le tableau 3.1. Le nom des catégories est donné en français, l'appellation habituellement utilisée dans la littérature scientifique est donnée entre parenthèses.

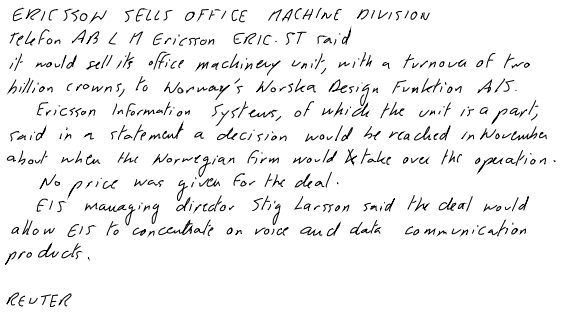
TABLE 3.1 : Distribution des documents manuscrits par catégorie et selon les ensembles d'entraînement et de test.

Catégorie	Entraînement	Test
Dividendes (earn)	734	135
Fusions-acquisitions (acq)	398	82
Céréales (grain)	139	53
Marché des changes (money-fx)	205	54
Pétrole (crude)	85	46
Taux d'intérêt (interest)	87	30
Commerce international (trade)	66	24
Fret maritime (ship)	62	18
Sucre (sugar)	37	11
Café (coffee)	34	10
Total	1 847	463

La figure 3.3 montre des exemples de documents manuscrits tirés du corpus collecté. Les documents sont composés d'un titre, la plupart du temps en majuscules, du corps de la dépêche et de la signature de l'agence Reuters®. Il est possible de remarquer à travers les différents exemples, la grande variété des styles d'écriture présents dans la base.

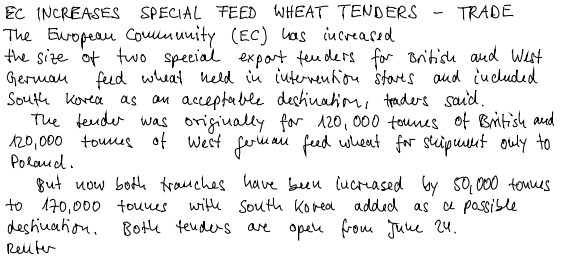

NATIONAL WESTMINSTER BANK PLC 1ST QTR NET
Net 17.7 mln vs 15.3 mln
NOTE: National Westminster Bank PLC subsidiary.
Loan loss provision 13.8 mln vs 13.0 mln
Investment securitiesd gains 2,003,000 dlrs vs 169,000 dlrs.
Figures in dollars.
Reuter.

(a) earn

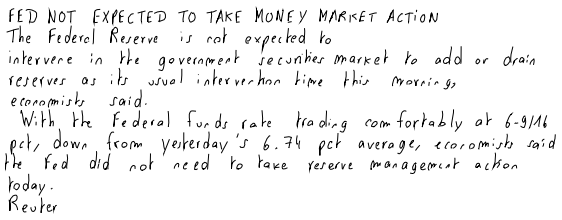

ERICSSON SELLS OFFICE MACHINE DIVISION
Telefon AB L M Ericsson ERIC.ST said
it would sell its office machinery unit, with a turnova of two billion crowns, to Norway's Norska Design Funktion A/S.
Ericsson Information Systems, of which the unit is a part, said in a statement a decision would be reached in November about when the Norwegian firm would take over the operation.
No price was given for the deal.
EIS managing director Stig Larsson said the deal would allow EIS to concentrate on voice and data communication products.

REUTER

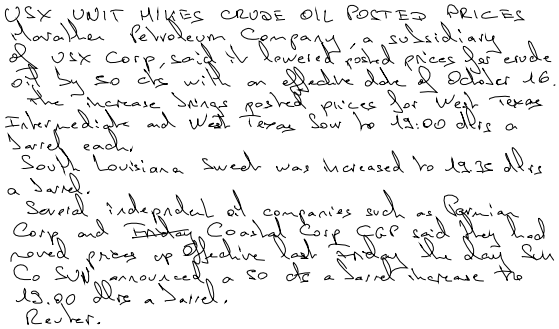
(b) acq


EC INCREASES SPECIAL FEED WHEAT TENDERS - TRADE
The European Community (EC) has increased
the size of two special export tenders for British and West German feed wheat held in intervention stores and included South Korea as an acceptable destination, traders said.
The tender was originally for 120,000 tonnes of British and 120,000 tonnes of West german feed wheat for shipment only to Poland.
But now both tranches have been increased by 50,000 tonnes to 170,000 tonnes with South Korea added as a possible destination. Both tenders are open from June 24.
Reuter

(c) grain

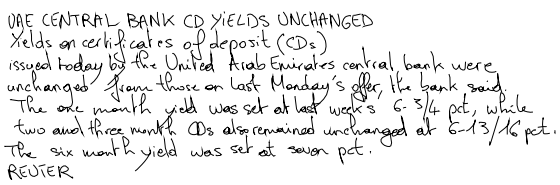

FED NOT EXPECTED TO TAKE MONEY MARKET ACTION
The Federal Reserve is not expected to
intervene in the government securities market to add or drain reserves as its usual intervention time this morning, economists said.
With the Federal funds rate trading comfortably at 6-9/16 pct, down from yesterday's 6.74 pct average, economists said the Fed did not need to take reserve management action today.
Reuter

(d) money-fx

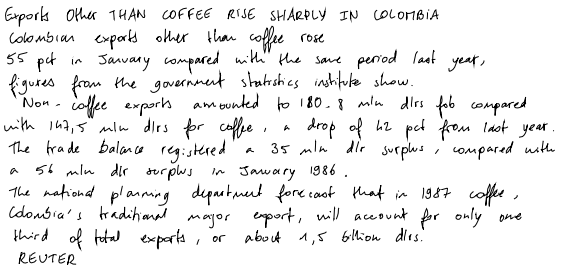

USX UNIT HIKES CRUDE OIL POSTED PRICES
Marathon Petroleum Company, a subsidiary of USX Corp, said it lowered posted prices for crude oil by 50 cts with an effective date of October 16.
The increase brings posted prices for West Texas Intermediate and West Texas Sour to 13.00 dlrs a barrel each.
South Louisiana Sweet was increased to 13.35 dlrs a barrel.
Several independent companies such as Permian Corp and Basin Coastal Corp CGP said they had moved prices up effective last Friday the day Sun Co SUN announced a 50 cts a barrel increase to 13.90 dlrs a barrel.
Reuter.

(e) crude

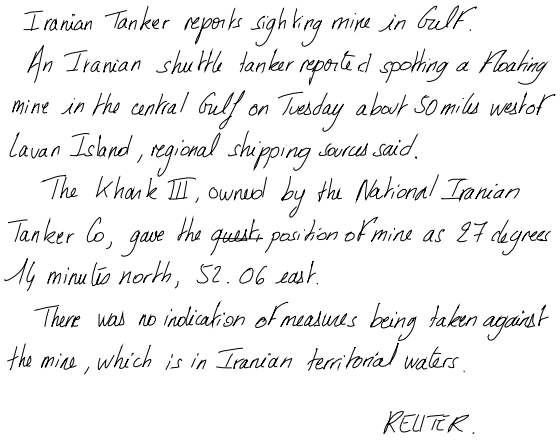
FIGURE 3.3 : Échantillons de documents manuscrits pour chacune des catégories de la base collectée.


UAE CENTRAL BANK CD YIELDS UNCHANGED
Yields on certificates of deposit (CDs)
issued today by the United Arab Emirates central bank were
unchanged from those on last Monday's offer, the bank said.
The one month yield was set at last week's 6-3/4 pct, while
two and three month CDs also remained unchanged at 6-13/16 pct.
The six month yield was set at seven pct.
REUTER

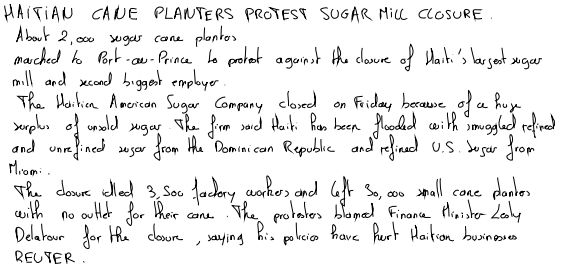
(f) interest


Exports Other THAN COFFEE RISE SHARPLY IN COLOMBIA
Colombian exports other than coffee rose
55 pct in January compared with the same period last year,
figures from the government statistics institute show.
Non-coffee exports amounted to 180.8 mln dlrs fob compared
with 143,5 mln dlrs for coffee, a drop of 42 pct from last year.
The trade balance registered a 35 mln dlr surplus, compared with
a 56 mln dlr surplus in January 1986.
The national planning department forecast that in 1987 coffee,
Colombia's traditional major export, will account for only one
third of total exports, or about 1,5 billion dlrs.
REUTER

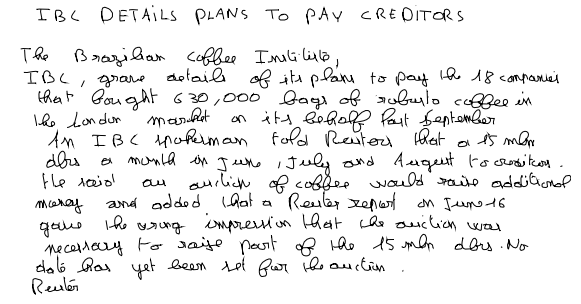
(g) trade


Iranian Tanker reports sighting mine in Gulf.
An Iranian shuttle tanker reported spotting a floating
mine in the central Gulf on Tuesday about 50 miles west of
Lavan Island, regional shipping sources said.
The Khark III, owned by the National Iranian
Tanker Co, gave the ~~quest~~ position of mine as 27 degrees
14 minutes north, 52.06 east.
There was no indication of measures being taken against
the mine, which is in Iranian territorial waters.
REUTER.

(h) ship


HAITIAN CANE PLANTERS PROTEST SUGAR MILL CLOSURE.
About 2,000 sugar cane planters
marched to Port-au-Prince to protest against the closure of Haiti's largest sugar
mill and second biggest employer.
The Haitian American Sugar Company closed on Friday because of a huge
surplus of unsold sugar. The firm said Haiti has been flooded with smuggled refined
and unrefined sugar from the Dominican Republic and refined U.S. sugar from
Miami.
The closure idled 3,500 factory workers and left 30,000 small cane planters
with no outlet for their cane. The protestors blamed Finance Minister Leslie
Delatour for the closure, saying his policies have hurt Haitian businesses.
REUTER.

(i) sugar


IBC DETAILS PLANS TO PAY CREDITORS
The Brazilian coffee Institute,
IBC, gave details of its plans to pay the 18 companies
that bought 630,000 bags of robusta coffee in
the London market on its behalf last September
An IBC spokesman told Reuters that a 15 mln
dlrs a month in June, July and August to creditors.
He said an auction of coffee would raise additional
money and added that a Reuter report on June 16
gave the wrong impression that the auction was
necessary to raise part of the 15 mln dlrs. No
date has yet been set for the auction.
Reuter

(j) coffee

FIGURE 3.3 : Échantillons de documents manuscrits pour chacune des catégories de la base collectée (suite et fin).

3.3 Corpus pour la RI

Les travaux rapportés dans cette thèse sont liés aussi bien à la catégorisation qu'à la recherche d'information ad-hoc (voir chapitre 6). Le corpus de données manuscrites collecté est un corpus pour la catégorisation, de ce fait il n'est pas adapté pour une tâche de recherche d'information. La recherche d'information nécessite un corpus composé d'un ensemble de requêtes, d'une collection de documents et de la liste de documents pertinents pour chacune des requêtes. D'un autre côté, la catégorisation nécessite un corpus composé d'un ensemble de catégories, d'une collection de documents et de la liste de documents pertinents pour chacune des catégories.

Il apparaît entre les concepts de requête et catégorie une certaine analogie. Il s'agit ici d'adapter automatiquement le corpus collecté pour la tâche de la RI en tirant parti de cette analogie.

Puisque les catégories sont assignées aux documents par un expert humain, elles peuvent être assimilées à des jugements de pertinence (voir §6.1). En se basant sur les catégories associées aux documents, des requêtes prototypiques de chaque catégorie peuvent être générées grâce à des techniques de contrôle de pertinence (relevance feedback). Suivant ce constat, Sanderson (1994) a réalisé des expériences de RI avec le corpus Reuters-21578. L'approche que nous proposons reprend les principes des expériences de Sanderson (1994). Il s'agit d'une approche en deux temps. La première étape consiste à sélectionner les termes pertinents pour chaque catégorie (§3.3.1). La deuxième étape consiste à générer les requêtes proprement dites (§3.3.2).

3.3.1 Sélection de termes

La génération de requêtes utilise les documents électroniques correspondant aux documents collectés. Les documents ont été divisés aléatoirement en deux sous-ensembles de taille égale :

- le sous-ensemble $\mathcal{Q}$ est utilisé pour générer les requêtes (1 157 documents) ;
- le sous-ensemble $\mathcal{T}$ est utilisé pour la recherche d'information (1 153 documents).

La formule de base du modèle d'indépendance binaire proposé par Robertson et Spärck Jones (1976) a été adaptée, afin de classer les *termes* par ordre de pertinence vis-à-vis de la catégorie (voir §4.2 pour une définition de la notion de terme).

En considérant $p_{t,c}$ comme la probabilité que t soit pertinent par rapport à la catégorie c, et $q_{t,c}$ la probabilité que t ne le soit pas, l'indice de pertinence d'un terme peut être calculé grâce au *log* du rapport de chances entre ces deux probabilités.

$$score_{t,c} = \log \frac{p_{t,c} \times (1 - q_{t,c})}{(1 - p_{t,c}) \times q_{t,c}} \tag{3.1}$$

En pratique les probabilités $p_{t,c}$ et $q_{t,c}$ sont estimées à partir du tableau 3.2. Un lissage des valeurs de la table de contingence (Salton et Buckley, 1990) est souvent effectué afin d'éviter les valeurs où soit le *log*, soit le rapport n'est pas défini. Les formules 3.2 et 3.3 donnent le moyen utilisé pour calculer ces probabilités dans le cadre de ce travail.

TABLE 3.2 : Tableau de contingence pour t et c (en nombre de documents).

	Doc. contenant t	Doc. ne contenant pas t
Doc. de c	A	B
Doc. pas de c	C	D

$$p_{t,c} = \frac{(A + 0{,}5)}{((A + B) + 1{,}0)} \tag{3.2}$$

$$q_{t,c} = \frac{(C + 0{,}5)}{((C + D) + 1{,}0)} \tag{3.3}$$

Les 100 termes les plus pertinents, au sens de la formule 3.1, forment l'ensemble de termes pour la génération des requêtes.

3.3.2 Génération de requêtes

La requête prototypique $\mathbf{q}_c$ d'une catégorie c donnée est calculée en soustrayant les fréquences d'apparitions des 100 termes choisis précédemment dans les documents de c et dans les autres. Cette idée se concrétise par la formule de contrôle de pertinence Ide-dec-hi (Ide, 1971) :

$$\mathbf{q}_c = \sum_{a=1}^{(A+B)} \mathbf{c}_a - \sum_{b=1}^{(A+B)} \overline{\mathbf{c}}_b \tag{3.4}$$

Dans cette formule, $\mathbf{c}_a$ est le vecteur des fréquences des termes dans le document a de catégorie c, et $\overline{\mathbf{c}}_b$ est le vecteur des fréquences des termes dans le document b n'appartenant pas à la catégorie c. Il est important de remarquer que tous les documents de

TABLE 3.3 : Requêtes générées pour les 10 catégories représentées dans le corpus manuscrit.

Catégorie	Requête
earn	vs ct net shr loss
acq	acquir stake acquisit complet merger
grain	tonn wheat grain corn agricultur
money-fx	stg monei dollar band bill
crude	oil crude barrel post well
interest	rate prime lend citibank percentag
trade	surplu deficit narrow trade tariff
ship	port strike vessel hr worker
sugar	sugar raw beet cargo kain
coffee	coffe bag ico registr ibc

la catégorie c interviennent dans le calcul, alors que seulement une partie des documents n'appartenant pas à c intervient. Lorsque le nombre de documents n'appartenant pas à c est supérieur à au nombre de document de la catégorie c, un sous-ensemble constitué du même nombre de documents est choisi au hasard. Dans le cas contraire, tous les documents n'appartenant pas à c sont utilisés. Une requête est composée des 5 termes ayant la plus grande fréquence dans le vecteur $\mathbf{q}_c$ (Sanderson, 1994).

Les requêtes générées pour chacune des catégories du corpus sont données dans le tableau 3.3. Lorsqu'une recherche sera effectuée, si la catégorie d'un document retrouvé correspond à la catégorie de la requête, alors ce document sera considéré comme pertinent. Ainsi, les performances des différents systèmes de recherche d'information peuvent être mesurées (§6.3). Lors de l'étape de recherche, les documents du corpus manuscrit correspondant à l'ensemble $\mathcal{T}$ sont utilisés.

Il est possible de remarquer que le nom de la catégorie est présent en tant que terme dans six des 10 requêtes générées. La plupart des termes générés peuvent être observés dans les différents échantillons du corpus donnés en figure 3.3.

Même si les requêtes semblent représentatives des différentes catégories d'un point de vue purement lexical, il est difficile d'estimer dans quelle mesure elles ont un sens d'un point de vue humain. Dans le cadre des expériences liées à cette thèse, cet aspect est négligé car il importe avant tout qu'un même ensemble de requêtes soit soumis aux différentes méthodes de recherche d'information. De plus, selon Sanderson (1994), le processus de génération de requêtes peut être assimilé à la première itération d'un processus de raffinement d'une recherche d'information dans les systèmes avec contrôle de pertinence.

3.4 Reconnaissance du corpus

Cette section présente les résultats de la reconnaissance du corpus selon les différentes ressources présentées en §2.3. La reconnaissance est évaluée avec le WER, c'est-à-dire selon la capacité du système à bien reconnaître les mots. En plus du taux d'erreur moyen par document, le tableau 3.4 fournit différentes mesures descriptives. Ces mesures sont obtenues à partir des ensembles d'entraînement et de test. Elles permettent de s'assurer, d'une part, que la moyenne est représentative de la tendance centrale et, d'autre part, qu'elle ne minore pas une différence de tendance entre les ensembles d'entraînement et de test.

TABLE 3.4 : Taux d'erreur au niveau mot (en pourcentage).

(a) Jeu d'entraînement

	`lk-free`	`lk-slex`	`lk-text`
Moyenne	51,62 %	25,18 %	20,90 %
Écart-type	17,41 %	12,15 %	12,82 %
Médiane	51,28 %	23,58 %	17,86 %
1er quartile	39,58 %	16,18 %	10,58 %
3e quartile	64,17 %	32,63 %	30,14 %

(b) Jeu de test

	`lk-free`	`lk-slex`	`lk-text`
Moyenne	52,02 %	25,41 %	21,15 %
Écart-type	17,34 %	12,53 %	13,50 %
Médiane	51,79 %	24,18 %	17,07 %
1er quartile	39,53 %	15,29 %	10,10 %
3e quartile	64,71 %	33,33 %	32,26 %

Le tableau 3.4 montre qu'il n'y a pas de différence significative du taux d'erreur entre les ensembles d'entraînement et de test. Cette absence de différence prouve qu'il n'y a pas de biais introduit par la sélection des ensembles, et que les ressources se comportent de manière cohérente sur les deux ensembles.

Une première remarque concernant le WER est que la ressource `lk-free` obtient des performances nettement inférieures aux deux autres ressources. Les performances des approches sans lexique ne sont pas représentatives des systèmes actuels. Cependant, elles permettent d'étudier le comportement de la catégorisation et de la recherche d'information avec des documents fortement dégradés.

Les deux autres ressources, `lk-slex` et `lk-text`, plus performantes, montrent des résultats assez proches. Cependant, la contrainte imposée par le modèle de langage permet de réduire le WER d'environ 5 % par rapport à la stratégie dirigée seulement par le lexique standard. Sachant que ni le lexique, ni le modèle de langage n'intègrent de connaissances spécifiques au corpus collecté, les transcriptions qui en résultent peuvent être considérées comme étant de bonne qualité.

Les valeurs de la moyenne et la médiane pour `lk-free` montrent que la distribution du taux d'erreur est symétrique. Pour `lk-slex`, bien que ses deux mesures de tendance centrale soient proches, la distribution est asymétrique (*cf.* figure 3.4). La différence importante entre la médiane et la moyenne de `lk-text` reflète une asymétrie de la distribution encore plus prononcée.

Paradoxalement, c'est pour les ressources obtenant les meilleures transcriptions que

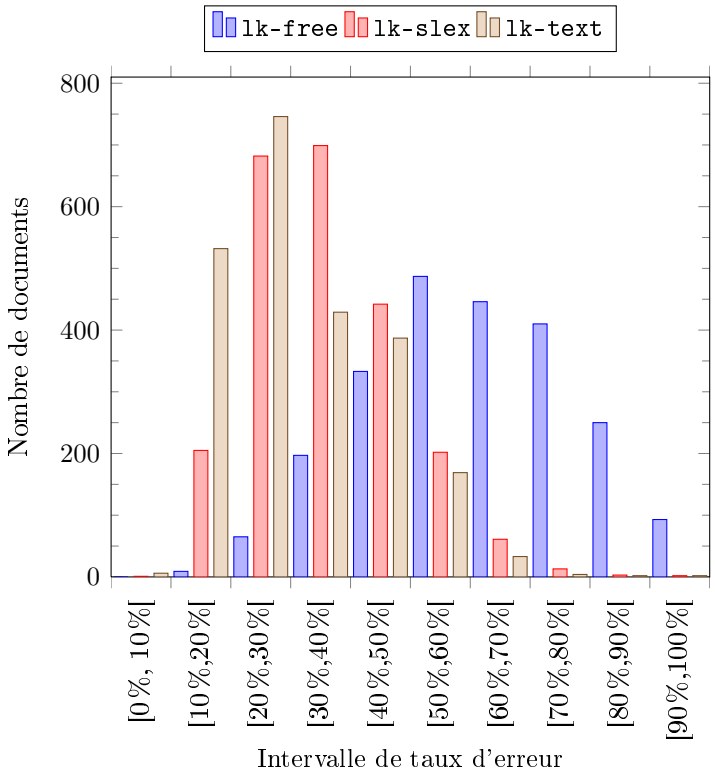
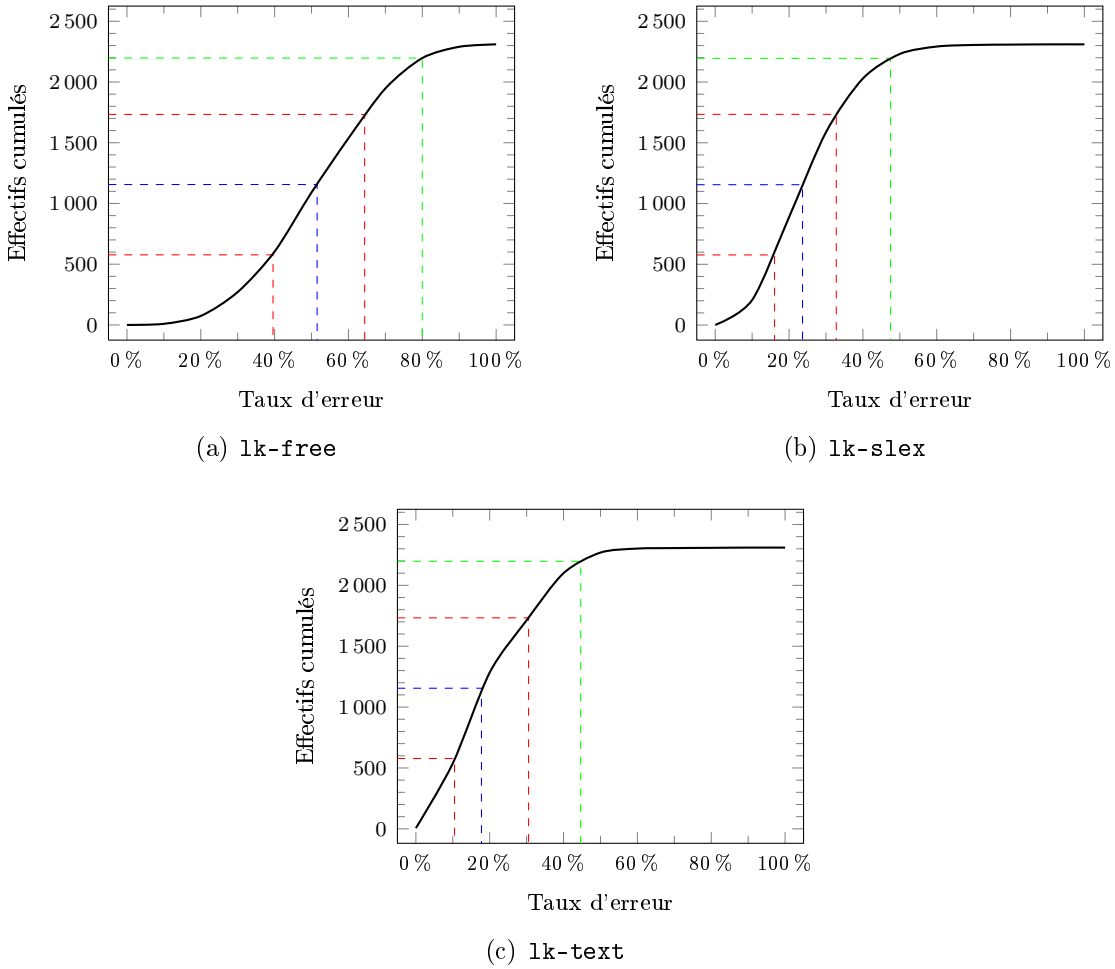


FIGURE 3.4 : Nombre de documents par intervalles du WER.

les données sont les plus hétérogènes. Dans le cas de `lk-slex`, les données s'étendent à droite de l'intervalle modal ([30 %, 40 %[). Pour la ressource `lk-text`, environ 32 % des effectifs sont concentrés dans l'intervalle modal ([20 %, 30 %[). Dans ces conditions, une analyse basée sur la valeur moyenne du WER sera faussée car elle n'est pas représentative des données.

En réalité, le taux d'erreur est difficile à résumer en une seule valeur. Cependant, dans une perspective d'analyse globale, c'est-à-dire en prenant les documents dans leur ensemble, la médiane est un meilleur indicateur pour illustrer l'ordre de grandeur du WER dans le corpus, car elle est représentative de l'intervalle modal des trois distributions (*cf.* figure 3.4).

Afin d'examiner la relation entre la qualité de la reconnaissance et les performances de la CAT ou la RI, il serait préférable de se référer aux quantiles lors de l'analyse individuelle des documents. L'intérêt des quantiles est de regrouper les effectifs par paliers. Le diagramme des effectifs cumulés (*cf.* figure 3.5) permet de visualiser quelques-uns de ces paliers.



(a) `lk-free` (b) `lk-slex`

(c) `lk-text`

FIGURE 3.5 : Diagrammes des effectifs cumulés. Les 1er et 3e quartiles sont représentés en rouge. La médiane en bleu et le 95e percentile en vert.

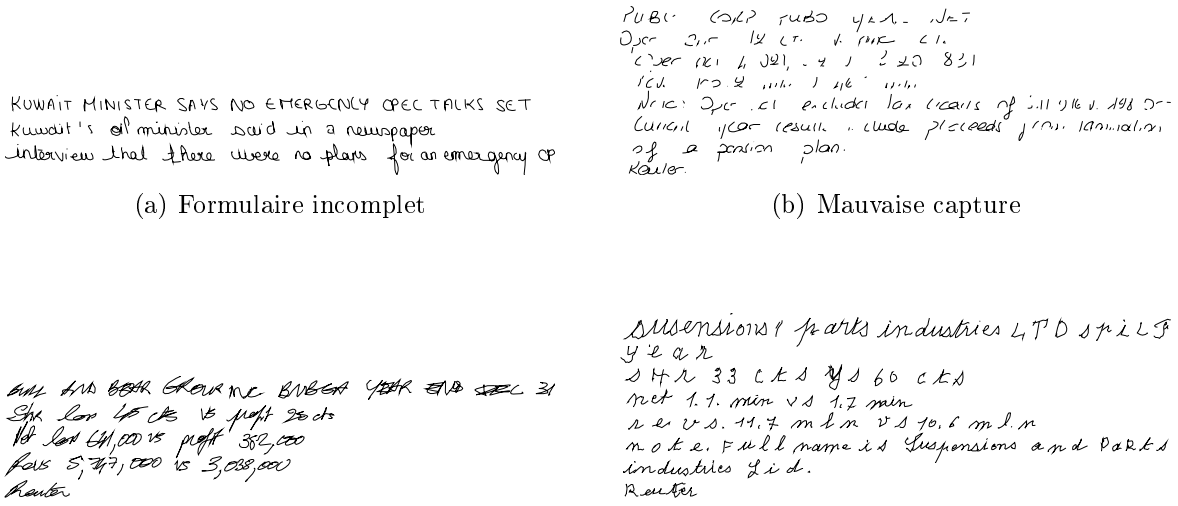
Par exemple, pour `lk-text`, 95 % des effectifs possèdent un WER inférieur à 50 % tandis que pour le même taux d'erreur, la ressource `lk-free` n'atteint qu'environ 35 % des effectifs.

Alors que pour `lk-free`, les effectifs sont distribués tout au long de l'intervalle du WER, pour les deux autres ressources ils sont regroupés dans la première moitié de l'intervalle. Pour chacune des ressources, le nombre de documents ne progresse pas de façon linéaire en fonction du WER. Cela pose un problème lors de l'analyse de l'impact du WER dans les applications en aval de la reconnaissance. En effet, il n'est pas possible de déterminer si cet impact est proportionnel au WER.

L'analyse par paliers prend alors tout son sens car elle permet de vérifier l'existence d'un impact progressif du taux d'erreur dans les performances de la CAT ou la RI. Cette vérification pourra être effectuée sans avoir recours à une introduction artificielle d'erreurs, permettant de contrôler l'évolution du WER, mais rendant les données peu représentatives des situations réelles (Mittendorf, 1998).

Un examen attentif des documents appartenant aux 5 derniers percentiles a permis de comprendre pourquoi certains documents ont des taux d'erreur très importants. Il y a principalement 3 raisons qui sont liées au contenu, à la capture et au scripteur. La première, relative au contenu, est le fait que certains documents soient incomplets (*cf.* figure 3.6(a)) par négligence du scripteur ou à cause d'une défaillance du stylo. La deuxième raison est liée à une mauvaise capture qui va rendre impossible la reconstruction du signal (*cf.* figure 3.6(b)). Enfin, la troisième raison, relative au scripteur, est l'influence négative de certains styles d'écriture (*cf.* figures 3.6(c) et 3.6(d)).

Malgré leur caractère défaillant ces documents n'ont pas été exclu du corpus. En effet, ces échantillons correspondent à des situations pouvant survenir dans le cadre d'applications réelles. Par conséquent, il a été choisi de les utiliser dans les expériences liées à la catégorisation (chapitre 5) et à la RI (chapitres 6 à 8).



(a) Formulaire incomplet

(b) Mauvaise capture

(c) Écriture difficile

(d) Écriture aérée

FIGURE 3.6 : Exemples de documents appartenant aux 5 derniers percentiles.

Chapitre 4

Catégorisation automatique de textes

« J'ai 7 marguerites et 3 roses, ai-je plus de marguerites que de fleurs ? »

Jean Piaget & Bärbel Inhelder

La psychologie cognitive définit la catégorisation comme un processus adaptatif fondamental pour l'appréhension du réel : catégoriser c'est comprendre le monde pour pouvoir agir sur lui. En effet, les catégories sont des entités discrètes qui interviennent dans toutes les formes d'interaction avec l'environnement. Dans la prise de décision par exemple, pour discriminer ce qui est mangeable de ce qui ne l'est pas. Dans la construction du sens également, lorsqu'un enfant est capable de distinguer entre des représentations de chats et celles de chiens, il a acquis les concepts de *chien* et *chat*.

Lorsqu'elle est appliquée aux textes, la catégorisation permet d'organiser l'information afin de rendre plus aisée son exploitation. Les applications de ces technologies sont nombreuses : recherche de documents par catégorie, organisation automatique, routage de documents, etc.

Ce chapitre est dédié à l'introduction de la catégorisation comme un problème d'apprentissage automatique. Sans bien entendu prétendre à l'exhaustivité, il tente de donner au lecteur les connaissances minimales pour pouvoir s'orienter dans les parties de ce mémoire liées à la catégorisation automatique. En partant des considérations théoriques (§4.1), seront abordés les éléments pratiques (§4.2) aboutissant à la mise en œuvre (§4.4) et à l'évaluation (§4.3) d'un système de catégorisation dont le but est d'assimiler « l'ensemble des traits nécessaires et suffisants qui définissent la(les) catégorie(s) » (Dubois, Giacomo, Gespin, Marcellesi, Marcellesi et Mével, 1999, pp. 66-67).

4.1 Définition

« Catégoriser, c'est constituer des classes d'équivalences, c'est être capable d'extraire des invariants tout en négligeant des caractéristiques non pertinentes » (Rossi, 2005, p. 129). Cette définition, empruntée aux sciences cognitives, fait le parallèle avec le modèle dominant dans la catégorisation de textes : l'*apprentissage supervisé*.

La catégorisation est un processus d'apprentissage. À l'instar des êtres vivants, c'est par l'expérience, par la rencontre des objets de la catégorie, que la capacité d'en extraire les caractéristiques s'acquiert. La catégorisation est également un processus supervisé. Par exemple, les catégories *chien* et *chat* préexistent à l'enfant qui les appréhende.

C'est la définition formelle de la tâche de catégorisation, au sens apprentissage automatique du terme, donnée par Sebastiani (2002) qui est adoptée dans ce travail.

Définition 4.1. *La* catégorisation automatique de textes (CAT) *consiste à assigner à chaque paire* $\langle d_j, c_i \rangle \in \mathcal{D} \times \mathcal{C}$ *une valeur booléenne. Avec* $\mathcal{D} = \{j | 1 \leq j \leq n\}$ *le domaine des documents et* $\mathcal{C} = \{i | 1 \leq i \leq |\mathcal{C}|\}$ *celui des catégories.*

Définition 4.2. *La* fonction de catégorisation exacte $\breve{\Phi} : \mathcal{D} \times \mathcal{C} \rightarrow \{0, 1\}$ *assignera une valeur de* 1 *à la paire* $\langle d_j, c_i \rangle$ *si* d_j *doit être classé dans* c_i*. Dans le cas contraire, la valeur assignée sera* 0*.*

Définition 4.3. *Un* algorithme de catégorisation *est une approximation de la fonction exacte* $\breve{\Phi}$*, notée* $\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{0, 1\}$*.*

Le processus d'induction de Φ se divise en trois phases : apprentissage, validation et test. L'objectif de la *phase d'apprentissage* est d'induire une approximation de $\breve{\Phi}$ à partir d'un jeu d'entraînement.

Définition 4.4. *L'apprentissage supervisé suppose l'existence d'un ensemble de documents* $\Omega = \{d_1, \ldots, d_{|\Omega|}\} \subset \mathcal{D}$ *dont la catégorie est connue à l'avance.* Ω *est appelé jeu d'entraînement.*

L'algorithme d'apprentissage extrait les éléments caractérisant les différentes catégories afin de pouvoir affecter tout nouveau document à une catégorie en fonction de ses caractéristiques. La qualité des modèles de catégorisation créés lors de la phase d'apprentissage dépend d'un ensemble de paramètres qu'il convient d'ajuster. Pendant la *phase de validation*, les paramètres du classifieur sont ajustés grâce à un jeu de validation. Il s'agit le plus souvent d'un sous-ensemble du jeu d'entraînement.

L'évaluation définitive du processus de catégorisation se fait durant la *phase de test*. Le jeu de test utilisé pour cette étape est un ensemble de documents pré-étiquetés dont l'information associée à la catégorie est utilisée pour estimer la qualité des réponses données par le classifieur. Il est distinct du jeu d'entraînement.

Contexte multi-catégorie, catégorisation multi-étiquette

La catégorisation est effectuée dans un contexte *multi-catégorie* lorsque $|\mathcal{C}| > 2$. Dans le cas contraire, le contexte est dit *binaire*, car les textes peuvent appartenir à

seulement deux catégories. Le processus de catégorisation est dit *multi-étiquette* lorsqu'il est possible de classer d_j dans plusieurs catégories. Lorsque cela n'est pas possible, le processus est appelé catégorisation *mono-étiquette*.

Le contexte des expériences effectuées dans le cadre de cette thèse est multi-catégorie. Sauf mention contraire, la catégorisation est, quant à elle, de type mono-étiquette. Ceci découle du choix du corpus de référence et de la méthodologie de collecte du corpus manuscrit (voir chapitre 3).

Catégorisation centrée-document, catégorisation centrée-catégorie

Lorsqu'il s'agit de trouver toutes les catégories $c_i \in \mathcal{C}$ pour un document d_j donné, le processus de catégorisation est centré sur le document. A contrario, lorsqu'il est question de retrouver tous les $d_j \in \mathcal{D}$ dont la catégorie est c_i, la catégorisation opère en mode centré-catégorie.

Même si la mise en œuvre de ces deux types de catégorisation diffère, il n'y pas de différence fondamentale d'un point de vue conceptuel. Le choix de mettre au centre du processus soit le document, soit la catégorie, n'a d'influence que sur les applications visées et les mesures de performance. Les aspects liés aux performances de la catégorisation seront développés en section 4.3.

4.2 Indexation

L'indexation de documents consiste à extraire les entités représentatives du sens d'un document. Les entités extraites sont représentées selon un formalisme prédéfini.

Dans le contexte de la catégorisation, le formalisme vectoriel (Salton, Wong et Yang, 1975) est le modèle de représentation d'un corpus de documents le plus utilisé. L'idée sous-jacente à ce modèle est que le sens d'un document est porté par l'ensemble d'unités lexicales qui le composent, indépendamment des relations contextuelles. Pour cette raison, cette représentation est appelée « sac de mots ».

À chaque document $j \in \mathcal{D}$ correspond un vecteur colonne, $\mathbf{d}_j$, où chaque élément du vecteur correspond à une unité lexicale. La valeur associée à cet élément indique son poids dans le document (*cf.* figure 4.1).

Définition 4.5. *En transposant les vecteurs des documents, le domaine des documents $\mathcal{D}$ peut être représenté par une matrice $\mathbf{A}$ de $n \times m$ éléments, où m est la taille de l'espace de représentation, c'est-à-dire du lexique. $\mathbf{A}$ est la matrice de la collection de documents.*

Les textes bruts en langue naturelle subissent un certain nombre de transformations afin de se conformer au modèle vectoriel. Un ensemble de pré-traitements linguistiques permet d'identifier les différentes unités porteuses de sens : les termes. La sélection de termes permet de construire un lexique permettant de caractériser les différentes catégories, « tout en négligeant des caractéristiques non pertinentes ». Enfin, l'étape de pondération permet de déterminer de manière quantitative la représentativité d'un terme. Ces transformations sont décrites ci-dessous.

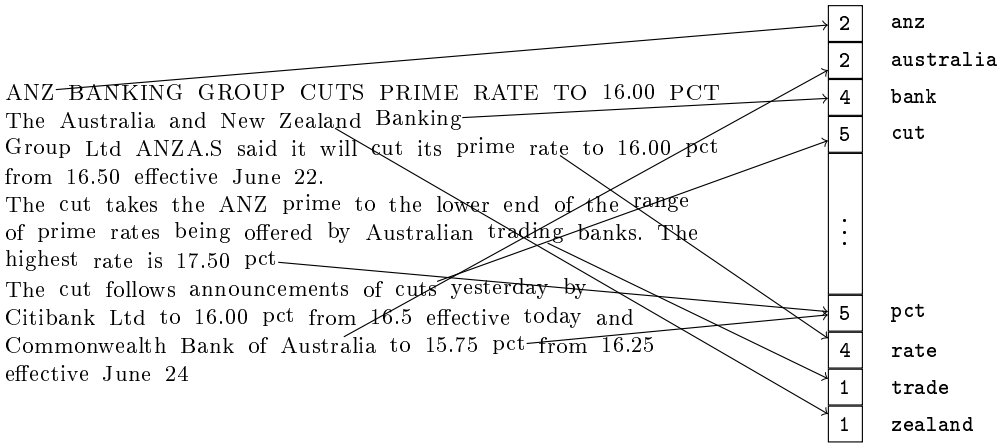


FIGURE 4.1 : Représentation vectorielle d'un texte.

4.2.1 Pré-traitements

L'ensemble des traitements effectués pour transformer un texte brut en une liste de formes canoniques que nous appellerons *termes d'indexation*, est donné par la figure 4.2. Ces traitements sont décrits ci-dessous.

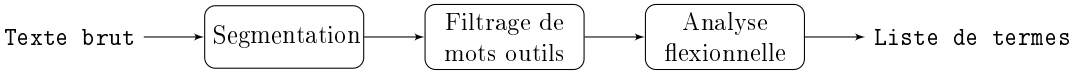


FIGURE 4.2 : Pré-traitements linguistiques.

Segmentation en occurrences de forme

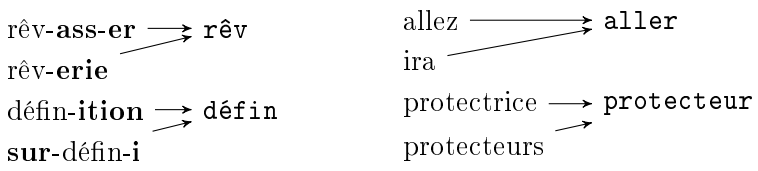
La segmentation ou tokenisation permet de séparer les éléments constituant le texte (*tokens*). Elle s'appuie généralement sur un ensemble de séparateurs tels que l'espace, la virgule ou le point.

Filtrage de mots outils

Après la segmentation, il est courant d'éliminer les unités lexicales qui ne portent pas de sens en elles-mêmes (*mots outils*). Une liste de mots outils est prévue pour cette tâche, elle est constituée de prépositions, conjonctions, déterminants, auxiliaires, etc.

Analyse flexionnelle

Dans les langues flexionnelles, les formes canoniques sont dérivées principalement selon leur nombre, genre ou mode. Nous pouvons supposer que deux formes dérivées véhiculent le même sens. Par exemple, *boit* et *boivent* partagent la notion de *boire*. Les deux unités linguistiques, couramment considérées comme représentatives du sens, sont la *racine* et le *lemme*. L'analyse flexionnelle permet de retrouver l'une de ces deux formes à partir d'une forme fléchie.



(a) Racinisation (b) Lemmatisation

FIGURE 4.3 : Analyse flexionnelle.

La racinisation consiste à supprimer tous les affixes d'un mot (*cf.* figure 4.3(a)), même si, de manière générale, les algorithmes existants n'effectuent que la désuffixation des mots (Porter, 1980 ; Savoy, 1999). La lemmatisation permet d'associer un lemme à chaque forme fléchie. Les entrées d'un dictionnaire peuvent être vues comme des lemmes (*cf.* figure 4.3(b)). La plupart des lemmatiseurs actuels effectuent, en plus de la lemmatisation, un étiquetage morpho-syntaxique (Schmid, 1994) ou un calcul des traits flexionnels d'un mot (Namer, 2000).

4.2.2 Sélection de termes

L'intérêt de la sélection de termes est triple. D'une part, elle permet d'écarter les termes non pertinents d'un point de vue statistique. D'autre part, elle permet d'éviter le surapprentissage (Sebastiani, 2002). Enfin, elle permet d'améliorer l'efficacité des algorithmes d'apprentissage ayant des difficultés à gérer un espace de représentation important.

Par le passé, diverses mesures ont été proposées afin de détecter les termes les plus discriminants des documents d'une catégorie donnée : la divergence de Kullback et Leibler (1951)[1], l'information mutuelle (Church et Hanks, 1989) ou le test du khi-deux (χ^2) (Yang et Pedersen, 1997). Le lecteur peut se référer aux travaux de Yang et Pedersen (1997), Galavotti, Sebastiani et Simi (2000) ou Forman (2003) pour des études comparatives des différentes méthodes.

Les expériences menées dans le cadre de cette thèse utilisent le test du χ^2 pour obtenir une liste de termes triés par ordre de pertinence pour chaque catégorie. Afin d'obtenir un espace de représentation commun à toutes les catégories, l'algorithme de Forman (2004) est utilisé. Le détail de la sélection de termes est donné en annexe A.

4.2.3 Pondération

La pondération de termes a pour but de déterminer de manière quantitative la représentativité d'un terme. L'hypothèse première de Salton et collab. (1975) est que la fréquence d'apparition d'un terme est liée à l'importance de celui-ci dans un document. Cependant, plus le terme apparaît dans l'ensemble des documents à indexer, moins il devient représentatif d'un document en particulier. La mesure $tf \times idf$ (Spärck Jones,

1. Connue également sous le nom de « gain d'information ».

1979) a été choisie dans cette étude pour évaluer l'importance d'un terme. Elle prend en compte sa fréquence locale, c'est-à-dire relative à un document (*term frequency*, tf) et sa fréquence globale, relative à un corpus (*inverse document frequency*, idf). Afin de réduire les effets engendrés par les différences de longueurs des documents, elle est divisée par la norme euclidienne de $\mathbf{d}_j$. Ainsi, le poids d'un terme i dans le document j est donné par la formule suivante :

$$\mathbf{A}_{j,i} = \frac{\mathrm{tf}(i,j) \times \mathrm{idf}(i)}{\|\mathbf{d}_j\|_2} \tag{4.1}$$

Avec $tf(i,j)$ la fréquence du terme i dans le document j. Le facteur $idf(i)$ est défini à partir du rapport entre la taille du corpus et le nombre de documents contenant le terme i ($n(i)$).

$$\mathrm{idf}(i) = \log\left(\frac{n}{n(i)}\right) \tag{4.2}$$

4.3 Évaluation des algorithmes de catégorisation

La CAT a hérité de toute l'artillerie issue de la recherche d'information, la façon de l'évaluer ne faisant pas exception. Les mesures classiques d'évaluation de la qualité d'un algorithme de classification sont la précision (π) et le rappel (ρ).

Pour une catégorie c_i donnée, S est l'ensemble des documents identifiés par un algorithme comme faisant partie de c_i et R est l'ensemble des documents faisant réellement partie de c_i. La figure 4.4 schématise ces deux ensembles. La précision et le rappel peuvent être dérivés à partir de ces ensembles.

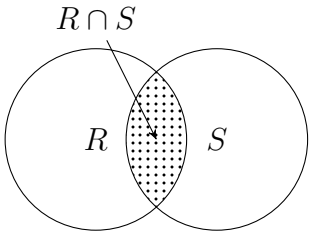


FIGURE 4.4 : Ensemble de documents supposés pertinents (S) et ensemble de documents réellement pertinents (R) pour une catégorie donnée.

Définition 4.6. *La* précision *mesure la proportion de documents correctement classés dans c_i parmi les documents de S. La précision évalue la capacité du classifieur à ne pas introduire de documents d'une autre catégorie dans c_i.*

$$\pi(c_i) = \frac{|R \cap S|}{|S|} \tag{4.3}$$

Définition 4.7. *Le* rappel *mesure la proportion de documents bien classés dans c_i parmi les documents de R. Le rappel évalue la capacité du classifieur à trouver tous les documents de c_i.*

$$\rho(c_i) = \frac{|R \cap S|}{|R|} \tag{4.4}$$

Il est possible d'augmenter la valeur de $\rho(c_i)$ en classant de plus en plus de documents dans c_i. Dans ce cas, la précision diminue car de plus en plus de documents non pertinents sont introduits. Inversement, il est possible d'augmenter la valeur de $\pi(c_i)$ en ne classant que les documents sûrs, au détriment du rappel. Ces deux mesures prises indépendamment l'une de l'autre ne permettent pas d'évaluer correctement un système de catégorisation.

Il existe donc un compromis à trouver entre $\pi(c_i)$ et $\rho(c_i)$. Ce compromis est accepté lorsqu'un seuil est défini (Yang, 2001) en fonction des besoins applicatifs. Lorsqu'il est nécessaire de donner une représentation détaillée du comportement du système, les courbes de précision *vs* rappel interpolées (Baeza-Yates et Ribeiro-Neto, 1999, pp. 76-78) doivent être utilisées. Le procédé d'interpolation maximise la précision à chaque niveau (Baeza-Yates et Ribeiro-Neto, 1999, p. 78), pour cette raison ces courbes constituent une borne supérieure des performances du système.

Une autre façon d'illustrer ce compromis est de donner une mesure prenant en compte l'importance relative de $\pi(c_i)$ et $\rho(c_i)$.

Définition 4.8. *La* F-mesure (van Rijsbergen, 1979, pp. 129-135) *est un indicateur prenant en compte la valeur relative de la précision et du rappel.*

$$F_\beta(c_i) = \frac{(1+\beta^2) \times \pi(c_i) \times \rho(c_i)}{\beta^2 \times \pi(c_i) + \rho(c_i)} \tag{4.5}$$

Une valeur de $\beta > 1$ donne plus de valeur à $\rho(c_i)$, une valeur comprise entre $0 < \beta < 1$ donne plus de valeur à $\pi(c_i)$. Lorsqu'il n'y a pas lieu de privilégier l'une des deux mesures, le paramètre β est fixé à 1.

Dans des contextes applicatifs où la catégorie d'un document est décidée par un expert, un système peut se contenter de trier les documents de S par ordre décroissant de pertinence. En considérant que l'expert humain ne s'intéresse qu'aux documents arrivant en tête, la précision peut être calculée pour un nombre fixe de documents. Inversement, les courbes de précision vs rappel supposent qu'un utilisateur s'intéresse à tous les documents de c_i.

Définition 4.9. *La* précision à n documents *évalue la capacité d'un algorithme à retrouver les documents de c_i dans les n premières positions.*

$$p@n(c_i) = \frac{|R \cap S|}{n} \tag{4.6}$$

Micro-moyenne, macro-moyenne

L'évaluation des algorithmes se faisant sur plusieurs catégories, il est nécessaire de résumer les mesures faites sur chaque catégorie en une seule valeur moyenne. Il y a deux façons de moyenner les valeurs de la précision, du rappel et de la f-mesure : la macro et la micro-moyenne.

Définition 4.10. *La* macro-moyenne *est donnée par la moyenne arithmétique des mesures par catégorie. Chaque catégorie a la même influence sur la moyenne. La macro-moyenne de la précision et du rappel est définie ci-dessous.*

$$\pi^M = \frac{\sum_{c=1}^{|C|} \pi(c)}{|C|} \qquad (4.7) \qquad\qquad \rho^M = \frac{\sum_{c=1}^{|C|} \rho(c)}{|C|} \qquad (4.8)$$

Définition 4.11. *La* micro-moyenne *de la précision et du rappel est calculée à partir de la somme des effectifs des différents ensembles. Dans la micro-moyenne chaque document a la même influence sur la moyenne.*

$$\pi^\mu = \frac{\sum_{c=1}^{|C|} |R_c \cap S_c|}{\sum_{c=1}^{|C|} |S_c|} \qquad (4.9) \qquad\qquad \rho^\mu = \frac{\sum_{c=1}^{|C|} |R_c \cap S_c|}{\sum_{c=1}^{|C|} |R_c|} \qquad (4.10)$$

En séparant les documents, la micro-moyenne donne plus d'importance aux classes les plus nombreuses. Comme celles-ci sont, en général, mieux classées, la micro-moyenne sera plus élevée que la macro-moyenne (Beney, 2008, p. 36). La moyenne de la précision à n documents est indépendante de la méthode de calcul car la taille de S est fixée.

Lorsque les méthodes de catégorisation sont évaluées dans une perspective centrée-catégorie, des courbes de précision vs rappel et des graphiques de précision à n documents seront présentés. Les courbes de précision vs rappel présentées sont données en macro-moyenne, elles ont été calculées grâce à l'utilitaire `trec_eval` (Voorhees et Harman, 2005, chap. 3).

Dans une perspective centrée-document, l'évaluation doit être faite en fonction des documents, c'est-à-dire en utilisant la micro-moyenne. De plus, comme les expériences présentées ici concernent la catégorisation mono-étiquette, les micro-moyennes de la précision et du rappel sont égales (Beney, 2008, p. 36). Une seule mesure, appelée *taux de classification*, sera alors utilisée pour l'évaluation du système selon cette perspective.

4.4 Expérimentations préliminaires

La dernière section de ce chapitre sera consacrée à la description du système de catégorisation adopté. Cette description omet volontairement une revue détaillée des algorithmes de classification existants, car l'objectif de cette thèse n'est pas lié à la définition d'un tel algorithme. Les algorithmes utilisés dans le cadre des expérimentations seront toutefois abordés d'un point de vue pragmatique.

La dernière partie de cette section présente les résultats d'une étude préliminaire réalisée sur la version $R(90)$ du corpus Reuters-21578. Cela a permis de vérifier la

robustesse du système par comparaison avec d'autres études utilisant le même corpus de référence.

4.4.1 Système de catégorisation

Le système de catégorisation adopté suit les étapes présentées dans la section 4.2. Ce système peut être décrit par l'ensemble des paramètres qui interviennent à chaque étape, le tableau ci-dessous résume la configuration du système pour l'utilisation de deux algorithmes d'apprentissage : un algorithme des k-plus proches voisins (k-PPV) (Mitchell, 1997, p. 234) et les séparateurs à vaste marge ou machines à vecteur de support (SVM) (Vapnik, 2000 ; Joachims, 2002).

TABLE 4.1 : Paramètres du système de catégorisation adopté.

Paramètre	Valeur
Mots outils	Liste de mots outils FreeWAIS (Pfeifer, Fuhr et Huynh, 1995).
Analyse flexionnelle	Racinisation par l'algorithme de Porter (1980).
Sélection de termes	χ^2 et algorithme de Forman (2004).
Espace de représentation	1 000 termes.
Pondération	$tf \times idf$
Algorithmes	k-PPV, SVM [2].

L'algorithme des k-PPV est basé sur l'hypothèse que si deux documents sont proches dans l'espace de représentation, alors ils ont une forte probabilité d'appartenir à la même catégorie. La décision d'attribuer une catégorie c_i à d_j est prise en fonction des classes des k plus proches voisins.

Les SVM sont des algorithmes d'apprentissage à base de noyaux (Schölkopf et Smola, 2002). Le noyau sert à projeter les données en entrée dans un espace de plus grande dimension. L'apprentissage servira à définir une frontière de séparation entre les documents de deux classes qui maximise la distance entre la frontière et les documents les plus proches (Burges, 1998). La décision d'attribuer une catégorie c_i à d_j est prise en fonction de la position de d_j par rapport à la surface de séparation.

4.4.2 Résultats

Afin de valider le système de catégorisation développé, la première expérience menée utilise la partie « cohérente » du corpus Reuters-21578, c'est-à-dire le sous-ensemble $R(90)$ (voir §3.1.2). Cela permet de positionner ce système par rapport aux résultats existants avec le même sous-ensemble du corpus.

2. L'implémentation de cette méthode a été réalisée grâce à l'application SVMlight V6.0 développée par Joachims (2002).

La valeur du paramètre k pour les k-PPV et le nombre de termes de l'espace vectoriel ont été déterminés par validation croisée avec 10 replis. La sélection de termes a été effectuée avec le χ^2 (voir annexe A). Le χ^2 est un test statistique robuste et a montré des résultats importants avec le corpus Reuters-21578 (Yang et Pedersen, 1997). De plus, il tient compte des interactions entre les termes et les catégories, tout en évitant de favoriser à outrance les termes rares, comme c'est le cas de l'information mutuelle (Yang et Pedersen, 1997).

Le nombre de termes est fixé à 1 000 et k est fixé à 30 pour l'algorithme des k-PPV. Les SVM sont utilisés avec un noyau linéaire et le paramètre c, qui contrôle le coût des erreurs à l'entraînement, est fixé à 1, il s'agit des paramètres par défaut de SVMlight. Ces paramètres n'ont pas été optimisés.

TABLE 4.2 : Micro-moyenne de la mesure F_1.

	Yang et Liu (1999)	Joachims (2002)	Système développé
kPPV	0,857	0,826	0,840
SVM	0,859	0,875	0,889

Le tableau 4.2 présente les résultats obtenus par le système implémenté, ainsi que ceux obtenus par Yang et Liu (1999) et Joachims (2002). Dans un souci de comparabilité, la micro-moyenne de la mesure F_1 est présentée.

Les résultats obtenus avec les k-PPV sont proches de ceux de référence, bien que légèrement en dessous de ceux de Yang et Liu (1999). En revanche, les résultats obtenus avec les SVM sont supérieurs à la F_1 de référence. Ceci permet d'attester du bon fonctionnement du système implémenté dans le cadre de cette thèse. Ce système va pouvoir être appliqué sur des données manuscrites, qui sont au centre de notre problématique. Le rapport de ces expériences fera l'objet du chapitre suivant.

Chapitre 5

Impact des erreurs de reconnaissance dans la catégorisation

Les erreurs de reconnaissance ont-elles un effet négatif sur la catégorisation de textes ? Oui, à titre d'exemple, lorsque la catégorisation est effectuée avec les documents reconnus avec `lk-free`, les performances du système baissent de plus de 10 %. Cela est une conséquence directe de la reconnaissance. En effet, elle a introduit plus de bruit que l'algorithme d'apprentissage ne peut tolérer.

Comme il s'agit, dans cette étude, d'utiliser un système de reconnaissance de l'écriture en amont de la catégorisation, notre problématique se situe dans le domaine de la catégorisation de documents bruités (Noisy Text Categorization, NTC). Mesurer l'impact des erreurs de reconnaissance dans un processus de catégorisation a fait l'objet d'un certain nombre d'études, en particulier dans le domaine de la reconnaissance optique de caractères (OCR) (Ittner, Lewis et Ahn, 1995 ; Junker et Hoch, 1998 ; Taghva, Nartker, Borsack, Lumos, Condit et Young, 2000 ; Murata, Busagala, Ohyama, Wakabayashi et Kimura, 2006). Cette question a également été abordée dans le domaine de la reconnaissance de la parole (Myers, Kearns, Singh et Walker, 2000 ; Agarwal, Godbole, Punjani et Roy, 2007). Le but de ces études est de répondre à une question d'ordre pratique, à savoir si la catégorisation peut être appliquée sans dégradation significative de ses performances par rapport à son application à des documents électroniques. Un autre point de vue sera adopté ici qui consiste à suivre la propagation des erreurs de reconnaissance dans tout le processus de catégorisation afin de comprendre pourquoi, et dans quelle mesure, la catégorisation de documents manuscrits est moins performante que celle de documents électroniques.

Ce chapitre reprend, révise et complète une partie des travaux déjà publiés dans le cadre de cette thèse (Peña Saldarriaga, Morin et Viard-Gaudin, 2008 ; Peña Saldarriaga, Viard-Gaudin et Morin, 2009b ; Peña Saldarriaga, Morin et Viard-Gaudin, 2009a ; Peña Saldarriaga, Viard-Gaudin et Morin, 2010). Sont présentés les principaux résultats des expériences de catégorisation avec le corpus manuscrit. Eu égard à la taille réduite du

corpus ayant servi pour la collecte des données manuscrites, les différents paramètres des classifieurs ont été ajustés par rapport à ceux rapportés dans le chapitre précédent. Ceci a été fait par validation croisée à 10 partitions (Hastie, Tibshirani et Friedman, 2008, pp. 241-245) sur le jeu d'entraînement électronique composée de 1 847 documents. Les SVM sont utilisés avec un espace de représentation composé de 1 000 termes et les kPPV avec $k = 15$ et 300 termes. Ces mêmes paramètres ont été conservés pour la catégorisation des documents manuscrits.

Dans un premier temps, une revue de l'état de l'art sur la catégorisation de documents bruités sera présentée (§5.1). La section 5.2 définit les principales mesures associées à la notion de « bruit » d'un document tandis que la section 5.3 livre un ensemble de données factuelles sur la quantité de bruit présent dans le corpus issu de la reconnaissance. Les sections 5.4 à 5.6 s'intéressent à l'impact du bruit sur la catégorisation, aussi bien au niveau de la représentation des documents (§5.4) qu'au niveau de l'entraînement (§5.5) et de la phase de test (§5.6). Enfin, la dernière section fera la synthèse de l'étude présentée dans ce chapitre.

5.1 Bilan des recherches sur la catégorisation de documents bruités

Il existe très peu de travaux en rapport avec la catégorisation de textes manuscrits. Cela est dû essentiellement à l'absence, ou à la difficulté de collecter de larges corpus de données manuscrites. La plupart des travaux sur la NTC se basent sur des données issues d'un processus de reconnaissance optique de caractères (Ittner et collab., 1995 ; Junker et Hoch, 1998 ; Taghva et collab., 2000 ; Murata et collab., 2006). Bien que les documents issus de l'OCR soient de même nature que ceux issus de la reconnaissance de l'écriture manuscrite, il existe une différence notable de difficulté entre les deux situations. En effet, lorsque les images sont de bonne qualité, les taux d'erreur habituels de l'OCR sont très faibles comparés à ceux obtenus avec la reconnaissance de l'écriture manuscrite. Selon la taille du lexique utilisé et la présence de modèles de langage, la reconnaissance de textes manuscrits peut conduire à des taux de reconnaissance exprimés au niveau caractère se situant entre 80 % à 90 % (Perraud, 2005) alors que, pour du texte imprimé, les meilleurs systèmes dépassent 99 % de reconnaissance (Belaïd et Cecotti, 2006).

Dès 1995, Ittner et collab. (1995) se sont intéressés à la catégorisation d'images de résumés d'articles de recherche. Les images, à l'origine à 300 dpi, ont été sous-échantillonnées pour atteindre la résolution classique des télécopieurs (200×100 dpi). Les taux d'erreur après l'étape d'OCR pouvaient aller de 4 % à 69 % avec une moyenne de 23 %. L'impact du bruit sur la catégorisation avec l'algorithme de Rocchio (Rocchio, 1971) semblait acceptable. La dégradation relative était de 16 % par rapport à l'expérience de contrôle, lorsque les documents électroniques étaient utilisés pour l'entraînement, et de seulement 7 % en utilisant des documents bruités à l'entraînement. Cela montrait qu'il valait mieux apparier des documents de même nature.

Afin de contourner le problème lié au bruit, Junker et Hoch (1998) appliquent une

technique de catégorisation à base de n-grammes (Cavnar et Trenkle, 1994), supposée être plus robuste vis-à-vis du bruit. Cette étude se focalise sur la manière d'indexer les documents : unigrammes, n-grammes, avec ou sans filtrage de mots outils, etc. En revanche, elle ne s'intéresse pas à l'impact du bruit dans les différentes représentations, et aucune comparaison avec une expérience de contrôle n'est rapportée.

Taghva et collab. (2000) ont également examiné les effets de l'OCR sur une méthode de classification bayésienne naïve (Maron, 1961) avec un corpus de la NRC (Nuclear Regulatory Commission). Leurs expériences ont montré que les méthodes de sélection de termes (gain d'information et seuil d'occurrences) permettent de réduire l'impact du bruit, qui devient alors quasi nul. Cependant ces conclusions semblent dépendantes du taux d'erreur observé (14 %), et surtout du nombre de catégories comptabilisées dans les performances (6 sur 52). De ce fait, ces conclusions sont difficilement généralisables à la problématique abordée ici.

Un travail récent sur l'OCR (Murata et collab., 2006) utilise une version imprimée puis scannée du corpus Reuters-21578. 750 documents ont été scannés avec des résolutions allant de 130 dpi à 300 dpi. Cette étude s'intéresse à différents moyens de pondérer les fréquences des termes d'indexation, mais pas à l'impact du bruit dans ces différentes pondérations. Aucune comparaison avec l'indexation sur les documents électroniques n'est mentionnée. En ce sens, ce travail peut être rapproché de celui de Junker et Hoch (1998).

Dans le domaine manuscrit, très peu de travaux existent ou ont fait l'objet d'une publication (Vinciarelli, 2005b ; Koch, 2006 ; Milewski, Govindaraju et Bhardwaj, 2009). Il faut noter que tous ces travaux s'intéressent à l'écriture hors-ligne. Les seuls travaux effectués s'intéressant à l'écriture en-ligne sont, à notre connaissance, ceux faisant l'objet de cette thèse.

Les travaux princeps de Vinciarelli (2005b) se placent dans un contexte mono-scripteur, c'est-à-dire où tous les documents ont été écrits par la même personne. La catégorisation est effectuée avec des SVM en utilisant des documents électroniques pour l'entraînement. La phase de test s'effectue sur un sous-ensemble de 200 documents issus du corpus Reuters-21578. Des expériences avec un corpus de synthèse sont également rapportées. L'intérêt de ces expériences est de pouvoir contrôler la quantité d'erreurs de reconnaissance injectée dans les documents (entre 10 % et 45 %). Cependant, comme le fait remarquer Mittendorf (1998), l'uniformité de ces injections ne fait pas des documents ainsi synthétisés des entités représentatives de situations réelles.

La catégorisation de courriers entrants a été explorée par Koch (2006) en utilisant un algorithme des k-PPV. Un système de reconnaissance de mots manuscrits isolés, assimilable à un système de word spotting, est utilisé pour la détection des termes d'indexation. Afin de « ne pas biaiser l'apprentissage par les erreurs de reconnaissance », l'entraînement est effectué sur la vérité terrain associée aux documents manuscrits. Le test s'effectue sur une base de courriers entrants manuscrits de 358 documents répartis en 6 catégories. Bien que la catégorisation s'effectue dans des « conditions réelles » (Koch, 2006), l'entraînement s'effectue toujours sur des documents parfaits. L'auteur conclut que, dans ces conditions, un système est suffisamment tolérant vis-à-vis du bruit pour « envisager son application aux documents manuscrits de type courriers entrants ».

Plus récemment, Milewski et collab. (2009) se sont intéressés à la catégorisation de

formulaires médicaux. La catégorisation se base sur l'analyse sémantique latente (LSA) (Deerwester, Dumais, Furnas, Landauer et Harshman, 1990). Ce travail se différencie de ceux qui viennent d'être abordés par un autre objectif que celui de la catégorisation. En effet, les catégories associées aux formulaires vont permettre de sélectionner un lexique adapté pour la reconnaissance du formulaire. Cela permet de ne pas utiliser des lexiques de trop grande taille qui peuvent avoir un impact négatif sur la reconnaissance (Xue et Govindaraju, 2002).

Les travaux présentés ici se distinguent de ces différentes études à plusieurs niveaux. Premièrement, par l'utilisation de l'écriture en-ligne, particularité de cette étude. Deuxièmement, par l'utilisation d'un corpus de grande taille, comparé à ceux utilisés dans les études sur l'écriture hors-ligne. De plus, ce corpus se base sur une collection standard dans le domaine de la catégorisation de textes. Troisièmement, contrairement aux études proches de celles présentées ici (Vinciarelli, 2005b ; Koch, 2006), des documents bruités sont utilisés aussi bien dans la phase d'entraînement que dans la phase de test, car il s'agit également de mesurer le biais introduit par les erreurs de reconnaissance dans les modèles de catégorisation. Enfin, cette étude se place d'une part dans une optique descriptive de la propagation du bruit dans toute la chaîne de catégorisation, et dans une optique comparative de l'effet du bruit sur deux algorithmes d'apprentissage de nature très différente, à savoir k-PPV et SVM.

5.2 Mesures du bruit

Comme il a été présenté en §2.4, les erreurs de reconnaissance sont habituellement mesurées au niveau caractère ou au niveau mot. Dans le cadre de la catégorisation, les mesures au niveau caractère n'ont d'intérêt que lorsque la représentation des documents se base sur les caractères, par exemple, dans le cas d'une représentation par n-grammes.

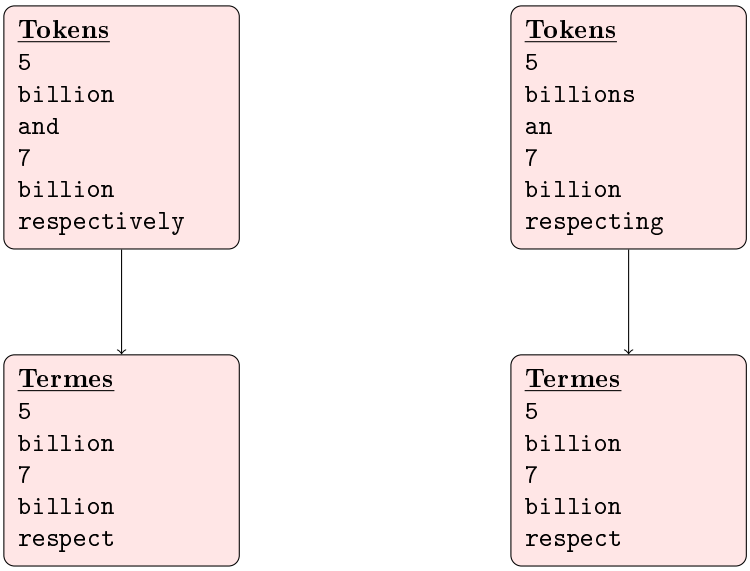
De la même manière, le taux d'erreur au niveau mot ne peut donner une estimation valable du bruit que lorsque la représentation des documents est faite au niveau mot. Or, comme il a été dit précédemment, l'unité porteuse de sens dans notre contexte est la racine, designée également comme terme ou terme d'indexation. En observant la figure 5.1, il peut être remarqué que certaines erreurs de reconnaissance n'ont aucun impact sur la représentation finale du document. En effet, la confusion entre deux mots dont la racine est la même, ou deux mots outils, est sans incidence sur la représentation du document.

En ce sens, le WER est une surestimation du bruit dans notre contexte. Une façon de tenir compte des pré-traitements est de mesurer le bruit au niveau terme.

5.2.1 Taux d'erreur au niveau terme

Le taux d'erreur au niveau terme est calculé de la même façon que le WER, c'est-à-dire grâce à un algorithme d'alignement (§2.4). Cependant, ce sont les termes qui sont alignés et non les mots.

Définition 5.1. *Le taux d'erreur au niveau terme (TER, Term Error Rate), représente le pourcentage de termes mal reconnus dans le document. Il est obtenu en normalisant*



(a) Vérité terrain (b) Reconnaissance

FIGURE 5.1 : Effets des pré-traitements sur les erreurs de reconnaissance.

le nombre d'erreurs (t_e) par le nombre de termes du document de référence (t_{ref}).

$$\mathsf{TER} = \frac{t_e}{t_{ref}} \tag{5.1}$$

En tenant compte des pré-traitements, le TER du document reconnu dans la figure 5.1 est égal à 0.

5.2.2 Plans de recouvrement

Le TER permet de mieux cerner l'impact du bruit sur la représentation des documents. Toutefois, son mode de calcul ne tient pas compte d'une partie importante des erreurs de reconnaissance : l'insertion de « faux » termes. Cet inconvénient peut être surmonté en comptant combien de termes sont correctement reconnus, et quelle est leur proportion par rapport au nombre total de termes de référence (Vinciarelli, 2005b). Dans ce but, Vinciarelli (2005b) introduit deux mesures du bruit supplémentaires : la précision-terme et le rappel-terme.

Définition 5.2. *La précision-terme (TP, Term Precision), représente la proportion de termes reconnus qui correspondent à des termes présents dans le document de la vérité terrain. Il est obtenu en normalisant le nombre de termes bien reconnus par le nombre de termes du document transcrit (t_{reco}).*

$$\mathsf{TP} = \frac{t_{ref} - t_e}{t_{reco}} \tag{5.2}$$

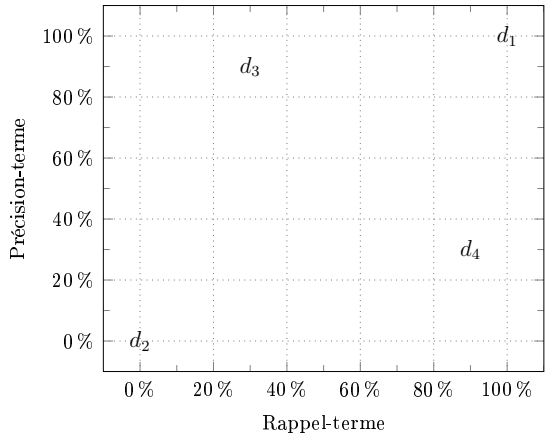


FIGURE 5.2 : Plan de recouvrement pour quatre documents.

Définition 5.3. *Le rappel-terme (TR, Term Recall), représente le taux de reconnaissance, c'est-à-dire la proportion de termes bien reconnus sur le nombre de termes à reconnaître. Il peut être calculé à partir du* TER.

$$\mathsf{TR} = 1 - \mathsf{TER} \tag{5.3}$$

Ces deux mesures permettent de représenter les documents dans un espace à deux dimensions, appelé plan de recouvrement (Coverage plan, CP), où chaque dimension correspond à une des mesures. La figure 5.2 montre un exemple de CP pour quelques documents. Dans cette figure, le document d_1 a la même représentation que sa version originale, tandis que le document d_2 ne contient aucun terme présent dans le document de la vérité terrain. Le document d_3 a une précision-terme très haute, mais il ne contient que 30 % des termes de référence. A contrario, le document d_4 contient 90 % de ces termes mais il introduit beaucoup de faux termes, d'où sa très basse précision-terme.

Toutefois, la figure 5.2 ne permet pas de visualiser deux informations importantes. D'une part, il est impossible de savoir si le document a été correctement catégorisé. D'autre part, il n'est pas possible de savoir si les erreurs de catégorisation sont dues au bruit. Afin de surmonter ces limites, lors de la présentation des plans de recouvrement, seuls les documents dont la catégorie change lorsque leur version manuscrite est utilisée y seront représentés. De plus ces documents seront représentés par des + lorsqu'ils sont bien catégorisés et par des • lorsqu'ils sont mal classés.

5.3 Bruit du corpus

Cette section présente les résultats de la reconnaissance du corpus manuscrit exposés dans le chapitre 3 avec les différentes ressources proposées en §2.3. La reconnaissance est évaluée dans la perspective du processus de catégorisation, c'est-à-dire au niveau terme.

TABLE 5.1 : Taux d'erreur exprimé au niveau terme, le WER est donné entre parenthèses.

	`lk-free`	`lk-slex`	`lk-text`
Moyenne	54,40 % (51,70 %)	26,21 % (25,23 %)	23,40 % (20,95 %)
Écart-type	19,32 % (17,39 %)	12,63 % (12,22 %)	13,13 % (12,96 %)
Médiane	54,10 % (51,52 %)	25,31 % (23,65 %)	21,57 % (17,77 %)
1er quartile	40,32 % (39,57 %)	16,82 % (16,07 %)	12,69 % (10,47 %)
3e quartile	69,65 % (64,38 %)	34,33 % (32,79 %)	33,33 % (30,52 %)

Le tableau 5.1 fournit diverses mesures descriptives de la distribution du TER dans le corpus. Tous les indicateurs, lorsqu'ils sont exprimés au niveau terme, sont supérieurs à ceux exprimés au niveau mot. Ce phénomène peut sembler paradoxal, étant donné que les pré-traitements devraient permettre, en théorie, de lisser certaines erreurs. Cependant, il s'explique par la bonne reconnaissance des mots outils dans les documents, lorsqu'ils sont supprimés ils ne contribuent plus à l'amélioration du taux d'erreur.

Le passage du WER au TER ne s'accompagne pas d'un changement important de la forme de la distribution (*cf.* figure 5.3). La médiane reste le meilleur indicateur du TER dans le corpus, car elle illustre bien l'ordre de grandeur du TER dans l'intervalle modal des trois distributions (*cf.* figure 5.3). Encore une fois, cette mesure ne doit être utilisée que dans une perspective d'analyse globale. Lorsqu'il est question d'analyser les documents individuellement, les quantiles s'avèrent être des indicateurs plus pertinents de l'évolution du bruit dans les documents.

Les mesures descriptives de la distribution du TER dans le corpus, données dans le tableau 5.1, ne permettent pas de mettre en évidence une différence de tendance du TER entre les documents des différentes catégories. Pourtant, il est important de connaître le TER par catégorie, lors de l'analyse des données, afin d'éviter deux artefacts. D'une part, si la catégorie regroupant le plus d'effectifs était la mieux reconnue, cette bonne performance pourrait occulter, en micro-moyenne, une perte de performances dans les autres catégories. D'autre part, une catégorie très mal reconnue pourrait aggraver, en macro-moyenne, une perte de performances liée à une seule catégorie. Le tableau 5.2 montre la médiane du TER par catégorie en fonction de la ressource utilisée pour la reconnaissance.

Le tableau 5.2 montre que, à l'exception de la classe *earn*, la distribution du TER dans les 9 autres catégories est globalement dans la même proportion. Les variations observées au sein de cet ensemble sont imputables à la distribution inégale des effectifs pour chaque catégorie et ne reflètent pas une réelle différence de tendance. En ce qui concerne la catégorie *earn*, il s'agit de la catégorie la mieux reconnue pour `lk-free`, alors qu'elle est la moins bien reconnue pour les deux autres ressources. Cela est dû à la présence importante de termes hors lexique de l'outil de reconnaissance dans les documents de cette catégorie.

Le tableau 5.3 s'intéresse à la TP [1]. La distribution de la précision-terme par catégo-

1. Le TR est volontairement omis car il est complémentaire du TER.

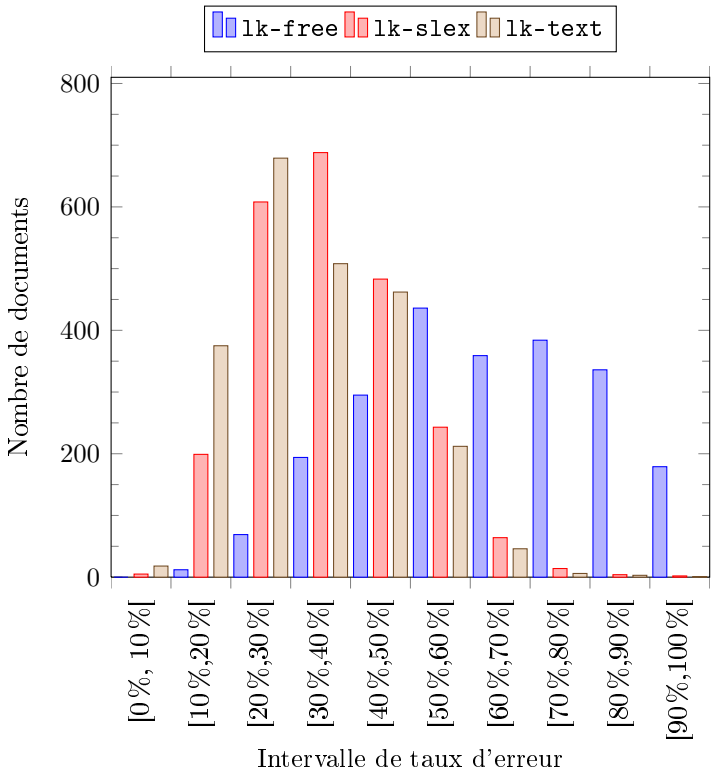


FIGURE 5.3 : Nombre de documents par intervalles du TER.

TABLE 5.2 : Médiane du TER par catégorie selon la ressource utilisée pour la reconnaissance.

	lk-free	lk-slex	lk-text
earn	40,35 %	31,91 %	33,33 %
acq	65,84 %	22,47 %	16,54 %
grain	58,58 %	17,71 %	14,29 %
money-fx	62,69 %	22,81 %	17,39 %
crude	65,79 %	21,21 %	17,72 %
interest	55,74 %	17,39 %	13,51 %
trade	64,08 %	19,86 %	13,67 %
ship	64,81 %	18,81 %	14,67 %
sugar	64,37 %	18,73 %	14,27 %
coffee	63,46 %	16,93 %	12,27 %

rie est consistante avec celle du TER. La précision-terme est très basse pour `lk-free` : la moitié des documents ont une TP inférieure à 42,55 %. Les deux autres ressources se retrouvent très au-dessus de ces performances. Il est important de remarquer que la différence entre la TP obtenue avec `lk-slex` et `lk-text` est d'environ 10 %, tandis que pour le WER, cette différence est d'à peine 3 %.

TABLE 5.3 : Médiane de la précision-terme (TP) par catégorie selon la ressource utilisée pour la reconnaissance.

	`lk-free`	`lk-slex`	`lk-text`
earn	61,19 %	71,01 %	78,95 %
acq	29,56 %	71,15 %	81,53 %
grain	36,43 %	76,35 %	83,74 %
money-fx	32,56 %	72,09 %	82,76 %
crude	31,82 %	74,00 %	82,56 %
interest	42,55 %	80,65 %	86,96 %
trade	29,77 %	73,69 %	85,47 %
ship	31,43 %	75,74 %	83,64 %
sugar	33,53 %	76,57 %	85,96 %
coffee	32,07 %	77,56 %	87,04 %
Médiane	42,55 %	72,50 %	81,82 %

En tenant compte de ces différents paramètres, il va s'agir d'évaluer l'impact du bruit à chaque maillon de la chaîne de catégorisation. C'est l'impact dit précoce, c'est-à-dire celui qui se manifeste dans la représentation des documents indépendamment des algorithmes de catégorisation, qui sera d'abord étudié (§5.4). Ensuite, le biais introduit dans les modèles de catégorisation sera analysé (§5.5). Enfin, la dernière section de ce chapitre fera le point sur l'impact du bruit sur la catégorisation (§5.6).

5.4 Impact précoce

Les recherches présentées précédemment (§5.1) ne s'intéressent qu'à l'impact final du bruit dans la catégorisation. Même si cela permet de répondre à une question d'ordre pratique, à savoir si la catégorisation peut être effectuée sans baisse significative des performances, cette démarche n'est pas explicative de tous les phénomènes induits par le bruit dans les documents. Pourtant, sans une compréhension approfondie de ces phénomènes, des tentatives pour améliorer les performances de la catégorisation peuvent se révéler infructueuses (Peña Saldarriaga, Morin et Viard-Gaudin, 2009b ; Peña Saldarriaga, Morin et Viard-Gaudin, 2009a).

Alors que nos premières études s'inscrivent dans une approche pratique (Peña Saldarriaga et collab., 2008, 2009b ; Peña Saldarriaga et collab., 2009a), il s'agit ici d'adopter une démarche descriptive et explicative. Cette démarche se concrétise par une analyse de l'impact du TER par étapes à l'aide de différentes mesures. Cela permet de

comprendre comment les erreurs de reconnaissance se propagent d'étape en étape, de la segmentation en occurrences de formes (tokenisation) jusqu'à la phase de catégorisation.

Les problèmes de reconnaissance se traduisent par une introduction très importante de mots dans la version transcrite du corpus manuscrit qui n'existent pas dans le corpus électronique et qui sont principalement des hapax (c'est-à-dire des mots ou termes qui n'apparaissent qu'une seule fois dans le corpus). Le tableau 5.4 montre l'importance de ces erreurs sur le lexique extrait des documents d'entraînement des différentes versions du corpus. Ces lexiques sont construits à partir des mots reconnus par MyScript® Builder, l'augmentation du nombre de mots correspond alors aux erreurs induites par la reconnaissance. Dans le tableau 5.4, seuls les documents d'entraînement sont considérés, mais les mêmes mesures peuvent être calculées à partir de l'ensemble de test.

TABLE 5.4 : Nombre de mots et termes d'indexation uniques en fonction de la ressource utilisée pour la reconnaissance. Le taux de recouvrement avec la vérité terrain (première colonne) est donné entre parenthèses.

	Réf.	`lk-free`	`lk-slex`	`lk-text`
Mots	8 805	49 436 (58,72 %)	14 666 (72,49 %)	13 828 (82,82 %)
Mots hors hapax	4 054	6 252 (51,43 %)	6 649 (86,06 %)	5 319 (89,17 %)
Termes	6 545	46 513 (62,83 %)	10 673 (67,33 %)	10 658 (80,69 %)
Termes hors hapax	2 931	6 034 (55,34 %)	5 377 (85,57 %)	4 167 (89,05 %)

Avec la ressource `lk-free`, le nombre de mots est multiplié par un peu plus de 5. Avec les deux autres ressources, le nombre de mots est à peine doublé. Lorsque les hapax ne sont pas pris en compte, le nombre de mots et termes distincts issus de la reconnaissance baisse dramatiquement.

L'introduction d'hapax va agir sur la bonne conservation du lexique de la vérité terrain. Alors que le nombre de termes hors hapax par rapport à la référence est presque doublé pour les ressource `lk-free` et `lk-slex`, le lexique de référence n'est conservé qu'à hauteur de 55,34 % pour `lk-free` et 85,57 % pour `lk-slex`. La ressource `lk-text` affiche, quant à elle, un taux de 89,05 %. La contrainte lexicale présente dans `lk-slex` et `lk-text` permet de mieux conserver le lexique d'origine.

Lors de la phase d'entraînement, après l'étape de pré-traitements, vient l'étape de sélection de termes. Celle-ci s'effectue sur les lexiques de termes hors hapax dont les caractéristiques sont données par la dernière ligne du tableau 5.4. L'introduction de termes hors du lexique d'origine se fait au détriment de ceux en faisant partie, se traduisant par une modification de la fréquence d'apparition de certains termes, nuisibles à la suite du processus de catégorisation.

5.4.1 Sélection de termes

La sélection de termes implémentée se base sur la statistique du χ^2. Le χ^2 permet d'obtenir pour un terme t et une catégorie c, un score qui reflète l'indépendance de t et

c. De manière générale, plus la valeur du χ^2 est élevée, plus t et c semblent dépendants l'un de l'autre. La première étape de la sélection consiste à calculer la distribution des scores du χ^2 pour chacune des catégories. L'impact du bruit sur ces distributions peut être mesuré grâce au coefficient de corrélation de rangs de Spearman (Myers et Well, 2003, p. 508). Le coefficient de Spearman mesure la concordance entre deux classements, plus sa valeur est proche de 1, plus les deux classements sont concordants, lorsqu'elle est égale à -1 les classements sont l'inverse l'un de l'autre. Les valeurs données dans le tableau 5.5 peuvent être interprétées comme une mesure de la concordance entre le classement des termes issu de la vérité terrain (texte électronique) et celui obtenu avec les termes issus de la reconnaissance.

TABLE 5.5 : Score de corrélation des distributions des scores du χ^2 par rapport à la vérité terrain.

	`lk-free`	`lk-slex`	`lk-text`
earn	0,52	0,80	0,82
acq	0,54	0,82	0,83
grain	0,51	0,76	0,79
money-fx	0,51	0,75	0,77
crude	0,51	0,76	0,78
interest	0,47	0,74	0,77
trade	0,48	0,75	0,77
ship	0,49	0,76	0,79
sugar	0,46	0,74	0,78
coffee	0,47	0,75	0,79
Moy.	0,50	0,76	0,79

Les scores donnés reflètent bien la qualité des documents dont ces distributions du χ^2 sont issues. Cependant, la mise en correspondance des valeurs des tableaux 5.2 et 5.5 montre qu'il n'y a pas de relation explicite entre le TER de la catégorie et l'impact qu'elle subit. La TP (tableau 5.3) ne semble pas non plus avoir de relation explicite avec les scores du tableau 5.5. En revanche, les deux catégories ayant le plus d'effectifs obtiennent le score de corrélation le plus élevé.

La différence entre les distributions des scores du χ^2 va se traduire par un écart entre les espaces vectoriels générés avec les documents manuscrits et l'espace vectoriel de référence. La tableau 5.6 montre cet écart.

Le premier indicateur de cet écart est le taux de recouvrement, c'est-à-dire, la proportion de termes partagés avec la vérité terrain. Le second est le taux d'orphelins. Par orphelins sont désignés les termes faisant partie de l'espace de représentation qui n'existent pas dans le lexique de la vérité terrain (hapax compris).

Comme attendu, plus les classements concordent, plus le recouvrement est important. De plus, le taux de recouvrement diminue lorsque le nombre de termes choisis augmente, mais dans des proportions différentes selon la ressource utilisée pour la reconnaissance. Pour la ressource `lk-text`, le recouvrement ne diminue que de 2 % tandis

TABLE 5.6 : Taux de recouvrement des espaces vectoriels sélectionnés avec le χ^2 et taux d'orphelins par rapport à la vérité terrain, le nombre de termes est donné entre parenthèses.

	`lk-free`	`lk-slex`	`lk-text`
Recouvr. à 300	56,33 % (169)	76,00 % (228)	80,33 % (241)
Recouvr. à 1 000	44,00 % (440)	72,40 % (724)	78,50 % (785)
Orphelins à 300	30,00 % (90)	8,00 % (24)	5,00 % (15)
Orphelins à 1 000	47,80 % (478)	17,50 % (175)	11,20 % (112)

que pour `lk-free` il diminue d'environ 12 %.

Le taux d'orphelins montre que les erreurs de reconnaissance se traduisent par l'apparition de « faux » termes dont la fréquence est suffisamment importante pour se retrouver dans l'espace de représentation final. Le nombre d'orphelins augmente en fonction du nombre de termes de l'espace de représentation. Ces termes orphelins constituent un biais dans la phase d'apprentissage. La sélection de 1 000 termes avec la ressource `lk-free` donne lieu à un espace de représentation dont presque la moitié est constituée de termes orphelins. Mais cela ne doit pas être systématiquement considéré comme nuisible. En effet, mieux vaut conserver des termes qui peuvent se reproduire dans les documents de test plutôt que des termes qui n'y apparaîtront jamais car le système de reconnaissance n'est pas en mesure de les reconnaître.

Le reste des termes composant l'espace vectoriel correspond aux termes qui, tout en faisant partie du lexique de référence, n'existent pas dans l'espace vectoriel de référence. La modification de l'espace de représentation des données est en elle-même une modification de la structure initiale des données comme il est montré ci-dessous.

5.4.2 Structure des données

L'impact sur l'espace de représentation des données est un impact sur l'algorithme d'apprentissage. Dans le cas des k-PPV, cela se traduit par une modification du voisinage local d'un document d. Le voisinage d'un document dépend d'une mesure de similarité, ici le cosinus, dont le résultat varie selon l'espace de représentation.

En considérant ℓ documents d'entraînement, la phase d'apprentissage des SVM revient à minimiser la fonction objectif suivante (Vapnik, 2000, p. 141) :

$$\mathcal{W}(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j \mathcal{K}(\mathbf{d}_i, \mathbf{d}_j) \tag{5.4}$$

sous les contraintes :

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0, \quad \alpha_i \geq 0 \tag{5.5}$$

Dans la formule 5.4 il faut souligner que l'expression à droite dépend des vecteurs des documents i et j. Un noyau linéaire, c'est-à-dire le produit scalaire $\mathbf{d}_i^\top \mathbf{d}_j$ entre les vecteurs, correspond au cosinus entre deux vecteurs de norme L_2.

Afin d'observer l'effet du bruit sur la structure des données, les documents sont projetés dans un espace à deux dimensions. Cette projection ne peut être qu'approximative et ne constitue pas une preuve formelle de ce qui est avancé. Elle a cependant l'avantage d'être intuitive et de fournir un support explicatif de l'impact du bruit sur la structure des données.

Les deux méthodes de projection utilisées (Belkin et Niyogi, 2002 ; van der Maaten et Hinton, 2008) représentent la collection $\mathcal{D}$ de n documents sous la forme d'un graphe pondéré : une matrice d'adjacence $\mathbf{G}$ de dimension de $n \times n$ dont la valeur de l'élément $\mathbf{G}_{ij}$, $i \neq j$ est donnée par une mesure de distance ou similarité. Le but de ces méthodes est de calculer une représentation de faible dimension tout en préservant les relations de proximité des données dans l'espace de départ.

Dans la première méthode (Belkin et Niyogi, 2002), la projection en deux dimensions se fait grâce aux vecteurs propres de la matrice laplacienne de G (voir §8.1). La méthode par t-SNE (t-Stochastic Neighbor Embedding) convertit les distances de l'espace de départ en probabilités, grâce à une gaussienne. Un ensemble de probabilités suivant une loi de Student est généré aléatoirement dans l'espace à deux dimensions. La suite de l'algorithme consiste à réduire l'écart entre les deux distributions, au sens de la divergence de Kullback et Leibler (1951), par une descente de gradient (van der Maaten et Hinton, 2008).

Les figures 5.5 et 5.6 montrent le résultat de la projection par la méthode laplacienne. Chaque couleur représente une catégorie, cette correspondance est donnée en figure 5.4.

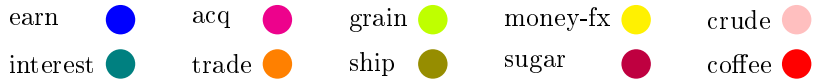
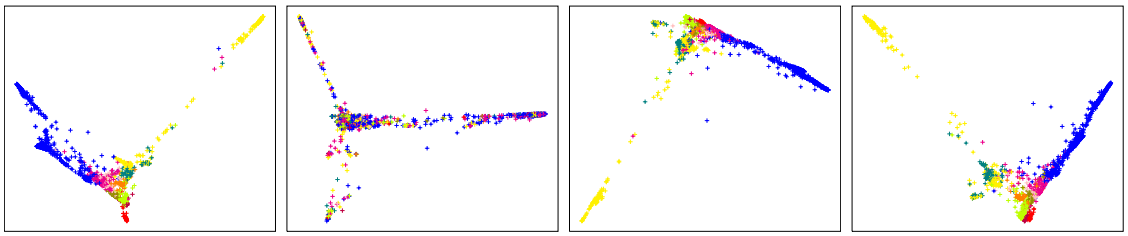


FIGURE 5.4 : Palette de couleurs pour les catégories du corpus.

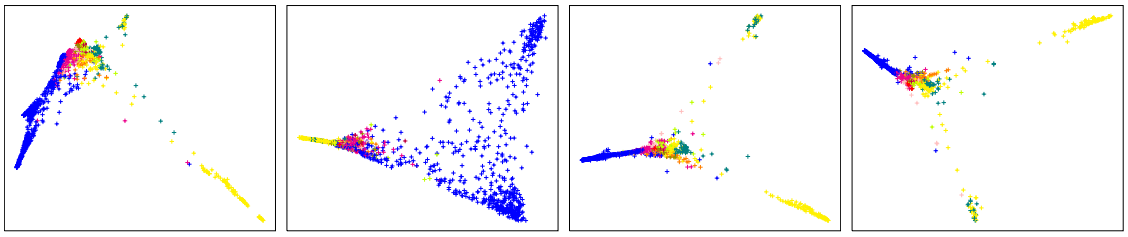
Les deux figures montrent un découpage assez grossier des documents, correspondant à une structure globale : d'une part, un ensemble de documents qui gravitent autour de la classe majoritaire, et d'autre part ceux de la classe *money-fx*. À une rotation/symétrie près, les projections pour `lk-slex` et `lk-text` sont très similaires à celles obtenues avec la vérité terrain. Cependant un rapprochement des documents dans l'angle central peut être observé. En comparaison avec les deux autres ressources, la projection avec la ressource `lk-free` ne montre aucune séparation entre les classes dans l'espace à 300 termes. En revanche, l'utilisation de 1 000 termes dans l'espace de représentation permet de séparer plus nettement la catégorie *earn*. Paradoxalement, 300 termes semblent plus discriminants des différentes classes avec `lk-slex` et `lk-text`.

La projection par t-SNE capture mieux la structure locale des données (*cf.* figures 5.7 et 5.8). L'impact du bruit sur le voisinage local semble moins important que sur les figures 5.6(b) et 5.5(b). Les erreurs de reconnaissance ont ici tendance à réduire la distance entre les classes plutôt qu'entre les documents. Si, à quelques documents



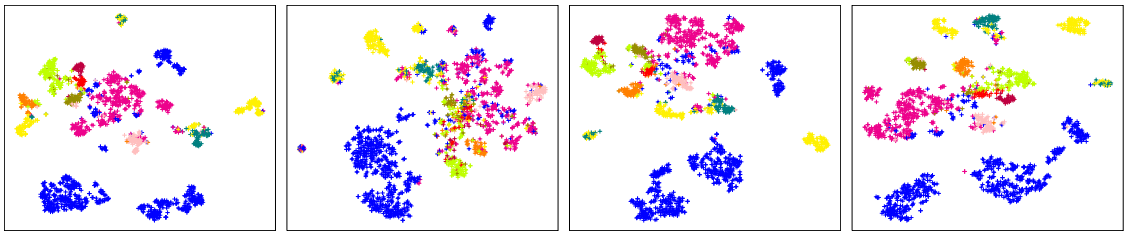
(a) Vérité terrain (b) `lk-free` (c) `lk-slex` (d) `lk-text`

FIGURE 5.5 : Projection dans l'espace propre de la matrice laplacienne avec 300 termes.



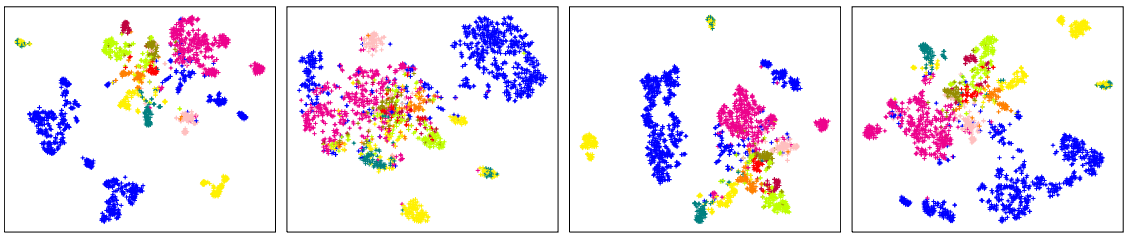
(a) Vérité terrain (b) `lk-free` (c) `lk-slex` (d) `lk-text`

FIGURE 5.6 : Projection dans l'espace propre de la matrice laplacienne avec 1 000 termes.



(a) Vérité terrain (b) `lk-free` (c) `lk-slex` (d) `lk-text`

FIGURE 5.7 : Projection par t-SNE avec 300 termes.



(a) Vérité terrain (b) `lk-free` (c) `lk-slex` (d) `lk-text`

FIGURE 5.8 : Projection par t-SNE avec 1 000 termes.

près, la séparation entre les différentes classes est très nette pour la vérité terrain, cette frontière diminue en fonction de la qualité de la reconnaissance. Des frontières assez nettes peuvent cependant être observées pour certains agglomérats de documents, y compris pour les documents de `lk-free`.

5.5 Biais introduit dans la phase d'apprentissage

Le biais introduit dans la phase d'apprentissage découle directement de l'impact précoce présenté dans la section précédente. Dans le cas des SVM, l'altération du modèle d'apprentissage peut être mesuré par le nombre de vecteurs de support. Il peut être montré que s'il existe une frontière qui sépare les documents d'entraînement sans erreur, l'espérance mathématique de la probabilité de commettre une erreur lors de la phase de test est bornée par (Cortes et Vapnik, 1995) :

$$\mathbb{E}\left[Pr(erreur)\right] \leq \frac{\mathbb{E}\left[\text{Nombre de vecteurs de support}\right]}{\text{Nombre d'exemples d'entraînement}} \tag{5.6}$$

D'après la formule 5.6, l'espérance est minimisée pour un faible nombre de vecteurs de support. L'augmentation du nombre de vecteurs de support, au sein du même système, peut indiquer une altération du modèle de catégorisation, même s'il a été mis en évidence que des systèmes qui génèrent moins de vecteurs de support, pour les mêmes données d'entraînement, ne vérifient pas toujours l'équation 5.6 (Burges, 1998).

Le tableau 5.7 montre que la qualité des documents influe sur le nombre de vecteurs de support. L'augmentation du TER s'accompagne d'une augmentation du nombre de vecteurs de support (à l'exception des catégories *money-fx* et *coffee* pour la ressource `lk-slex`).

TABLE 5.7 : Nombre de vecteurs support par catégorie.

	Ref.	`lk-free`	`lk-slex`	`lk-text`
earn	277	423	310	301
acq	345	517	377	362
grain	225	336	254	246
money-fx	245	338	241	248
crude	201	262	224	223
interest	165	226	185	186
trade	162	228	180	180
ship	202	246	218	210
sugar	111	205	139	115
coffee	177	197	166	179

Afin de vérifier dans quelle mesure l'altération du modèle influe sur la catégorisation, la capacité de généralisation des modèles bruités va être évaluée par validation croisée à 10 partitions (Hastie et collab., 2008, pp. 241-245). Les 10 partitions ont été choisies

aléatoirement et reproduites à l'identique pour toutes les expériences de façon à ce que les résultats soient comparables.

Le tableau 5.8 montre la précision, le rappel et le taux de classification pour l'algorithme des k-PPV. La ressource `lk-free` montre une baisse de près de 10 % du taux de classification en comparaison avec la vérité terrain. Les deux autres ressources obtiennent des performances très proches, la capacité de généralisation reste préservée. Cependant des variations plus ou moins importantes de la précision et du rappel selon les catégories sont observées. Cela montre les effets du changement du voisinage local des documents. En effet, certains documents « s'éloignent » de la catégorie prédite avec la vérité terrain. Pour la catégorie *coffee* par exemple, cela se traduit par une baisse du rappel, alors que la précision ne change pas en fonction des ressources.

TABLE 5.8 : k-PPV. Précision (π), rappel (ρ) et taux de classification (Moy$^{\mu}$.) pour l'ensemble d'entraînement avec validation croisée à 10 partitions.

	Vérité terrain		`lk-free`		`lk-slex`		`lk-text`	
	π	ρ	π	ρ	π	ρ	π	ρ
earn	99,44 %	96,19 %	90,46 %	91,69 %	99,43 %	94,69 %	97,93 %	96,59 %
acq	90,71 %	95,73 %	77,56 %	87,69 %	87,87 %	96,48 %	90,21 %	94,97 %
grain	90,58 %	89,93 %	78,23 %	82,73 %	89,44 %	91,37 %	89,58 %	92,81 %
money-fx	85,59 %	95,61 %	79,07 %	82,93 %	85,59 %	92,68 %	85,71 %	93,66 %
crude	87,50 %	82,35 %	84,93 %	72,94 %	89,04 %	76,47 %	88,89 %	75,29 %
interest	84,93 %	71,26 %	69,57 %	55,17 %	74,42 %	73,56 %	81,48 %	75,86 %
trade	86,57 %	87,88 %	83,33 %	60,61 %	91,80 %	84,85 %	96,30 %	78,79 %
ship	81,97 %	80,65 %	78,85 %	66,13 %	86,44 %	82,26 %	80,00 %	77,42 %
sugar	94,44 %	91,89 %	94,44 %	45,95 %	94,44 %	91,89 %	94,59 %	94,59 %
coffee	100,00 %	97,06 %	100,00 %	91,18 %	100,00 %	94,12 %	100,00 %	94,12 %
Moy$^{\mu}$.	92,85 %		83,70 %		91,93 %		92,31 %	

Le même comportement est observé avec les SVM, les performances de ces derniers étant cependant supérieures à celles de l'algorithme des k-PPV. Les erreurs de reconnaissance ont un impact certain sur les performances, mais à ce stade, une corrélation entre qualité de la reconnaissance et performances de la catégorisation ne peut être établie.

Ces résultats montrent que le biais introduit dans la phase d'apprentissage pour les ressources `lk-slex` et `lk-text` n'est pas, dans un premier temps, particulièrement dommageable pour les performances. Cependant, il introduit une certaine imprévisibilité dans le comportement des algorithmes ainsi qu'une non-prédictibilité de l'issue du processus, à la lumière des mesures du bruit et de l'impact précoce.

5.6 Impact sur la catégorisation

L'impact sur la catégorisation peut être mesuré de deux façons. La première consiste à mesurer la dégradation des performances suite à la perte de termes. Pour cela, un ensemble d'entraînement électronique et des documents de test bruités sont utilisés. La

TABLE 5.9 : SVM. Précision (π), rappel (ρ) et taux de classification (Moy$^{\mu}$.) pour l'ensemble d'entraînement avec validation croisée à 10 partitions.

	Vérité terrain		`lk-free`		`lk-slex`		`lk-text`	
	π	ρ	π	ρ	π	ρ	π	ρ
earn	99,72 %	97,68 %	97,04 %	93,87 %	99,31 %	98,09 %	98,63 %	98,23 %
acq	93,10 %	98,24 %	82,41 %	92,96 %	93,30 %	97,99 %	93,92 %	96,98 %
grain	93,06 %	96,40 %	84,14 %	87,77 %	93,66 %	95,68 %	93,10 %	97,12 %
money-fx	90,91 %	97,56 %	85,02 %	85,85 %	90,09 %	97,56 %	90,09 %	97,56 %
crude	92,31 %	84,71 %	88,46 %	81,18 %	93,33 %	82,35 %	92,11 %	82,35 %
interest	94,37 %	77,01 %	80,00 %	73,56 %	93,15 %	78,16 %	93,15 %	78,16 %
trade	92,31 %	90,91 %	83,58 %	84,85 %	90,91 %	90,91 %	96,77 %	90,91 %
ship	86,67 %	83,87 %	86,54 %	72,58 %	86,21 %	80,65 %	83,33 %	80,65 %
sugar	94,59 %	94,59 %	96,30 %	70,27 %	97,14 %	91,89 %	97,06 %	89,19 %
coffee	100,00 %	97,06 %	93,75 %	88,24 %	100,00 %	97,06 %	100,00 %	97,06 %
Moy$^{\mu}$.	95,34 %		89,17 %		95,18 %		95,07 %	

seconde consiste à mesurer la dégradation lorsque les documents bruités sont utilisés aussi bien à l'entraînement qu'au test. Les deux premières parties de cette section s'attachent à l'étude de l'impact dans ces deux configurations. Enfin, une dernière partie tente de livrer une analyse individuelle des documents dont l'issue de la catégorisation change avec les erreurs de reconnaissance.

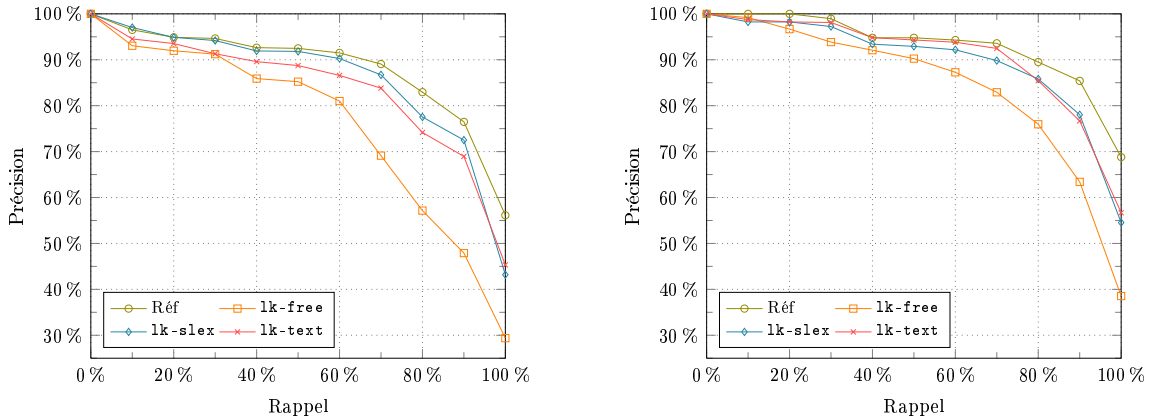
5.6.1 Impact avec entraînement électronique

La figure 5.9 montre les courbes de précision vs rappel, en macro-moyenne, pour les k-PPV et SVM. Les résultats présentés ont été obtenus avec l'ensemble d'entraînement de la vérité terrain. La dégradation des performances est donc due aux erreurs de reconnaissance dans l'ensemble de test. Ainsi, les faibles performances de la ressource `lk-free` apparaissent naturelles car celle-ci préserve peu le lexique d'origine. Il faut noter cependant que jusqu'à 30 % de rappel, ses performances restent proches de la référence, et ce, avec les deux algorithmes. La différence devient réellement visible pour des taux de rappel supérieurs à 30 %.

Pour l'algorithme des k-PPV avec la ressource `lk-text`, cette différence est marquée à partir de 50 % de rappel, tandis que pour `lk-slex`, elle ne l'est qu'à partir de 80 %. Étonnamment, la ressource avec les indicateurs du bruit les moins favorables obtient des meilleures performances.

Avec les SVM, `lk-slex` et `lk-text` montrent des performances très similaires, qui ne se distinguent de la référence qu'à partir de 80 % de rappel. Les tableaux 5.10 et 5.11 confirment ces premiers résultats.

Avec les k-PPV, la ressource `lk-slex` obtient les performances les plus proches de la référence. La baisse du taux de classification est d'environ 4 % avec `lk-text` et 8 % avec `lk-free`. Avec les SVM, la dégradation est de 5 % pour `lk-slex` et `lk-text`, et de 7 % pour `lk-free`.



(a) k-PPV (b) SVM

FIGURE 5.9 : Précision vs rappel avec ensemble d'entraînement électronique et documents de test manuscrit (macro-moyenne).

TABLE 5.10 : k-PPV. Précision (π), rappel (ρ) et taux de classification (Moy$^{\mu}$.) pour l'ensemble de test manuscrit avec jeu d'entraînement électronique.

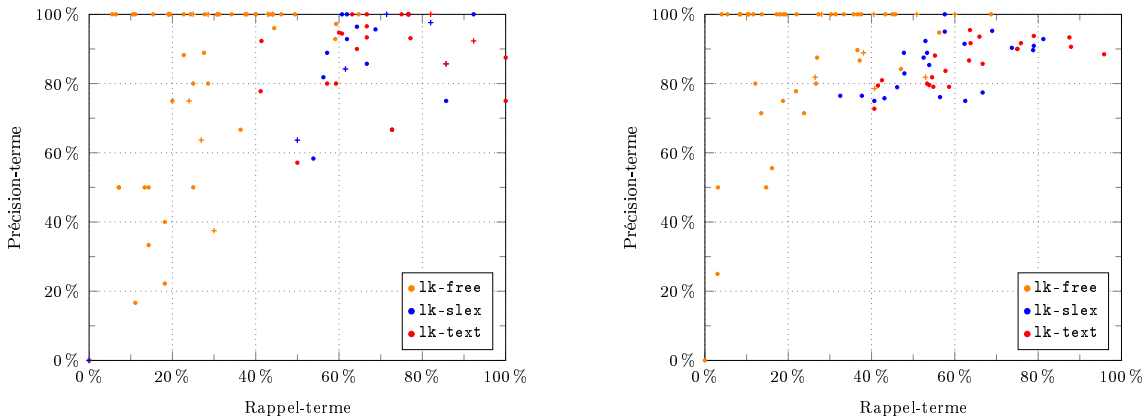
	Vérité terrain		`lk-free`		`lk-slex`		`lk-text`	
	π	ρ	π	ρ	π	ρ	π	ρ
earn	100,00 %	98,54 %	99,26 %	84,81 %	96,30 %	98,48 %	93,33 %	96,18 %
acq	95,12 %	92,86 %	86,59 %	79,78 %	93,90 %	88,51 %	93,90 %	84,62 %
grain	96,23 %	94,44 %	84,91 %	95,74 %	98,11 %	94,55 %	98,11 %	92,86 %
money-fx	75,93 %	75,93 %	55,56 %	73,17 %	77,78 %	77,78 %	74,07 %	75,47 %
crude	93,48 %	87,76 %	89,13 %	80,39 %	89,13 %	85,42 %	93,48 %	84,31 %
interest	66,67 %	80,00 %	60,00 %	78,26 %	66,67 %	74,07 %	60,00 %	75,00 %
trade	79,17 %	82,61 %	62,50 %	71,43 %	79,17 %	79,17 %	66,67 %	76,19 %
ship	55,56 %	90,91 %	50,00 %	75,00 %	61,11 %	91,67 %	61,11 %	91,67 %
sugar	100,00 %	84,62 %	63,64 %	87,50 %	81,82 %	81,82 %	81,82 %	81,82 %
coffee	100,00 %	76,92 %	100,00 %	76,92 %	100,00 %	76,92 %	100,00 %	76,92 %
Moy$^{\mu}$.	90,28 %		82,07 %		88,77 %		86,83 %	

TABLE 5.11 : SVM. Précision (π), rappel (ρ) et taux de classification (Moy$^{\mu}$.) pour l'ensemble de test manuscrit avec jeu d'entraînement électronique.

	Vérité terrain		`lk-free`		`lk-slex`		`lk-text`	
	π	ρ	π	ρ	π	ρ	π	ρ
earn	100,00 %	98,54 %	99,26 %	95,04 %	91,11 %	98,40 %	91,11 %	98,40 %
acq	93,90 %	92,77 %	91,46 %	84,27 %	93,90 %	87,50 %	93,90 %	85,56 %
grain	98,11 %	100,00 %	88,68 %	90,38 %	98,11 %	98,11 %	98,11 %	98,11 %
money-fx	85,19 %	83,64 %	59,26 %	74,42 %	81,48 %	83,02 %	81,48 %	83,02 %
crude	93,48 %	84,31 %	91,30 %	80,77 %	84,78 %	79,59 %	86,96 %	80,00 %
interest	73,33 %	78,57 %	73,33 %	81,48 %	73,33 %	84,62 %	73,33 %	78,57 %
trade	83,33 %	95,24 %	70,83 %	77,27 %	75,00 %	85,71 %	75,00 %	90,00 %
ship	61,11 %	84,62 %	55,56 %	62,50 %	55,56 %	37,04 %	61,11 %	47,83 %
sugar	100,00 %	100,00 %	54,55 %	85,71 %	81,82 %	100,00 %	81,82 %	100,00 %
coffee	100,00 %	83,33 %	100,00 %	71,43 %	100,00 %	83,33 %	100,00 %	83,33 %
Moy$^{\mu}$.	92,22 %		85,31 %		87,26 %		87,69 %	

Les différentes ressources peuvent favoriser ou défavoriser les différentes catégories. Dans certains cas, les performances par catégorie peuvent dépasser les performances de référence. Dans toutes les configurations, la même imprévisibilité que précédemment peut être observée (§5.5). Ceci est particulièrement visible pour la catégorie *ship*. Les ressources `lk-slex` et `lk-text` améliorent la précision et le rappel par rapport à la référence. Avec les SVM, les mêmes ressources dégradent de façon très importante le rappel alors que les mesures du bruit ne permettaient pas de le prédire.

La figure 5.10 montre le plan de recouvrement des données dont la catégorie change lorsque les données manuscrites sont utilisées. Il faut noter que les TR et TP sont calculés seulement sur les termes de l'espace de représentation car ce sont les seuls qui



(a) 300 termes

(b) 1000 termes

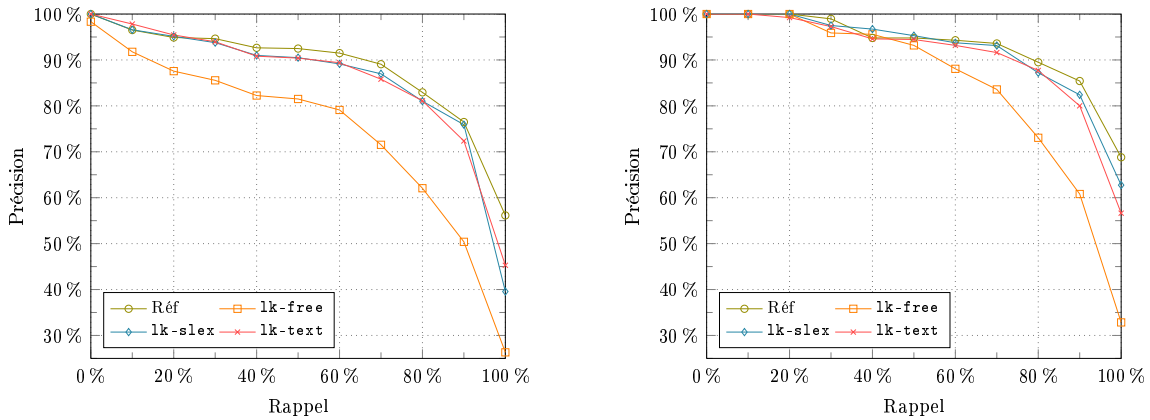
FIGURE 5.10 : Plan de recouvrement selon les différentes ressources et le nombre de termes de l'espace vectoriel.

influent sur la représentation des documents.

Cette figure montre que la précision-terme est haute pour la plupart des documents. Dans l'espace vectoriel à 1 000 termes, seulement 5 documents ont une TP inférieure à 70 %, malgré cela, la plupart des documents sont mal classés. La distribution du TR est, quant elle, plus étalée. Il faut noter également que certains documents sont mal classés malgré un TR et une TP élevés. De plus, avec l'algorithme des k-PPV, certains documents sont bien catégorisés dans leur version manuscrite alors qu'ils ne l'étaient pas dans la version originale. Cela montre également l'instabilité des résultats avec les documents bruités.

5.6.2 Impact avec entraînement bruité

Cette section présente les résultats des expériences menées avec des ensembles d'entraînement et de test issus de la reconnaissance. Les courbes de précision vs rappel (*cf.* figure 5.11) montrent que les performances de la ressource `lk-slex` et `lk-text` sont meilleures lorsque les documents d'entraînement bruités sont utilisés. En revanche, avec la ressource `lk-free`, elles sont encore plus dégradées. Cela s'explique par l'absence de contrainte lexicale dans cette ressource. En effet, la contrainte lexicale favorise la reconnaissance des termes du lexique dans les ensembles d'entraînement et test, ce qui n'est pas le cas de la ressource `lk-free`.



(a) k-PPV (b) SVM

FIGURE 5.11 : Précision vs rappel avec entraînement et test manuscrit (macro-moyenne).

Les tableaux 5.12 et 5.13 confirment également qu'il y a peu de différences entre les résultats de référence et ceux de `lk-slex` et `lk-text`. Avec les deux algorithmes, la différence au niveau du taux de classification est de moins de 1 %.

Ces résultats vont dans le sens des conclusions d'Ittner et collab. (1995) selon lesquelles il vaut mieux apparier des documents de même nature. Par ailleurs, cela correspond à un cadre applicatif plus proche de la réalité, où il faut catégoriser des documents manuscrits en se basant sur des documents eux-mêmes manuscrits.

TABLE 5.12 : k-PPV. Précision (π), rappel (ρ) et taux de classification (dernière ligne) avec entraînement et test manuscrit.

	Vérité terrain		`lk-free`		`lk-slex`		`lk-text`	
	π	ρ	π	ρ	π	ρ	π	ρ
earn	100,00 %	98,54 %	99,26 %	82,21 %	99,26 %	98,53 %	98,52 %	98,52 %
acq	95,12 %	92,86 %	86,59 %	73,96 %	95,12 %	95,12 %	93,90 %	89,53 %
grain	96,23 %	94,44 %	81,13 %	86,00 %	98,11 %	94,55 %	100,00 %	94,64 %
money-fx	75,93 %	75,93 %	57,41 %	70,45 %	74,07 %	78,43 %	72,22 %	76,47 %
crude	93,48 %	87,76 %	82,61 %	86,36 %	91,30 %	85,71 %	95,65 %	84,62 %
interest	66,67 %	80,00 %	50,00 %	68,18 %	70,00 %	72,41 %	63,33 %	70,37 %
trade	79,17 %	82,61 %	50,00 %	85,71 %	79,17 %	79,17 %	79,17 %	86,36 %
ship	55,56 %	90,91 %	38,89 %	77,78 %	61,11 %	78,57 %	55,56 %	90,91 %
sugar	100,00 %	84,62 %	45,45 %	62,50 %	81,82 %	81,82 %	81,82 %	90,00 %
coffee	100,00 %	76,92 %	100,00 %	76,92 %	100,00 %	83,33 %	100,00 %	76,92 %
Moy$^{\mu}$.	90,28 %		79,05 %		89,85 %		89,20 %	

TABLE 5.13 : SVM. Précision (π), rappel (ρ) et taux de classification (dernière ligne) avec entraînement et test manuscrit.

	Vérité terrain		`lk-free`		`lk-slex`		`lk-text`	
	π	ρ	π	ρ	π	ρ	π	ρ
earn	100,00 %	98,54 %	99,26 %	94,37 %	99,26 %	98,53 %	100,00 %	98,54 %
acq	93,90 %	92,77 %	92,68 %	80,00 %	95,12 %	89,66 %	95,12 %	90,70 %
grain	98,11 %	100,00 %	88,68 %	92,16 %	98,11 %	98,11 %	98,11 %	98,11 %
money-fx	85,19 %	83,64 %	74,07 %	75,47 %	90,74 %	84,48 %	85,19 %	85,19 %
crude	93,48 %	84,31 %	86,96 %	85,11 %	89,13 %	83,67 %	95,65 %	81,48 %
interest	73,33 %	78,57 %	63,33 %	90,48 %	70,00 %	84,00 %	73,33 %	84,62 %
trade	83,33 %	95,24 %	70,83 %	80,95 %	79,17 %	95,00 %	75,00 %	85,71 %
ship	61,11 %	84,62 %	44,44 %	66,67 %	61,11 %	78,57 %	55,56 %	90,91 %
sugar	100,00 %	100,00 %	63,64 %	87,50 %	81,82 %	100,00 %	81,82 %	100,00 %
coffee	100,00 %	83,33 %	100,00 %	76,92 %	100,00 %	83,33 %	100,00 %	83,33 %
Moy$^{\mu}$.	92,22 %		85,96 %		91,58 %		91,58 %	

Lorsque les documents bruités sont utilisés pour l'entraînement, le plan de recouvrement n'est pas pertinent. En effet, même si un document de test est parfaitement reconnu, rien ne garantit qu'il sera catégorisé de la même façon, et ce, à cause de la modification de la structure des données d'entraînement (§5.4.2).

En dernier lieu, une analyse qualitative des documents va être réalisée. Cette inspection du contenu des documents transcrits va permettre de mieux cerner pourquoi ils sont mal classés. Les résultats de cette analyse sont livrés dans la sous-section suivante.

5.6.3 Analyse qualitative des documents

L'écart entre les taux de classification de la référence et les ressources `lk-slex` et `lk-text` (tableaux 5.12 et 5.13) est dû à très peu de documents. En effet, dans le cas des

SVM, il s'agit de seulement 8 documents mal classés supplémentaires par rapport à la référence. Avec l'algorithme des k-PPV, sont décomptés 8 documents avec la ressource `lk-slex` et 11 avec `lk-text`.

La plupart de ces documents ont un TER inférieur à la médiane, avec les termes de l'espace vectoriel préservés. S'ils ne sont pas catégorisés de la même façon que leur version électronique c'est principalement en raison de la modification de la structure des données d'entraînement.

Les documents restants ont un TER inférieur ou égal à 56 %. Il s'agit de textes courts qui préservent peu les termes de l'espace vectoriel, ils ne préservent même parfois qu'un seul terme. Il apparaît naturel que la perte des termes de l'espace de représentation ait plus d'impact que celle des autres termes, cette hypothèse a été formulée dans un travail précédent (Peña Saldarriaga et collab., 2009b) mais elle n'a pas été pleinement validée expérimentalement.

L'analyse détaillée des documents n'a pas permis d'extraire des caractéristiques particulières pouvant déterminer l'issue du processus de catégorisation pour un document donné. La catégorisation devient incertaine lorsque les documents bruités sont utilisés. En effet, la catégorisation montre des variations, de nature presque aléatoire, plus ou moins importantes, des ensembles de documents correctement et mal classés. Il apparaît ici que les erreurs de reconnaissance engendrées avec les ressources `lk-slex` et `lk-text` introduisent des variations « systémiques », c'est-à-dire une modification implicite des stratégies de tokenisation et filtrage de mots outils par exemple. Il a été montré, par ailleurs, que l'utilisation de différents algorithmes de tokenisation ou de racinisation a une influence sur les résultats de la RI ad-hoc, ce qui a constitué un des points de départ des techniques de fusion de résultats en RI (voir chapitre 7).

5.7 Conclusion

Dans ce chapitre, l'influence du bruit sur une tâche de catégorisation a été étudiée. La première partie a tenté d'évaluer l'impact des erreurs de reconnaissance dans la représentation des documents. La seconde, quant à elle, a comparé les résultats de deux algorithmes de catégorisation, à savoir les k-PPV et les SVM. Les résultats ont été mesurés sur trois versions transcrites du corpus manuscrit ainsi que sur sa version électronique.

En comparant les performances obtenues avec les documents électroniques et les documents manuscrits, il a été constaté qu'il n'y a pas de baisse importante des performances pour les ressources `lk-slex` et `lk-text` : moins de 1 % pour un TER compris entre 23 et 26 %. Il a également été constaté qu'un TER trop élevé, d'environ 55 %, dégrade de façon considérable les performances de la catégorisation.

Il peut raisonnablement être estimé que pour des TER inférieurs à 23 %, peu de variations seront observées au moment de la catégorisation. Il est difficile d'évaluer le comportement des algorithmes de classification dans l'intervalle $]26\,\%, 55\,\%[$ sans des versions du corpus représentatives du continuum du TER dans cet intervalle.

Les expériences ont également montré la nature imprévisible, quasi-aléatoire, des performances de la catégorisation et, par conséquence, de la difficulté des méthodes

expérimentales à déterminer des indices prédictifs des baisses de performances attendues. Ces éléments conduisent à penser que le bruit devrait également être étudié dans une perspective théorique en le modélisant, par exemple, par un processus aléatoire qui pourrait être intégré directement dans les modèles de classification.

En parallèle de ce chapitre sur la catégorisation, le chapitre suivant ouvre le second volet de cette thèse, à savoir la recherche d'information dans des documents manuscrits en-ligne.

Chapitre 6

À la recherche du document manuscrit en-ligne

« Les gens qu'on interroge, pourvu qu'on les interroge bien, trouvent d'eux-mêmes les bonnes réponses. »

SOCRATE

Alors que les travaux s'intéressant à la recherche de documents électroniques existent depuis de nombreuses années, les travaux dans le domaine des documents manuscrits - y compris hors-ligne - sont relativement récents (Lopresti et Tomkins, 1994 ; Aref, Barabará et Vallabhaneni, 1995b ; Manmatha, Chengfeng, Riseman et Croft, 1996b ; Manmatha, Chengfeng et Riseman, 1996a ; Kwok, Perrone et Russell, 2000 ; Russell, Perrone et Chee, 2002 ; Jain et Namboodiri, 2003 ; Rath et Manmatha, 2003a,b ; Rath, Manmatha et Lavrenko, 2004 ; Vinciarelli, 2004 ; Srihari, Huang et Srinivasan, 2005 ; Vinciarelli, 2005a ; Schimke et Vielhauer, 2006 ; Cao et Govindaraju, 2007 ; Jawahar, Balasubramanian, Meshesha et Namboodiri, 2009 ; Terasawa et Tanaka, 2009 ; Perronnin et Rodriguez-Serrano, 2009 ; Cheng, Zhu, Chen et Nakagawa, 2009).

Jusqu'aux travaux récents de Vinciarelli (2004, 2005a), le paradigme dominant de la recherche de documents manuscrits était le *word spotting*. Les différentes méthodes utilisées peuvent alors être caractérisées par le niveau de représentation sur lequel elles se basent. Schématiquement, il est possible de classer ces méthodes selon qu'elles nécessitent une reconnaissance explicite ou pas. Cependant, il n'y a pas de distinction conceptuelle entre ces méthodes car le paradigme sous-jacent est le même : la recherche de mots clés. Depuis les travaux de Vinciarelli (2005a), il devient nécessaire de distinguer ce qui relève de la recherche de données et de la recherche d'information (*cf.* figure 6.1).

Les méthodes de word spotting sont des algorithmes de recherche de sous-chaînes, tels ceux de Aho et Corasick (1975) ou Boyer et Moore (1977), conçus pour le domaine

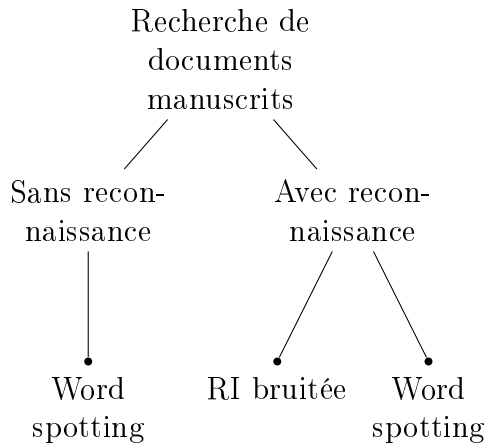


FIGURE 6.1 : Typologie des méthodes de recherche de documents manuscrits.

manuscrit. Leur but est d'identifier les documents qui contiennent des mots clés donnés. Au contraire, l'objectif d'un algorithme de recherche d'information est d'identifier des documents dont l'information peut être pertinente pour un utilisateur, sans pour autant utiliser les mots clés dans le document. C'est là que se trouve la différence conceptuelle entre la RI et le word spotting.

Dans le domaine en-ligne, le word spotting se justifie par la possibilité d'étendre les capacités de recherche du modèle à des dessins, mais surtout par l'idée selon laquelle l'encre numérique doit être considérée comme un type primitif (Aref, Barabará et Lopresti, 1995a ; Aref, Kamel et Lopresti, 1995c), c'est-à-dire un type de données qui peut être traité directement. Dans le domaine hors-ligne, il se justifie principalement par la difficulté à reconnaître des documents historiques (Manmatha et collab., 1996b,a). La figure 6.2 montre un exemple issu de la collection des lettres de George Washington [1] utilisé pour leurs expériences.

Toutefois, l'idée défendue dans cette thèse est que le word spotting ne doit plus être vu comme une fin mais comme un moyen pour la RI. Face à la masse de données hétérogènes qu'il est possible de rencontrer à l'heure actuelle, la RI doit être privilégiée dans tous les domaines où, soit la reconnaissance de l'écriture, soit le word spotting, sont suffisamment robustes.

Ce travail s'inscrit dans le cadre de la recherche d'information. Celle-ci peut être définie comme l'ensemble des méthodes et techniques permettant de « trouver des objets (principalement des documents) de nature non structurée (principalement textuelle) qui satisfont un besoin d'information » [2] à partir de collections de données (Manning, Raghavan et Schütze, 2008, chap. 1).

Ce chapitre a pour vocation de présenter ce qu'est la recherche d'information afin de permettre notamment la compréhension des chapitres suivants. Il se décompose en

1. Plus d'informations sur ces documents peuvent être trouvées à l'adresse suivante : http://gwpapers.virginia.edu/

2. *« Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need ».*

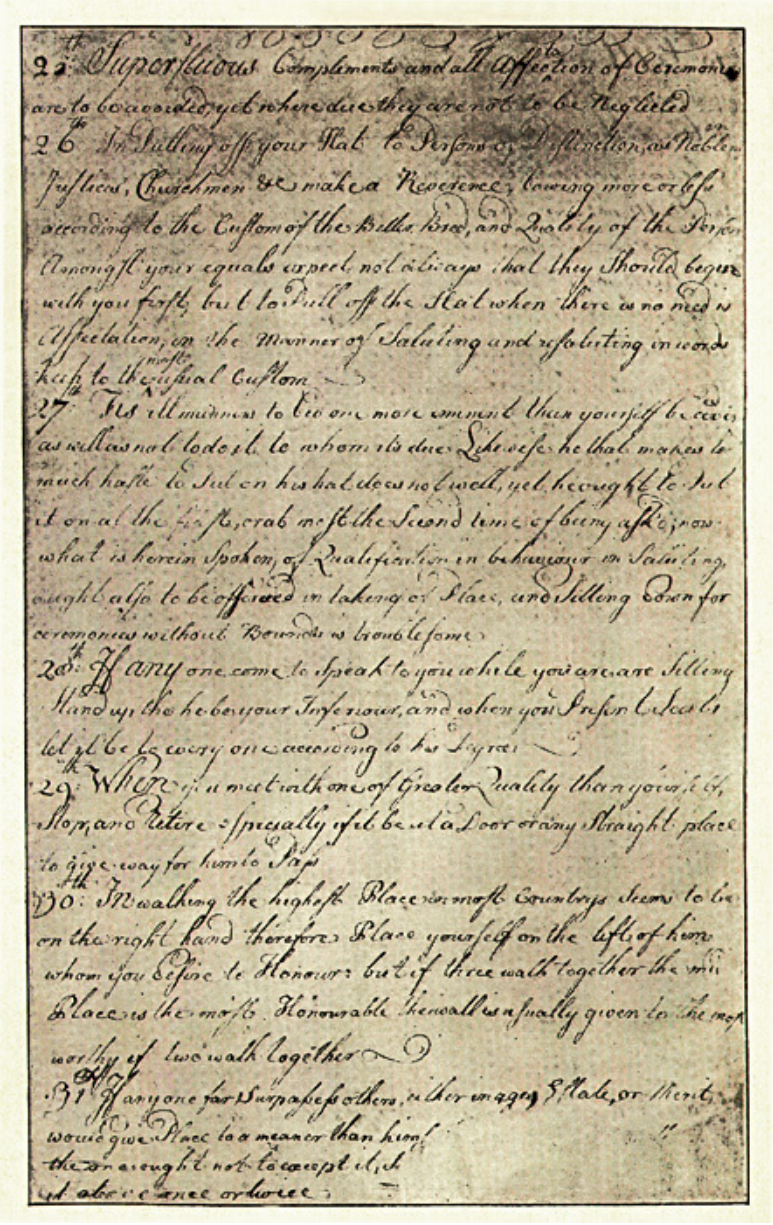


FIGURE 6.2 : Extrait des lettres de George Washington utilisé dans des expériences sur la recherche de documents historiques.

quatre sections. Une première aborde les concepts clés de la RI (§6.1), la deuxième décrit les principaux modèles de recherche d'information et de word spotting (§6.2), en effet la plupart de ces méthodes serviront de référence pour les expériences liées à la RI. La troisième section s'intéresse à l'évaluation de la RI, ce qui va permettre de mesurer l'adéquation entre le besoin d'information exprimé par l'utilisateur et l'ensemble des objets restitués par le système de RI. Enfin, la dernière section de ce chapitre présente les expériences validant la mise en œuvre des différentes méthodes de référence, et livre une première analyse de l'impact des erreurs de reconnaissance sur les performances de la RI.

6.1 Concepts

Dans un système de recherche d'information (SRI), un utilisateur exprime un besoin d'information sous la forme d'une courte phrase ou d'une séquence de mots clés. Le SRI retourne alors un ensemble de documents qu'il considère comme répondant à ce besoin. Dans ce cas, la RI est dite « ad-hoc ». Dans la suite de ce document, il sera fait référence à ce type de RI.

Requête, collection de documents et pertinence sont trois concepts clés de la recherche d'information. L'objectif de cette section est de définir ces trois concepts dans la perspective classique de la RI tout en les replaçant dans le contexte du word spotting.

6.1.1 Requête

Lorsqu'un utilisateur veut interroger un SRI, il doit produire une requête. Cela implique qu'il doit être capable de produire un ensemble restreint de mots clés qui « véhiculent la sémantique du besoin d'information »[3] (Baeza-Yates et Ribeiro-Neto, 1999, p. 4).

La *requête* constitue le point de départ de la RI dont la finalité est de retrouver, à l'intérieur d'une *collection*, les documents les plus *pertinents* vis-à-vis de la requête.

Dans le cas du word spotting, la requête est limitée à un seul motif qui peut se traduire par un mot manuscrit ou dactylographié, ou par une expression régulière. Le système agit alors comme un algorithme de recherche de sous-chaînes et retrouve les documents qui satisfont la contrainte imposée par le motif.

6.1.2 Collection de documents

Une collection de documents est une source d'information dans laquelle un utilisateur pourra satisfaire *son* besoin particulier. Cette source d'information est représentée de la même façon que dans la catégorisation (chapitre 4) à savoir par une matrice $\mathbf{A}$ de $n \times m$ éléments, où n est la taille de la collection, et m celle du lexique. Les étapes d'indexation restent identiques (§4.2), à l'exception de l'étape de sélection de termes qui n'a pas lieu dans les expériences liées à la RI. L'étape d'indexation dans le contexte de la RI inclut la création d'un *index terminologique* ou fichier inverse (Baeza-Yates et

3. *« convey the semantics of the information need ».*

Ribeiro-Neto, 1999, chap. 8). Il s'agit d'une mise en correspondance entre les termes et les documents dans lesquels ils apparaissent (*cf.* figure 6.3). Cela permet un accès rapide à un fichier spécifique à partir de ses termes sans avoir à parcourir séquentiellement l'ensemble des documents. Voilà qui appelle deux remarques.

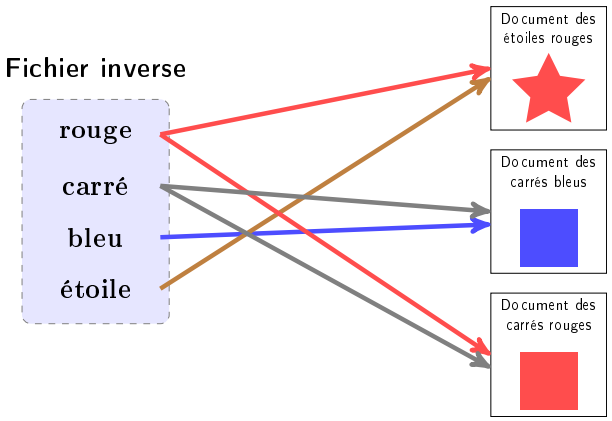


FIGURE 6.3 : Vision schématique d'un index terminologique.

D'une part, la représentation matricielle ne convient ni à tous les modèles de RI, ni aux méthodes de word spotting. Les représentations liées à un modèle particulier seront abordées au moment de sa description. La phase d'indexation dans le cas du word spotting correspond principalement à une segmentation des mots de l'image ou du tracé en ligne. Des tentatives pour imiter les mécanismes de racinisation et le filtrage de mots outils étaient jusqu'à très récemment inexistantes (Jawahar et collab., 2009).

D'autre part, sauf cas marginaux (Aref et collab., 1995b ; Jawahar et collab., 2009), l'étape de segmentation ne donne pas lieu à la création d'un ficher inverse ou d'une structure similaire. En effet, la complexité des algorithmes de word spotting reste dépendante de la taille de la collection.

Un problème que le word spotting et les modèles de RI ont en commun est la mise en correspondance entre un ensemble de documents et une requête. Cette mise en correspondance consiste à attribuer un score de confiance au mot clé, détecté dans un document pour le word spotting, et un score, reflétant la pertinence du document par rapport à la requête, pour les modèles de RI. Le concept de pertinence est défini ci-après tandis que les techniques de mise en correspondance sont décrites dans la section 6.2.

6.1.3 Pertinence

L'efficacité d'un SRI réside dans sa capacité à retrouver les documents pertinents pour le besoin d'information d'un utilisateur donné. « Un document est pertinent vis-à-vis d'un besoin d'information si et seulement si, il contient au moins une phrase qui répond à ce besoin »[4] (van Rijsbergen, 1979, p. 114). Cependant, la pertinence est une

4. « *A document is relevant to an information need if and only if it contains at least one sentence which is relevant to that need.* »

notion subjective. Pour un même besoin d'information, deux utilisateurs peuvent avoir des avis différents quant à la pertinence d'un ou plusieurs documents.

L'évaluation des modèles théoriques se fait à l'aide de corpus standardisés. Ces corpus sont composés d'un ensemble de requêtes, d'une collection de documents et de la liste de documents pertinents pour chacune des requêtes. Cela permet de faire abstraction du caractère subjectif de la pertinence, en s'attachant à un ensemble de jugements de référence fourni généralement par un panel d'experts (van Rijsbergen, 1979, p. 113).

À l'aide de ces corpus, l'évaluation revient à comparer les réponses d'un SRI avec les jugements de référence. Il faut remarquer que le corpus collecté dans le cadre de cette étude (chapitre 3) n'est pas un corpus de RI. Cependant, une méthode a été proposée afin de produire un ensemble de requêtes et de jugements de pertinence à partir des catégories associées par un expert à chacun des documents du corpus (§3.3). Ces jugements constituent la base de l'évaluation des méthodes de recherche de documents manuscrits en-ligne présentées dans la prochaine section.

6.2 Modèles pour la recherche de documents manuscrits en-ligne

Cette section présente différents modèles de référence aussi bien de la RI classique que du word spotting pour les documents en-ligne. Cette présentation n'a pas la prétention d'être exhaustive, elle se concentre sur des modèles classiques de la RI et sur ceux qui représentent les tendances actuelles du word spotting dans le domaine en-ligne.

6.2.1 Modèles de RI

Étant donnée une requête q, un modèle peut être représenté par une fonction $f_q : \mathcal{D} \rightarrow \mathbb{R}$. Le score associé à un document est appelé *retrieval status value* (rsv). Chaque modèle possède une méthode pour calculer l'ensemble $\tau \in \mathbb{R}^n$ des scores de la collection, ces scores déterminent l'ordre de présentation des résultats.

6.2.1.1 Modèle vectoriel

Dans le modèle vectoriel (Salton, 1968 ; Salton et collab., 1975), la requête est considérée comme un court document. Ainsi q et $\mathcal{D}$ sont représentés dans le même espace à m dimensions. Le rsv d'un document est calculé par le produit scalaire de $\mathbf{q}$ et $\mathbf{d}_j$ sur la norme euclidienne des vecteurs.

En considérant $\mathbf{q}$ et $\mathbf{A}$ normalisés au préalable, τ peut être calculé grâce à la formule suivante :

$$\tau = \mathbf{A}\mathbf{q} \tag{6.1}$$

6.2.1.2 Modèle probabiliste

Le modèle probabiliste (Spärck Jones, Walker et Robertson, 2000a,b) estime la probabilité que le document d_j soit pertinent pour la requête q. Cela revient à remplacer la pondération $tf \times idf$ (équation 4.1) par la pondération BM25, couramment appelée formule Okapi :

$$\mathbf{A}_{j,i} = \underbrace{\frac{\mathrm{tf}(i,j) \times (k+1)}{\mathrm{tf}(i,j) + k \times \left((1-b) + b\left(\frac{|d_j| \times n}{\sum_{x=1}^{n} |d_x|}\right)\right)}}_{\text{facteur } tf} \times \underbrace{\log\left(\frac{(n+0{,}5) - \mathrm{n}(i)}{0{,}5 + \mathrm{n}(i)}\right)}_{\text{facteur } idf} \tag{6.2}$$

Dans la formule 6.2, k et b sont des hyperparamètres habituellement fixés à 2 et 0,75 respectivement. $|d_j|$ est la longueur du document en nombre de termes.

En considérant, encore une fois, que $\mathbf{q}$ et $\mathbf{A}$ ont été préalablement normalisés, le calcul de τ s'effectue grâce à la formule 6.1.

6.2.1.3 Modèles de langage

La modélisation probabiliste du langage est particulièrement appliquée au domaine de la recherche d'information depuis les travaux de Ponte et Croft (1998). Un document est représenté par une distribution multinomiale estimée à partir des occurrences des termes d'indexation, c'est-à-dire son modèle unigramme de langage, θ_d.

Une manière d'estimer θ_d consiste à appliquer le principe de maximum de vraisemblance (*maximum likelihood estimate*) :

$$\theta_d = \frac{\mathbf{d}}{\|\mathbf{d}\|_1} \tag{6.3}$$

L'inconvénient de cette estimation est qu'elle n'attribuera de probabilité qu'aux termes présents dans le document. Cela pose un problème lorsqu'un terme de la requête n'y apparaît pas. Il existe des techniques de lissage permettant de remédier à ce problème (Zhai et Lafferty, 2001).

Calculer le rsv revient à mesurer une vraisemblance entre le modèle de la requête θ_q et celui du document θ_d. En choisissant la divergence de Kullback et Leibler (1951) comme distance, et selon l'intuition que plus la distance est importante, moins le document est pertinent, le rsv d'un document donné est calculé grâce à la formule suivante (Lafferty et Zhai, 2001) :

$$\mathrm{rsv}(q,d) = -\sum_{w \in q} \mathrm{p}(w|\theta_q) \log \mathrm{p}(w|\theta_d) \tag{6.4}$$

En supposant que $\mathbf{d}_j$ est le vecteur des fréquences des termes, l'équation 4.1 peut être remplacée par :

$$\mathbf{A}_{j,i} = \frac{\mathrm{tf}(i,j)}{\|\mathbf{d}_j\|_1} \tag{6.5}$$

De plus, en considérant que le vecteur de la requête $\mathbf{q}$ a été normalisé par sa norme L_1, le calcul de τ s'effectue grâce à la formule suivante :

$$\tau = -(\log \mathbf{A})\mathbf{q} \tag{6.6}$$

6.2.2 Bilan des recherches sur la RI bruitée

La RI bruitée a toujours été assimilée à la RI dans des collections de documents issus d'un processus de reconnaissance de la parole mais surtout, de reconnaissance optique de caractères (OCR) (Voorhees et Harman, 2005, p. 11). Le travail de la RI bruitée à partir de sources manuscrites se distingue de cette conception classique de la RI bruitée pour essentiellement deux raisons.

La première tient à l'étude de l'impact du bruit dans des approches booléennes de la RI (Taghva, Borsack, Condit et Erva, 1994 ; Lopresti et Zhou, 1996 ; Lopresti, 1996), conceptuellement proches du word spotting (Vinciarelli, 2006). Elle tient également à l'utilisation de corpus électroniques dégradés artificiellement (Croft, Harding, Taghva et Borsack, 1994 ; Schäuble et Glavitsch, 1994 ; Lopresti et Zhou, 1996), les résultats de ces expériences ayant alors une valeur relative (Mittendorf et Schäuble, 1996).

La seconde raison est liée aux CER habituels de l'OCR. En effet, les meilleurs systèmes montrent des CER inférieurs à 1 %. En ce qui concerne les collections de RI bruitée existantes, le CER observé est de l'ordre de 5 % seulement (Kantor et Voorhees, 2000) bien qu'avec une diminution volontaire de la résolution de l'image à reconnaître, ou l'injection artificielle d'erreurs, ils puissent atteindre 20 % (Kantor et Voorhees, 2000) ou plus (Croft et collab., 1994 ; Lopresti et Zhou, 1996). Si 5 % peut être un taux d'erreur représentatif de ceux rencontrés dans les applications réelles de l'OCR (Voorhees et Harman, 2005, chap. 8), il est clairement inférieur à ceux habituellement rencontrés dans le domaine de la reconnaissance de l'écriture où le CER se situe entre 10 % et 20 % selon les lexiques et les modèles de langage utilisés (Perraud, 2005).

La conséquence principale de cette distinction est que les conclusions tirées des études utilisant des corpus issus de l'OCR, ou d'une dégradation artificielle, ne sont pas directement extrapolables à la problématique de ce travail. Ces conclusions peuvent se résumer au titre d'une publication de Mittendorf et Schäuble (2000) : « *Information Retrieval can Cope with Many Errors* ». Avec ce léger bémol cependant : les erreurs ont très peu d'effet sur les résultats de la RI, tant que la taille des documents est suffisamment importante pour que la redondance permette de contrebalancer une partie des erreurs de reconnaissance.

Par conséquent, une partie de ce travail s'attache à étudier l'impact des erreurs de reconnaissance afin de vérifier la validité des conclusions précitées dans le domaine en-ligne (§ 6.4).

6.2.3 Word spotting

Étant donnée un motif κ, un algorithme de word spotting peut être représenté comme une fonction $f_\kappa : \mathcal{D} \rightarrow \mathbb{R}^n$. Le résultat de cette fonction est l'ensemble Υ des scores associés à chacune des occurrences du motif dans un document. Chaque méthode

propose une façon différente de calculer Υ. Chaque méthode se base également sur un niveau de représentation abstrait du signal différent (§2.2.1). Cette sous-section s'intéresse à trois modèles appliqués au domaine en-ligne : (1) un modèle de word spotting au niveau des points normalisés de l'encre numérique (Jain et Namboodiri, 2003 ; Namboodiri, 2004, pp. 109-114) ; (2) un modèle de word spotting au niveau textuel, c'est-à-dire des mots reconnus (Kwok et collab., 2000 ; Russell et collab., 2002) ; (3) et un moteur industriel de word spotting.

6.2.3.1 Word spotting au niveau point

La méthode proposée par Namboodiri (2004) se décompose en trois étapes décrites ci-dessous.

Segmentation en mots

Afin d'effectuer la détection du motif, les documents doivent être préalablement segmentés en lignes puis en mots. La localisation des lignes d'écriture procède par une analyse de l'histogramme des projections sur l'axe des y (Ratzlaff, 2000). L'ordonnancement temporel du tracé est utilisé, afin de regrouper correctement les traits d'écriture s'étendant au-delà des limites estimées des lignes.

À partir des lignes, les mots sont également segmentés par une analyse d'histogramme, mais cette fois-ci, la projection s'effectue sur l'axe des x. Le processus de segmentation se termine par un ensemble de traitements : (1) ré-échantillonnage des points afin de s'affranchir de la vitesse d'écriture en les rendant équidistants ; (2) lissage des traits d'écriture par un filtre gaussien passe-bas ; (3) ré-échantillonnage à nouveau.

Extraction des caractéristiques

Namboodiri (2004) propose de représenter l'écriture en-ligne par des courbes polynomiales paramétriques de degré 3 (splines cubiques). Cela permet d'approcher les contours complexes de l'écriture et de s'assurer que l'approximation passe par tous les points de la séquence originale.

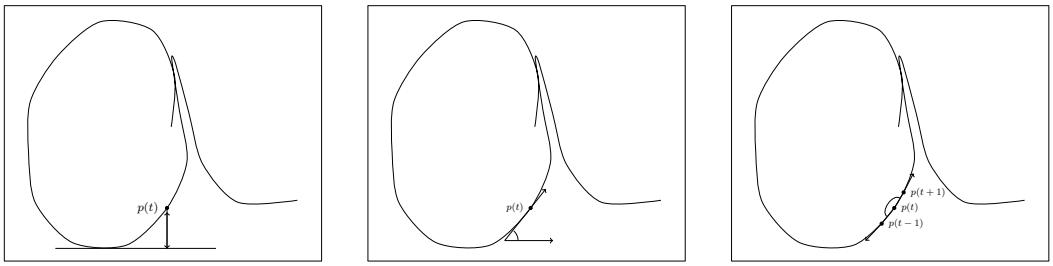
À partir de cette représentation, chaque point du tracé est décrit par un vecteur à 3 caractéristiques (*cf.* figure 6.4) :

1. La hauteur du point, c'est-à-dire la distance par rapport au point le plus bas.

2. La direction au point $\mathrm{p}(t)$ caractérisée par l'angle anti-horaire entre la tangente à $\mathrm{p}(t)$ et l'axe horizontal $\mathrm{x}(t)$.

3. La courbure au point $\mathrm{p}(t)$ caractérisée par l'angle anti-horaire à l'intersection entre le segment $[\mathrm{p}(t-1), \mathrm{p}(t)]$ et $[\mathrm{p}(t), \mathrm{p}(t+1)]$.

Aussi bien les mots du corpus que les motifs de recherche sont alors représentés par la séquence de vecteurs caractéristiques de chacun de leurs points. Cette méthode accepte seulement des requêtes sous forme manuscrite.



(a) Hauteur (b) Direction (c) Courbure

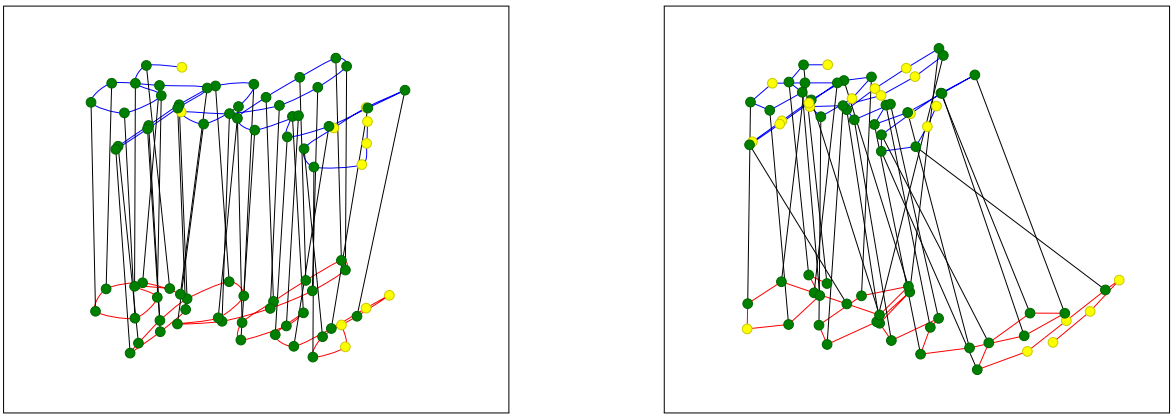
FIGURE 6.4 : Caractérisation des points d'un tracé en-ligne.

Mise en correspondance

La mise en correspondance entre le motif κ et les mots de la base s'effectue grâce à une distance élastique ou algorithme de déformation temporelle dynamique (Sakoe et Chiba, 1978) (Dynamic Time Warping, DTW). L'algorithme DTW cherche un alignement entre les séquences qui minimise une fonction de coût locale en tenant compte des compressions et extensions temporelles (*cf.* figure 6.5). La fonction de coût locale est donnée par la distance euclidienne pondérée des vecteurs caractéristiques de deux points a et b :

$$\mathrm{d}(a,b) = \sum_{i \in \{h,dir,courb\}} w_i \times (a_i - b_i)^2 \tag{6.7}$$

Avant l'application de l'algorithme d'alignement, les mots du document sont mis à l'échelle afin qu'ils soient de la même hauteur que le motif. Une translation est également effectuée pour que mots et motif aient le même barycentre. La recherche s'effectue par comparaison du motif avec chacun des mots de chacun des documents.



(a) $\mathrm{d}(split, split) = 0{,}12$ (b) $\mathrm{d}(split, said) = 0{,}65$

FIGURE 6.5 : Alignement entre deux paires de mots par déformation temporelle dynamique. Les points alignés sont signalés en vert.

6.2.3.2 Word spotting au niveau textuel

Russell et collab. (2002) proposent une méthode basée sur une reconnaissance des documents, avec prise en compte des candidats-mots à la reconnaissance. Cette méthode accepte des requêtes manuscrites ou dactylographiées. Dans le cas des requêtes manuscrites, la requête subit un processus de reconnaissance.

Reconnaissance des documents

La première étape de cette méthode consiste à effectuer la reconnaissance des documents (Subrahmonia, Nathan et Perrone, 1996). Les documents sont représentés par les listes de n-best correspondant à chaque mot manuscrit (*cf.* figure 6.6). Si la requête est de type manuscrit, elle est représentée de la même façon. Chaque liste de n-best peut être considérée comme un document de longueur n.

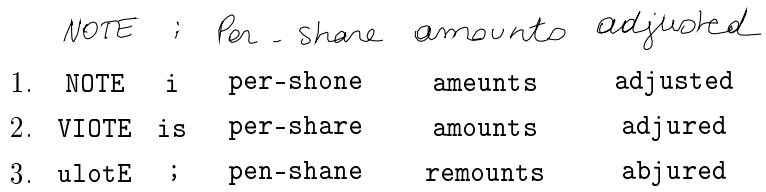


FIGURE 6.6 : Exemple de reconnaissance avec liste des n-best.

Mise en correspondance

La mise en correspondance se fait grâce à des mesures de similarité entre la requête et les listes de n-best candidats. Cette similarité peut être obtenue grâce à un score de corrélation, à la mesure cosinus, et même en utilisant le modèle probabiliste décrit précédemment (Kwok et collab., 2000).

6.2.3.3 MyScript® InkSearch®

MyScript® InkSearch® est une extension du SDK MyScript® Builder (§ 2.3) pour la recherche de mots clés dans des documents manuscrits. Ce système se base sur le moteur de MyScript® Builder, par conséquent il attend que des requêtes dactylographiées lui soient fournies. La recherche ne se limite pas à des simples mots clés car MyScript® InkSearch® permet également la recherche d'expressions régulières ou booléennes.

L'étape d'indexation avec MyScript® InkSearch® correspond à la création d'un index par document. La recherche s'effectue en parcourant chacun des index créés précédemment. Lors de la recherche, chaque occurrence du motif recherché se voit attribuer un indice de confiance. Cet indice reflète la similarité entre le motif et l'occurrence comme le montre la figure 6.7.

MyScript® InkSearch® a été choisi comme système de référence du word spotting dans les expériences réalisées dans le cadre de ce travail. En effet il s'agit de l'outil de word spotting développé par le partenaire industriel et livré avec leur SDK.

6.2.3.4 RSV et word spotting

Le word spotting a été défini par une fonction $f_\kappa : \mathcal{D} \rightarrow \mathbb{R}^n$. Puisque f_κ est définie de $\mathcal{D}$ dans $\mathbb{R}^n$, le résultat de la fonction ne peut pas être interprété comme un rsv. Comme le word spotting réalise une recherche de motifs, il ne peut calculer de rsv que pour une occurrence d'un motif donné. Cela apparaît clairement lorsque la fonction est définie du domaine des motifs κ dans celui des réels : $f_\kappa : \kappa \rightarrow \mathbb{R}$.

L'absence de rsv rend impossible la comparaison des méthodes venant du domaine de la RI et celles venant du domaine de la reconnaissance de formes. En effet, ces dernières doivent être évaluées par leur capacité à retrouver des motifs et non pas des documents. Ainsi, il est proposé de définir un rsv à partir de f_κ en additionnant les scores pour chaque occurrence de chacun des motifs composant la requête.

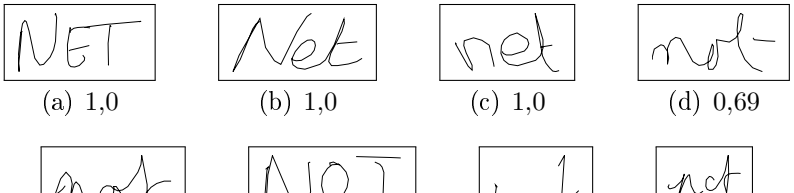
Soit la requête $q = \{\kappa_1, \kappa_2, \ldots, \kappa_n\}$ composée de n motifs, et Υ^i l'ensemble des scores des m occurrences du motif κ_i dans un document d, le rsv de d peut être calculé grâce à la formule suivante :

$$\mathrm{rsv}(q, d) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \Upsilon^i(j) \tag{6.8}$$

C'est cette formule qui est utilisée dans toutes les expériences présentées dans ce chapitre et les suivants. Elle est indépendante du système de word spotting utilisé. Elle suppose que les indices de confiance des occurrences soient comparables entre eux. En effet, il semble raisonnable de faire l'hypothèse que la même stratégie de recherche a été utilisée pour tous les documents de la collection.

6.3 Évaluation de la RI

Plusieurs mesures standard en RI peuvent être utilisées pour évaluer les performances des SRI. La précision, le rappel, la F-mesure et la précision à n documents ont



(a) 1,0 (b) 1,0 (c) 1,0 (d) 0,69

(e) 0,37 (f) 0,36 (g) 0,35 (h) 0,28

FIGURE 6.7 : Résultats de recherche pour le motif `net` et indices de confiance donnés par MyScript® InkSearch®.

déjà été présentés (§4.3).

En dehors de ces mesures, la mesure standard utilisée dans les campagnes d'évaluation de la RI est la moyenne des précisions moyennes (Mean Average Precision, MAP). Elle évalue la capacité d'un SRI à retrouver les documents pertinents rapidement, tout en intégrant la notion de rappel. Sa stabilité a également été démontrée (Buckley et Voorhees, 2000).

La définition des mesures ci-dessous adopte la notation introduite en §4.3. Bien entendu, la notion de catégorie est remplacée par celle de requête.

Définition 6.1. *Pour une requête q, la* précision moyenne *est la somme des précisions non interpolées sur l'ensemble des documents pertinents.*

$$\overline{\pi}(q) = \frac{1}{|R|} \sum_{i=1}^{|S|} \mathrm{p@}i(q) \times \mathrm{rel}(i) \tag{6.9}$$

Dans la formule 6.9, p@$i(q)$ est la précision à i documents et $rel(i)$ est la fonction indicatrice de la pertinence de i :

$$\mathrm{rel}(i) = \begin{cases} 1, & \text{si } i \in R \\ 0, & \text{sinon} \end{cases} \tag{6.10}$$

Comme $rel(i)$ annule la précision pour les documents non pertinents, les documents pertinents arrivés en tête de liste ont plus de poids sur la moyenne. En utilisant le nombre attendu de documents pertinents $|R|$ pour la moyenne, cette mesure introduit une notion de rappel.

À l'instar de la catégorisation, l'évaluation des SRI s'effectue sur un ensemble de requêtes Q. Dans ce cas, une valeur moyenne des différentes mesures doit être calculée. En RI, la moyenne des mesures sur un ensemble de requêtes est toujours la macro-moyenne.

Définition 6.2. *La* MAP *est la moyenne arithmétique des précisions moyennes.*

$$\mathrm{MAP}(Q) = \frac{1}{|Q|} \sum_{q \in Q} \overline{\pi}(q) \tag{6.11}$$

Les mesures dites de haute précision sont couramment employées pour l'évaluation des SRI. Il s'agit de la précision à n documents avec $n \in \{5, 10, 15, 20, 25, 30\}$.

Dans le cadre de cette thèse, l'évaluation de la RI utilise `trec_eval` (Voorhees et Harman, 2005, chap. 3). Il s'agit du logiciel d'évaluation utilisé dans les campagnes TREC. Il permet de calculer toute une batterie de mesures de la qualité d'un SRI. Dans cette étude la MAP et les mesures de haute précision seront privilégiées lors de la présentation des résultats des expériences.

6.4 Expériences préliminaires

Les premières expériences concernant la RI ont été effectuées avec quatre méthodes de référence sur le sous-ensemble $\mathcal{T}$ du corpus avec les requêtes prototypiques générées pour chacune de catégories (§3.3). Ces requêtes sont rappelées dans le tableau ci-dessous.

TABLE 6.1 : Requêtes générées pour les 10 catégories représentées dans le corpus manuscrit.

Catégorie	Requête
earn	vs ct net shr loss
acq	acquir stake acquisit complet merger
grain	tonn wheat grain corn agricultur
money-fx	stg monei dollar band bill
crude	oil crude barrel post well
interest	rate prime lend citibank percentag
trade	surplu deficit narrow trade tariff
ship	port strike vessel hr worker
sugar	sugar raw beet cargo kain
coffee	coffe bag ico registr ibc

L'ensemble de ces requêtes est soumis aux différentes méthodes de référence. Les 4 méthodes considérées ici et dans le chapitre suivant sont :

- le modèle vectoriel, abrégé VSM ;
- le modèle probabiliste, abrégé BM25 ;
- les modèles de langage, abrégés LM ;
- MyScript® InkSearch®, abrégé IS.

Les méthodes de RI Bruitée se basent sur les trois versions transcrites du corpus manuscrit utilisées également pour la catégorisation, tandis que IS se base directement sur le tracé manuscrit.

Les tableaux 6.2 à 6.4 présentent les résultats individuels pour chacune des trois méthodes de RI bruitée. Les résultats pour IS sont données dans la dernière colonne de chacun des tableaux.

L'impact des erreurs de reconnaissance dans la précision moyenne par catégorie ainsi que dans la MAP est évident. Selon la qualité du corpus et la méthode utilisée, une baisse de la MAP entre 3 et 10 points peut être observée. Les méthodes VSM et BM25 obtiennent des performances très similaires pour chacune des colonnes correspondant au corpus manuscrit. La MAP obtenue avec la méthode LM est très basse comparée aux autres en particulier parce qu'elle ne trouve aucun document pertinent pour la catégorie *trade*. Il faut également remarquer que la MAP obtenue avec IS est très similaire à celle des méthodes VSM et BM25 avec les ressources `lk-slex` et `lk-text`.

Un autre point important c'est la précision moyenne de la catégorie *earn* avec la ressource `lk-free` dans les tableaux 6.2 et 6.4, ainsi qu'avec MyScript® InkSearch®.

TABLE 6.2 : Précision moyenne par catégorie et MAP pour le modèle vectoriel (VSM).

	Réf.	lk-text	lk-slex	lk-free	IS
earn	0,9354	0,8108	0,8418	0,8979	0,8913
acq	0,6648	0,6474	0,6452	0,2452	0,6503
grain	0,9383	0,9255	0,9193	0,6818	0,9047
money-fx	0,4521	0,4301	0,4369	0,4017	0,5214
crude	0,9224	0,8968	0,8852	0,8004	0,035
interest	0,4379	0,4279	0,4378	0,1995	0,3647
trade	0,2853	0,2824	0,2873	0,1391	0,2282
ship	0,2433	0,2043	0,2177	0,1441	0,840
sugar	0,8934	0,8235	0,7641	0,5029	0,8593
coffee	0,9422	0,9200	0,8637	0,8574	0,8497
MAP	0,6715	0,6369	0,6299	0,4870	0,6357

TABLE 6.3 : Précision moyenne par catégorie et MAP pour le modèle probabiliste (BM25).

	Réf.	lk-text	lk-slex	lk-free	IS
earn	0,8351	0,8099	0,8414	0,7810	0,8913
acq	0,5971	0,6101	0,6501	0,2283	0,6503
grain	0,9401	0,9264	0,9188	0,6586	0,9047
money-fx	0,4531	0,4236	0,4365	0,3916	0,5214
crude	0,9166	0,8932	0,8898	0,8063	0,9035
interest	0,4332	0,4356	0,4382	0,1933	0,3647
trade	0,3468	0,3265	0,3408	0,1647	0,2282
ship	0,1750	0,1750	0,1949	0,1114	0,1840
sugar	0,9057	0,8549	0,7668	0,5019	0,8593
coffee	0,9471	0,9270	0,8603	0,8529	0,8497
MAP	0,6550	0,6382	0,6338	0,4690	0,6357

TABLE 6.4 : Précision moyenne par catégorie et MAP pour les modèles de langage (LM).

	Réf.	lk-text	lk-slex	lk-free	IS
earn	0,8837	0,8024	0,8262	0,8385	0,8913
acq	0,2546	0,2761	0,2596	0,1267	0,6503
grain	0,9045	0,8658	0,8647	0,6059	0,9047
money-fx	0,2944	0,1819	0,0986	0,2521	0,5214
crude	0,9142	0,8832	0,8870	0,8025	0,9035
interest	0,3213	0,3318	0,3144	0,2269	0,3647
trade	0,0000	0,0000	0,0000	0,0000	0,2282
ship	0,1500	0,1464	0,1656	0,0850	0,1840
sugar	0,8364	0,7465	0,7134	0,8364	0,8593
coffee	0,4875	0,4703	0,3184	0,4825	0,8497
MAP	0,5047	0,4705	0,4448	0,4320	0,6357

Dans ces trois configurations, la précision moyenne de la catégorie *earn* est supérieure à celle des ressources `lk-slex` et `lk-text`. Cela s'explique par le taux très important de termes hors lexique dans cette catégorie. De plus, parmi les 5 termes de la requête, 3 termes ne sont pas présents dans le lexique de reconnaissance. Il s'agit là d'un cas particulier où l'absence de contrainte lexicale peut renforcer la détection de termes hors lexique.

Il est évident que la façon de générer (absence/présence de mots-clés) et d'exprimer le besoin d'information (mots-clés racinisés) peut favoriser l'approche booléenne du word spotting. Cela n'est pas sans conséquence dans les performances de IS.

6.5 Conclusion

Dans ce chapitre, les principales approches de la RI dans des collections de documents manuscrits ont été présentées. Il a été montré qu'il existe une différence conceptuelle entre les approches de RI classique appliquées couramment aux documents bruités et le word spotting. En effet, le word spotting, paradigme dominant de la recherche de documents dans le domaine manuscrit, n'effectue pas une RI mais une recherche de données, c'est-à-dire qu'il cherche seulement à déterminer quels documents contiennent un motif donné. Il a été proposé d'utiliser le word spotting comme un support permettant de restituer à un utilisateur un ensemble de documents répondant à un besoin d'information particulier.

La dernière partie de ce chapitre a présenté les résultats des expériences de RI avec le corpus manuscrit. Ces résultats se basent sur un ensemble de 10 requêtes générées à partir des données lexicales de chaque catégorie. Ces premiers résultats montrent que le bruit a un faible impact sur les performances. En comparant les performances obtenues avec les documents électroniques et les documents transcrits avec les ressources `lk-slex` et `lk-text`, une baisse d'environ 3 points seulement est observée au niveau de la MAP avec les méthodes VSM et BM25. La méthode LM montre, quant à elle, des écarts plus importants car elle est incapable de fournir le moindre document pertinent pour la catégorie *trade*.

Il a également pu être observé l'intérêt que peuvent présenter des méthodes de recherche sans reconnaissance ou basées sur une reconnaissance sans lexique lorsque les documents contiennent beaucoup de termes qui n'auraient pas pu être reconnus avec un lexique restreint.

En comparant les performances des méthodes VSM et BM25 avec les ressources `lk-slex` et `lk-text` avec celles de IS, il est difficile de déterminer laquelle de ces approches est la meilleure. Ce constat est le point de départ du chapitre suivant : plutôt que d'opposer ces méthodes, pourquoi ne pas chercher à les combiner de façon à tirer le meilleur parti des points forts de chacune ?

Afin d'améliorer les performances de la RI, la fusion de résultats en RI et son application aux documents manuscrits en-ligne ont été explorées. Notre hypothèse est qu'en combinant les différents systèmes, les chances d'obtenir des documents pertinents pour une requête donnée sont alors augmentées. La validation expérimentale de cette hypothèse fera l'objet du chapitre suivant.

Chapitre 7

Fusion de résultats en RI et son application au domaine de l'écriture en-ligne

La fusion de résultats de recherche en RI, également appelée fusion de données ou métarecherche, consiste à combiner différentes sources d'information, souvent hétérogènes. L'objectif est de fusionner les listes, renvoyées par plusieurs systèmes de RI, en une liste unique, afin d'obtenir un système combiné qui soit plus performant que les systèmes individuels.

L'application de la métarecherche au domaine en-ligne se justifie par l'existence de deux grandes familles d'approches d'indexation et de recherche pour les documents manuscrits (*cf.* §6.1). La première approche consiste à appliquer des méthodes standard en recherche d'information (RI) aux transcriptions obtenues grâce à un moteur de reconnaissance de l'écriture (Vinciarelli, 2005a). Dans ce cas, elle est nommée *RI bruitée* car les transcriptions contiennent des erreurs. La seconde approche évite un processus explicite de reconnaissance et tente d'identifier les mots-clés soumis par un utilisateur dans une distribution donnée par un automate de Markov à états cachés (Kwok et collab., 2000 ; Russell et collab., 2002), ou directement dans le tracé manuscrit (Lopresti et Tomkins, 1994 ; Jain et Namboodiri, 2003 ; Schimke et Vielhauer, 2006 ; Jawahar et collab., 2009). Dans ce cas elle est nommée *word spotting* (voir chapitre précédent).

Chacune de ces approches possède ses avantages propres. Dans le cas du word spotting, sa robustesse pour détecter les mots-clés est souvent mise en avant. En revanche, il est souvent reproché à ce type de méthodes d'avoir une approche binaire (présence/absence de mots-clés) de la RI (Vinciarelli, 2006). D'un autre côté, les méthodes standard de RI possèdent des schémas de pondération favorisant des termes en fonction de leur importance. De plus, elles peuvent considérer les variations morphologiques d'un même mot comme une seule entité. Dans le cas du word spotting des tentatives pour imiter ce comportement ne sont apparues que très récemment (Jawahar et collab., 2009). Cependant, les performances des méthodes standards risquent d'être pénalisées par une quantité importante d'erreurs de reconnaissance.

En pratique, il est difficile de déterminer *a priori* laquelle de ces approches est la meilleure. Il a été proposé à la fin du chapitre précédent, d'appliquer des méthodes de fusion afin d'améliorer les performances de la recherche. La fusion peut tirer parti des fortes différences systémiques entre les deux familles d'approches et de la diversité des résultats qu'elles peuvent engendrer. Tant que les ensembles de documents pertinents retournés par deux systèmes différents sont suffisamment disjoints, les résultats peuvent être améliorés en termes de rappel. D'un autre côté, même les documents pertinents communs peuvent être source d'amélioration, au niveau de la précision, s'ils se retrouvent classés dans le haut de la liste après la fusion. Ce chapitre présente les résultats obtenus grâce à l'application des méthodes de fusion de données pour la recherche de documents manuscrits en-ligne (Peña Saldarriaga, Viard-Gaudin et Morin, 2010b ; Peña Saldarriaga, Morin et Viard-Gaudin, 2010a ; Peña Saldarriaga, Morin et Viard-Gaudin, 2010).

La section 7.1 introduit les notions de base nécessaires à la compréhension de l'étude présentée dans ce chapitre. Ensuite, une section sera consacrée à la définition des différentes méthodes de fusion utilisées dans le cadre de nos expériences (§7.2). La section 7.3 s'attaque à l'évaluation empirique de la fusion de données appliquée à la RI de documents manuscrits en-ligne. Dans la dernière section, sera dressé un bilan de l'étude expérimentale présentée dans ce chapitre.

7.1 Concepts

Comme il a été vu dans le chapitre précédent, les algorithmes de recherche d'information calculent un ensemble de scores pour les documents d'un corpus, noté τ. Ces scores induisent une relation d'ordre entre les documents de τ, de ce fait, il s'agit d'un ensemble ordonnée $\tau = \{d_a \prec d_b \prec d_c \prec \ldots \prec d_{|\tau|}\}$. Par rapport au domaine des documents $\mathcal{D}$, $\tau \subseteq \mathcal{D}$ est un sous-ensemble ordonné des documents du domaine.

Définition 7.1. *Pour tout document $i \in \tau$, $\tau(i)$* est le rang de *i, le rang de i est petit lorsque sa position dans le classement est haute.*

Le rang d'un document est déterminé par son indicateur de pertinence, ou *retrieval status value* (rsv).

Définition 7.2. *Pour tout document $i \in \tau$, $\mathrm{s}^{\tau}(i)$* est le score du document *i. Sans restreindre la généralité, nous considérons que le rang de i est est une fonction décroissante de son score.*

Lorsque $\tau = \mathcal{D}$, τ est une *liste complète*, c'est-à-dire qu'elle induit une relation d'ordre entre tous les documents de la collection. Cependant, la plupart du temps, les résultats d'une recherche sont des *listes partielles*. En effet, les SRI ne peuvent classer qu'un sous-ensemble strict de documents $\tau \subset \mathcal{D}$, par exemple, ceux qui contiennent les mots-clés de la requête. En plus de retourner des listes partielles, différents SRI peuvent retourner des scores qui ne sont pas comparables entre eux. Une étape de normalisation de scores, préalable à la fusion, devient alors nécessaire.

Définition 7.3. *Pour tout document $i \in \tau$, $\omega^\tau(i)$* est le score normalisé du document *i, la normalisation employée ici a été proposée par Montague et Aslam (2001). Le score normalisé de i est donné par :*

$$\omega^\tau(i) = \frac{\mathrm{s}^\tau(i)}{\sum_{j \in \tau} \mathrm{s}^\tau(j)} \tag{7.1}$$

Dans certaines situations, notamment dans le cas de la fusion de résultats de recherches issus du web, les scores ne sont pas toujours disponibles. Dans ce cas, seuls sont disponibles les rangs des documents. Pour chaque document, un score en fonction de son rang peut alors être calculé.

Définition 7.4. *Pour tout document $i \in \tau$, $\mathrm{r}^\tau(i)$* est le score de rang normalisé du document *i. Le score normalisé basé sur le rang de i est donné par :*

$$\mathrm{r}^\tau(i) = 1 - \frac{\tau(i) - 1}{|\tau|} \tag{7.2}$$

Enfin, le dernier concept clé pour la fusion de données est la notion d'*accord* (Lee, 1997).

Définition 7.5. *Soit $\mathcal{R} = \{\tau_1, \tau_2, \ldots \tau_{|\mathcal{R}|}\}$ un ensemble de résultats retournés par différents systèmes de RI. Soit $\mathcal{U} = \{\tau_1 \cup \tau_2 \cup \ldots \cup \tau_{|\mathcal{R}|}\}$ l'ensemble issu de l'union de tous les documents des différentes listes partielles appartenant à $\mathcal{R}$. Pour tout document $i \in \mathcal{U}$, $\mathrm{h}(i, \mathcal{R})$* est le nombre d'accords dans $\mathcal{R}$ pour le document *i. Formellement, le nombre d'accords est donné par :*

$$\mathrm{h}(i, \mathcal{R}) = |\{\tau \in \mathcal{R} : i \in \tau\}| \tag{7.3}$$

Concrètement, il s'agit du nombre de listes partielles qui contiennent le document i, l'hypothèse sous-jacente à cette notion étant que plus un document est retrouvé par des systèmes différents, plus il a des chances d'être pertinent (Lee, 1997).

7.2 Méthodes de fusion de résultats

Cette section décrit les méthodes de fusion de résultats utilisées dans nos expériences. La fusion de données est un domaine de recherche très actif en recherche d'information. Il existe une quantité innombrable de méthodes et de nouvelles sont proposées régulièrement (Shaw et Fox, 1994 ; Lee, 1997 ; Savoy, Le Calvé et Vrajitoru, 1997 ; Vogt et Cottrell, 1999 ; Aslam et Montague, 2001 ; Manmatha, Rath et Feng, 2001 ; Dwork, Kumar, Naor et Sivakumar, 2001 ; Montague et Aslam, 2002 ; Renda et Straccia, 2003 ; Beitzel, Jensen, Chowdury, Grossman, Frieder et Goharian, 2004 ; Wu et McClean, 2005 ; Farah et Vanderpooten, 2007).

Très vite, Belkin, Cool, Croft et Callan (1993) ont montré que différentes représentations d'une même requête pouvaient donner lieu à des résultats de recherche très

différents. Les travaux de Shaw et Fox (1994) ont tenté de tirer parti de la diversité des résultats obtenus par différentes stratégies de recherche. Ils ont introduit un ensemble de stratégies de fusion basées sur les opérateurs *min* et *max*, ainsi que sur une combinaison linéaire des rsv. Vogt et Cottrell (1999) ont proposé une méthode de combinaison linéaire où chaque système considéré pour la fusion se voit affecter un poids qui reflète la confiance qui lui est portée. Cette méthode a l'inconvénient de nécessiter des données d'entraînement pour estimer les poids optimaux. Wu et McClean (2005) ont proposé d'utiliser des mesures de corrélation entre les différents systèmes afin d'éviter l'étape d'entraînement. Jusqu'à aujourd'hui, la manière d'estimer le poids accordé à chaque système reste une question ouverte (Wu, Bi, Zeng et Han, 2009).

Des modèles statistiques ont également été proposés. Manmatha et collab. (2001) ont proposé de modéliser les résultats de recherche par une fonction de densité, issue d'une combinaison linéaire d'une gaussienne et d'une loi exponentielle. Les scores combinés sont ensuite calculés par la moyenne des probabilités estimées. Un modèle de fusion basé sur la loi de Bayes a été proposé par Aslam et Montague (2001). La régression linéaire a également été employée pour ce type de tâches (Savoy et collab., 1997). Il faut noter que ces dernières méthodes nécessitent des données d'entraînement pour l'estimation de leurs paramètres respectifs.

D'autres auteurs ont considéré le problème de la fusion de résultats comme une procédure de scrutin. À partir des résultats donnés par les différents systèmes, il s'agit de trouver un consensus. Des méthodes classiques en théorie des choix collectifs comme la méthode de Borda (1781) ou Condorcet (1785) ont déjà été appliquées (Aslam et Montague, 2001 ; Montague et Aslam, 2002). Des modes de scrutin à base de chaînes de Markov (Dwork et collab., 2001 ; Renda et Straccia, 2003) ou des méthodes multicritères d'aide à la décision ont également été explorés (Farah et Vanderpooten, 2007). Un état de l'art plus détaillé des méthodes existantes peut être trouvé dans les travaux de Beitzel et collab. (2004) ou Farah et Vanderpooten (2007).

Comme signalé par Aslam et Montague (2001), ces méthodes peuvent être classées selon qu'elles utilisent les rangs ou les scores pour la fusion. L'autre critère permettant de classer ces méthodes est le besoin de données d'entraînement pour l'estimation de paramètres (*cf.* figure 7.1).

Dans ce travail, seules des méthodes simples, ne nécessitant pas de données d'entraînement, seront abordées. Seront laissées de côté les méthodes à droite du spectre de répartition des méthodes de fusion (*cf.* figure 7.1). Ces méthodes, tout en étant plus complexes montrent, la plupart du temps, des résultats mitigés selon les corpus utilisés. A contrario, des méthodes très simples, comme l'opérateur CombMNZ, sont devenues des standards dans le domaine car elles permettent d'obtenir de très bons résultats, tout en étant robustes vis-à-vis des différents corpus, même si dans des contextes particuliers leur efficacité a été mise à défaut (Beitzel et collab., 2004).

Les sous-sections suivantes décrivent les différentes méthodes de fusion de résultats considérées dans les expériences menées dans le cadre de cette thèse. Le tableau 7.1 récapitule les différents éléments de base pour la définition de ces méthodes. La notation est celle adoptée également par Renda et Straccia (2003).

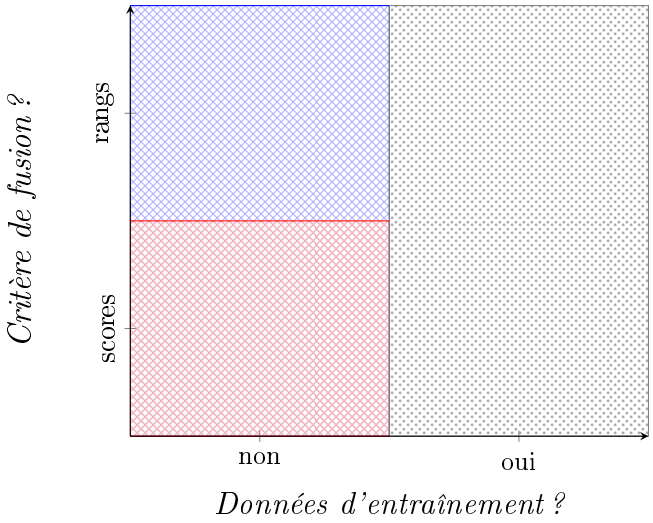


FIGURE 7.1 : Caractérisation des méthodes de fusion de résultats en RI.

TABLE 7.1 : Éléments notationnels pour la définition des méthodes de fusion.

Symbole	Définition
i	Un document
τ	Classement de documents
$\tau\,(i)$	Rang du document i
$\omega^{\tau}\,(i)$	Score normalisé du document i
$\mathrm{r}^{\tau}\,(i)$	Score basé sur le rang de i
$\mathcal{R}$	Ensemble de résultats à fusionner
$\mathcal{U}$	Union de tous les $i \in \tau$, $\tau \in \mathcal{R}$
$\mathrm{h}(i, \mathcal{R})$	Nombre d'accords pour le document i
$\mathrm{s}^{\hat{\tau}}\,(i)$	Score combiné du document i

7.2.1 Méthodes basées sur les scores

Les trois premières méthodes considérées se basent sur une combinaison des scores normalisés des documents.

CombSUM

L'opérateur CombSUM introduit par Shaw et Fox (1994) consiste à combiner de façon linéaire les scores. Les différents ensembles considérés dans la fusion reçoivent le même poids :

$$\mathrm{s}^{\hat{\tau}}(i) = \sum_{\tau \in \mathcal{R}} \omega^{\tau}(i) \tag{7.4}$$

CombMNZ

L'opérateur CombMNZ est également basé sur une combinaison linéaire des scores (Shaw et Fox, 1994), mais cette fois, les scores de documents ayant été retrouvés par plus d'un des systèmes considérés pour la fusion sont renforcés en multipliant par le nombre d'accords :

$$\mathrm{s}^{\hat{\tau}}(i) = \mathrm{h}(i, \mathcal{R}) \times \sum_{\tau \in \mathcal{R}} \omega^{\tau}(i) \tag{7.5}$$

Ce renforcement se base sur l'hypothèse selon laquelle plus un document est retrouvé par des systèmes différents, plus il a des chances d'être pertinent.

CombHMEAN

Même si le raisonnement qui a conduit à l'opérateur CombMNZ se révèle être bénéfique expérimentalement, il peut également conduire à des effets pervers lorsque les ensembles considérés dans la fusion partagent un nombre important de documents non-pertinents. En effet, ceux-ci voient leurs scores renforcés et leur classement amélioré.

Dans cette étude, il est également proposé de combiner les scores en prenant la moyenne harmonique. Comme la moyenne harmonique tend vers la plus petite des valeurs moyennées, elle permet de pénaliser les désaccords entre les différents systèmes à combiner. L'accord, cette fois-ci n'est pas donné par $\mathrm{h}(i, \mathcal{R})$ mais par les scores eux-mêmes. Le score combiné par l'opérateur CombHMEAN est donné par :

$$\mathrm{s}^{\hat{\tau}}(i) = \frac{|\mathcal{R}|}{\sum_{\tau \in \mathcal{R}} \frac{1}{\omega^{\tau}(i) + \mu}} \tag{7.6}$$

Comme la moyenne harmonique n'est définie que pour $\omega^{\tau}(i) > 0$, c'est-à-dire que l'opérateur CombHMEAN ne peut pas gérer les listes partielles, un paramètre d'ajustement, μ, est introduit. Dans les expériences décrites dans la section 7.3, μ a été fixé à 10^{-10}.

7.2.2 Méthodes basées sur les rangs

Les trois dernières méthodes considérées se basent sur les rangs des documents. Les deux premières méthodes correspondent aux opérateurs de combinaison linéaire agissant cette fois sur les rangs. La troisième, quant à elle, est une méthode classique en théorie des choix collectifs : la méthode de Borda (1781).

RankCombSUM

L'opérateur RankCombSUM est une combinaison linéaire des scores de rang normalisés pour chacun des documents :

$$\mathrm{s}^{\hat{\tau}}(i) = \sum_{\tau \in \mathcal{R}} \mathrm{r}^{\tau}(i) \tag{7.7}$$

RankCombMNZ

L'opérateur RankCombMNZ est le résultat de RankCombSUM multiplié par le nombre d'accords pour chacun des documents :

$$\mathrm{s}^{\hat{\tau}}(i) = \mathrm{h}(i, \mathcal{R}) \times \sum_{\tau \in \mathcal{R}} \mathrm{r}^{\tau}(i) \tag{7.8}$$

Méthode de Borda

La méthode de Borda est une méthode de scrutin pondérée. En plus d'être utilisée pour la fusion basée sur les rangs, elle a également été utilisée pour la combinaison de classifieurs dans le contexte de la reconnaissance de l'écriture (van Erp et Schomaker, 2000).

Chacun des ensembles considérés pour la fusion peut être vu comme un électeur, et leurs classements respectifs comme des bulletins indiquant leurs préférences vis-à-vis des documents, c'est-à-dire des candidats. Chaque document candidat i se voit attribuer un certain nombre de points en fonction de son rang, le premier candidat obtient $|\tau|$ points, le second obtient $|\tau| - 1$ points, et ainsi de suite. Lors du travail avec des listes partielles, les candidats qui n'ont pas été classés par un électeur se voient attribuer le nombre de points restants divisés équitablement.

La figure 7.2 illustre le calcul des scores de Borda, $\mathrm{b}^{\tau}(i)$, pour trois classements différents, ainsi que le classement final obtenu après leur fusion, $\hat{\tau}$.

Définition 7.6. *Soit* $\mathcal{R} = \{\tau_1, \tau_2, \ldots \tau_{|\mathcal{R}|}\}$ *un ensemble de résultats retournés par différents systèmes de RI. Soit* $\mathcal{U} = \{\tau_1 \cup \tau_2 \cup \ldots \cup \tau_{|\mathcal{R}|}\}$ *l'ensemble issu de l'union de tous les documents des différentes listes partielles appartenant à* $\mathcal{R}$*. Pour tout document* $i \in \mathcal{U}$, $\mathrm{b}^{\tau}(i)$ est le score de Borda pour le document i*. Formellement, il est calculé grâce à la formule suivante :*

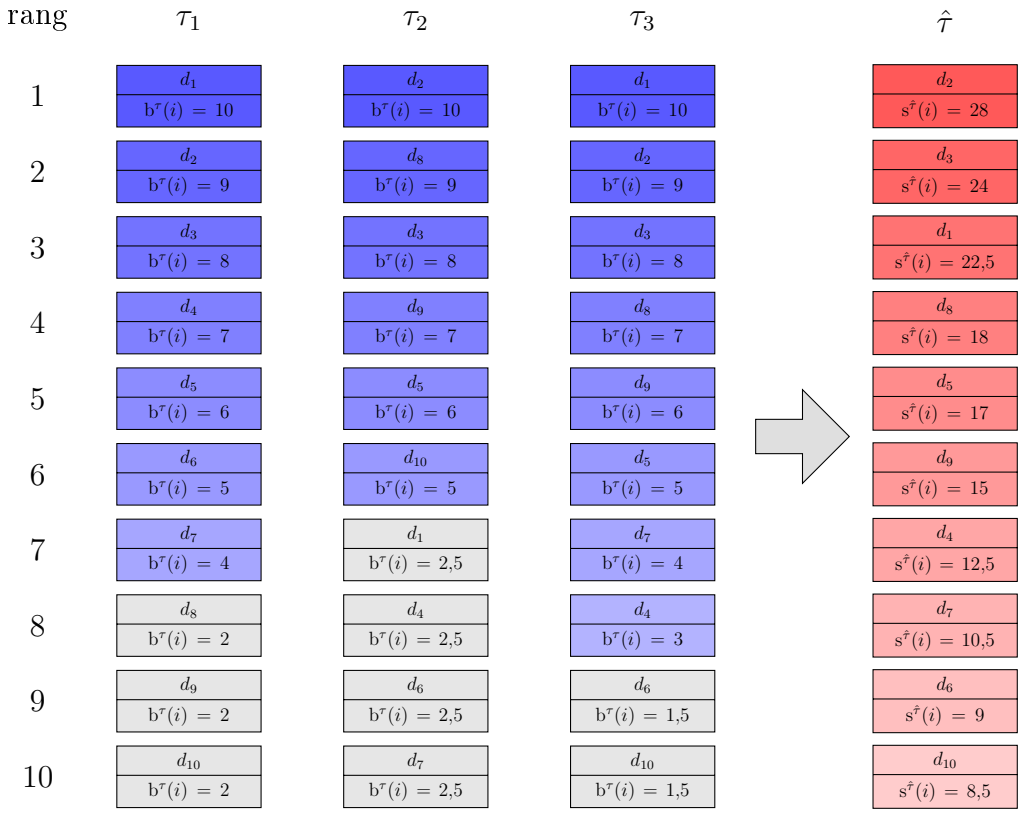


FIGURE 7.2 : Calcul du score de Borda pour trois classements différents. Les documents en gris sont ceux qui n'ont pas été classés par les différents systèmes. Le classement après fusion est donné en rouge.

$$\mathrm{b}^\tau(i) = \begin{cases} |\mathcal{U}| - \tau(i) + 1, & si\ i \in \tau \\ \dfrac{(|\mathcal{U}| - |\tau|) \times (|\mathcal{U}| - |\tau| + 1)}{2 \times (|\mathcal{U}| - |\tau|)}, & sinon \end{cases} \tag{7.9}$$

La fusion proprement dite s'effectue ensuite en additionnant les scores :

$$\mathrm{s}^{\hat{\tau}}(i) = \sum_{\tau \in \mathcal{R}} \mathrm{b}^\tau(i) \tag{7.10}$$

7.2.3 Prédire la réussite de la fusion

Malgré l'existence d'une littérature riche en méthodes de fusion et succès expérimentaux, peu de recherches se sont intéressées, sinon à la prédiction, du moins à l'étude des conditions nécessaires à la réussite ou l'échec de la fusion.

Les expériences menées par Lee (1997) se basaient sur l'observation selon laquelle des systèmes de SRI ont tendance à restituer les mêmes documents pertinents, mais des documents non pertinents différents. Ainsi, la fusion pouvait être fructueuse tant

que les systèmes en jeu partageaient un nombre important de documents pertinents et un nombre faible de documents non pertinents. Deux mesures ont été proposées pour calculer le nombre de documents partagés.

Définition 7.7. *Soit $\mathcal{P}$ l'ensemble de documents pertinents correspondant à une requête et $\mathcal{I}$ son ensemble de documents non pertinents, tels que $\mathcal{P} \cup \mathcal{I} = \mathcal{D}$.* Le taux de chevauchement de documents pertinents, noté $r_{\text{pertin.}}$, est défini par :

$$r_{\text{pertin.}} = \frac{|\mathcal{R}| \times \left|\mathcal{P} \cap \tau_1 \cap \tau_2 \cap \ldots \cap \tau_{|\mathcal{R}|}\right|}{\sum_{\tau \in \mathcal{R}} |\mathcal{P} \cap \tau|} \tag{7.11}$$

Définition 7.8. Le taux de chevauchement de documents non pertinents, noté $r_{\text{npertin.}}$, est calculé de la même façon :

$$r_{\text{npertin.}} = \frac{|\mathcal{R}| \times \left|\mathcal{I} \cap \tau_1 \cap \tau_2 \cap \ldots \cap \tau_{|\mathcal{R}|}\right|}{\sum_{\tau \in \mathcal{R}} |\mathcal{I} \cap \tau|} \tag{7.12}$$

Les conclusions tirées de l'étude de Lee (1997) ont été contredites plus tard par les travaux de Beitzel et collab. (2004). Ils ont montré que la fusion de différentes stratégies de recherche partageant les mêmes algorithmes de segmentation, racinisation, etc., n'apporte pas d'amélioration de la MAP. Ils ont mis à défaut les mesures proposées par Lee (1997) et proposé de corréler les améliorations attendues de la fusion par le coefficient de corrélation de Spearman (Myers et Well, 2003, p.508)[1]. De plus, Beitzel et collab. (2004) ont proposé d'utiliser le coefficient moyen de déplacement de rangs pour montrer que la fusion n'était pas à même d'améliorer les résultats voire qu'elle pouvait les dégrader. Il s'agit des différences absolues entre les rangs avant et après fusion.

Définition 7.9. Le coefficient moyen de déplacement de rang pour les documents pertinents, noté $rdc_{\text{pertin.}}$, est défini par :

$$rdc_{\text{pertin.}} = \frac{1}{|\mathcal{P}|} \sum_{i \in (\hat{\tau} \cap \mathcal{P})} \hat{\tau}(i) - \tau(i) \tag{7.13}$$

Définition 7.10. Le coefficient moyen de déplacement de rang pour les documents non pertinents, noté $rdc_{\text{npertin.}}$, est défini par :

$$rdc_{\text{npertin.}} = \frac{1}{|\mathcal{I}|} \sum_{i \in (\hat{\tau} \cap \mathcal{I})} \hat{\tau}(i) - \tau(i) \tag{7.14}$$

L'une des conclusions du travail de Beitzel et collab. (2004) est que la fusion ne peut améliorer les performances que lorsque les documents pertinents non communs aux différents systèmes sont reclassés dans le haut du classement. Or, le coefficient de

1. Voir également §5.4.1.

déplacement de documents pertinents ne peut être calculé que si $\hat{\tau} \cap \mathcal{P} = \tau \cap \mathcal{P}$, alors que c'est la partie disjointe des ensembles qui peut améliorer les performances. Sans restreindre la généralité, il est raisonnable de penser que la partie disjointe de $\hat{\tau}$ et τ est classée après le dernier élément de τ. En appliquant le même raisonnement aux documents non pertinents, ces deux mesures ne peuvent être calculées que si $\tau = \mathcal{U}$.

Par conséquent, plus les ensembles de documents non pertinents avant et après fusion sont disjoints, plus les documents non pertinents vont gagner des places en moyenne car *ils ne figuraient pas dans le classement auparavant* tout en étant plus nombreux. De plus, selon l'hypothèse de Lee (1997), c'est-à-dire que les ensembles de documents pertinents sont plutôt joints alors que ceux des documents non pertinents tendent à être disjoints, cela veut dire que ces mesures sont fortement biaisées vers la confirmation de l'hypothèse de Beitzel et collab. (2004). En effet, les documents non pertinents vont gagner plus de places en moyenne que les documents pertinents.

Une partie de notre contribution expérimentale consistera à vérifier que différentes mesures de prédiction des améliorations se généralisent au domaine manuscrit en-ligne. Eu égard aux objections théoriques qui viennent d'être effectuées concernant les coefficients de déplacement, et dans l'absence de détails concernant leur calcul avec des listes partielles, les résultats seront analysés à la lumière des taux de chevauchement ainsi que du ρ de Spearman.

7.3 Résultats

Cette section présente et discute les résultats obtenus après fusion des résultats de IS avec les méthodes de RI bruitée.

7.3.1 Méthodes de référence

Les performances obtenues par les différentes méthodes de référence, sur l'ensemble des requêtes générées pour notre corpus, sont présentées ici. L'ensemble des requêtes a été présenté dans le tableau 3.3 et rappelé dans le chapitre précédent (*cf.* tableau 6.1). La figure 7.3 montre les résultats individuels, en termes de précision moyenne, pour chacune des méthodes de référence selon la version transcrite du corpus utilisée.

La MAP obtenue avec le word spotting est, pour l'ensemble de nos requêtes, très proche de celles qui peuvent être obtenues en appliquant les approches VSM et BM25 sur les transcriptions obtenues avec `lk-slex` et `lk-text`. Indépendamment de la qualité du corpus utilisé, l'approche par modèles de langage est largement dépassée par toutes les autres méthodes. Le TER très important des documents de `lk-free` se traduit par une baisse de 15 points en moyenne au niveau de la MAP. La MAP de IS reste constante tout au long de la figure 7.3 car cette méthode utilise le tracé manuscrit.

Des éléments supplémentaires des performances obtenues avec les méthodes de référence sont donnés par les mesures de haute précision (*cf.* figure 7.4). L'impact des erreurs de reconnaissance est moins important que sur la précision moyenne pour les méthodes de RI bruitée. Les écarts observés pour les documents de `lk-free`, par rap-

port à la vérité terrain, sont de moindre importance. Encore une fois, LM arrive toujours en dernière position indépendamment de la version du corpus utilisée.

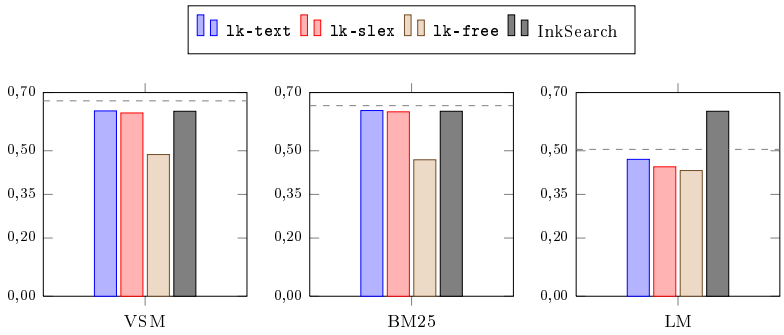
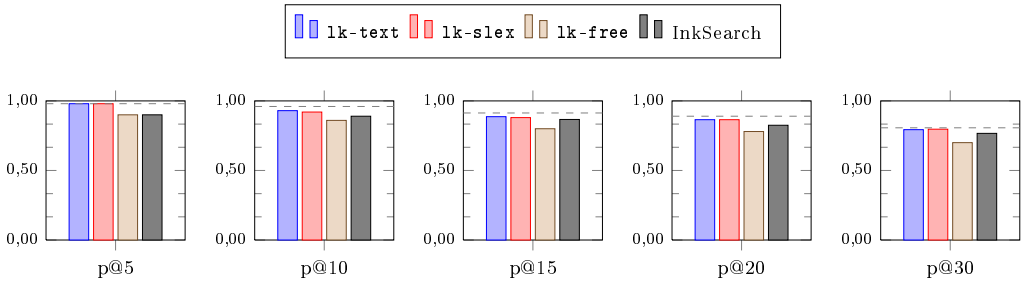
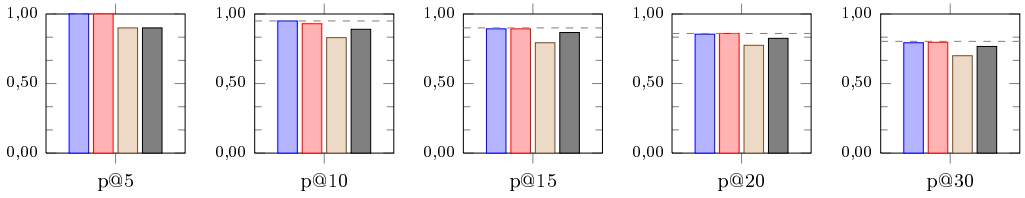


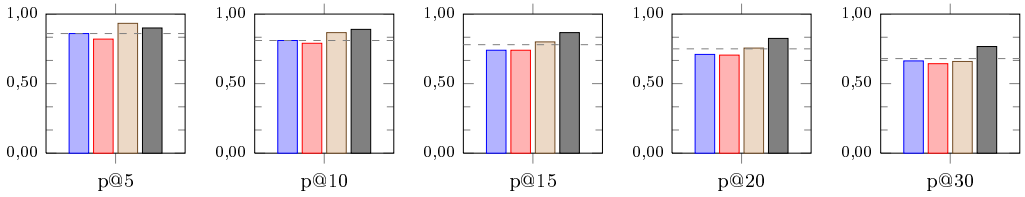
FIGURE 7.3 : Précision moyenne pour les méthodes de référence en fonction de la ressource utilisée pour la reconnaissance des documents. La ligne pointillée indique les performances obtenues avec les documents de la vérité terrain.



(a) VSM



(b) BM25



(c) LM

FIGURE 7.4 : Précision à n documents pour les méthodes de référence.

7.3.2 Fusion de résultats

Les mesures choisies pour évaluer l'intérêt de la fusion dans le domaine manuscrit en-ligne sont la précision moyenne, les courbes de précision-rappel ainsi que les mesures dites de haute précision.

La précision moyenne après fusion est présentée dans le tableau 7.2. La fusion s'effectue toujours entre IS et chacune des autres méthodes de référence. Les tableaux 7.3(a) et 7.3(b) montrent les effets positifs de la fusion sur la précision moyenne, à l'exception notable des résultats combinés avec LM. De plus, dans certains cas les performances après fusion sont supérieures à celles qui seraient obtenues en n'utilisant que les documents de la vérité terrain. Dans toutes les autres configurations où des améliorations sont observées, les MAP obtenues sont très proches de la vérité terrain. L'opérateur de fusion que nous avons proposé, CombHMEAN, obtient des résultats proches des opérateurs classiques CombSUM et CombMNZ. Dans le cas des documents de `lk-slex`, notre opérateur surpasse les performances de tous les autres.

Les résultats de la fusion avec les documents de `lk-free` sont plus mitigés. Ce qui peut sembler normal, eu égard à la qualité des transcriptions. Seulement quelques configurations obtiennent une MAP supérieure à celle des méthodes de référence prises individuellement. En comparant les résultats pour les trois versions du corpus (`lk-text` et `lk-slex` *versus* `lk-free`), il est observé que la différence entre les résultats fusionnés est de l'ordre de 3 % à 5 % alors que celle entre les résultats individuels est de 10 %

TABLE 7.2 : Précision moyenne après fusion des résultats. Les chiffres en gras indiquent une amélioration des performances par rapport à IS, tandis que ceux en italique indiquent une dégradation. Les résultats marqués d'une † indiquent que la MAP est supérieure à celle obtenue avec les documents de la vérité terrain.

(a) `lk-text`

	VSM	BM25	LM
CombSUM	**0,6694**	**0,6728**†	*0,6242*
CombMNZ	**0,6707**	**0,6727**†	*0,6242*
CombHMEAN	**0,6667**	**0,6680**†	*0,6202*
Borda	**0,6640**	**0,6671**†	*0,6325*
RankCombSUM	**0,6621**	**0,6680**†	*0,6249*
RankCombMNZ	**0,6656**	**0,6710**†	*0,6245*

(b) `lk-slex`

	VSM	BM25	LM
CombSUM	**0,6718**†	**0,6734**†	*0,6238*
CombMNZ	**0,6667**	**0,6686**†	*0,6120*
CombHMEAN	**0,6734**†	**0,6742**†	*0,6221*
Borda	**0,6674**	**0,6677**†	*0,6284*
RankCombSUM	**0,6662**	**0,6683**†	*0,6219*
RankCombMNZ	**0,6696**	**0,6714**†	*0,6221*

(c) `lk-free`

	VSM	BM25	LM
CombSUM	**0,6456**	**0,6371**	*0,6199*
CombMNZ	*0,6280*	*0,6198*	*0,5944*
CombHMEAN	**0,6440**	**0,6366**	*0,6277*
Borda	**0,6467**	**0,6374**	*0,6213*
RankCombSUM	*0,6354*	*0,6334*	*0,6186*
RankCombMNZ	**0,6371**	*0,6335*	*0,6181*

en moyenne et peut aller jusqu'à 17 % (*cf.* tableau 6.3). Cela veut dire que dans le cas de la version `lk-free`, même si IS domine le processus de fusion, la recherche dans les documents fortement bruités peut apporter de la pertinence aux résultats du word spotting.

En ce qui concerne les mesures de haute précision, leur amélioration globale suit la même tendance que la MAP (*cf.* figures 7.5 à 7.7). Les stratégies de fusion, dans lesquelles la recherche à base de modèles de langage intervient, n'améliorent pas la p@n par rapport à IS seul. En revanche, les figures 7.5 et 7.6 montrent que la fusion des

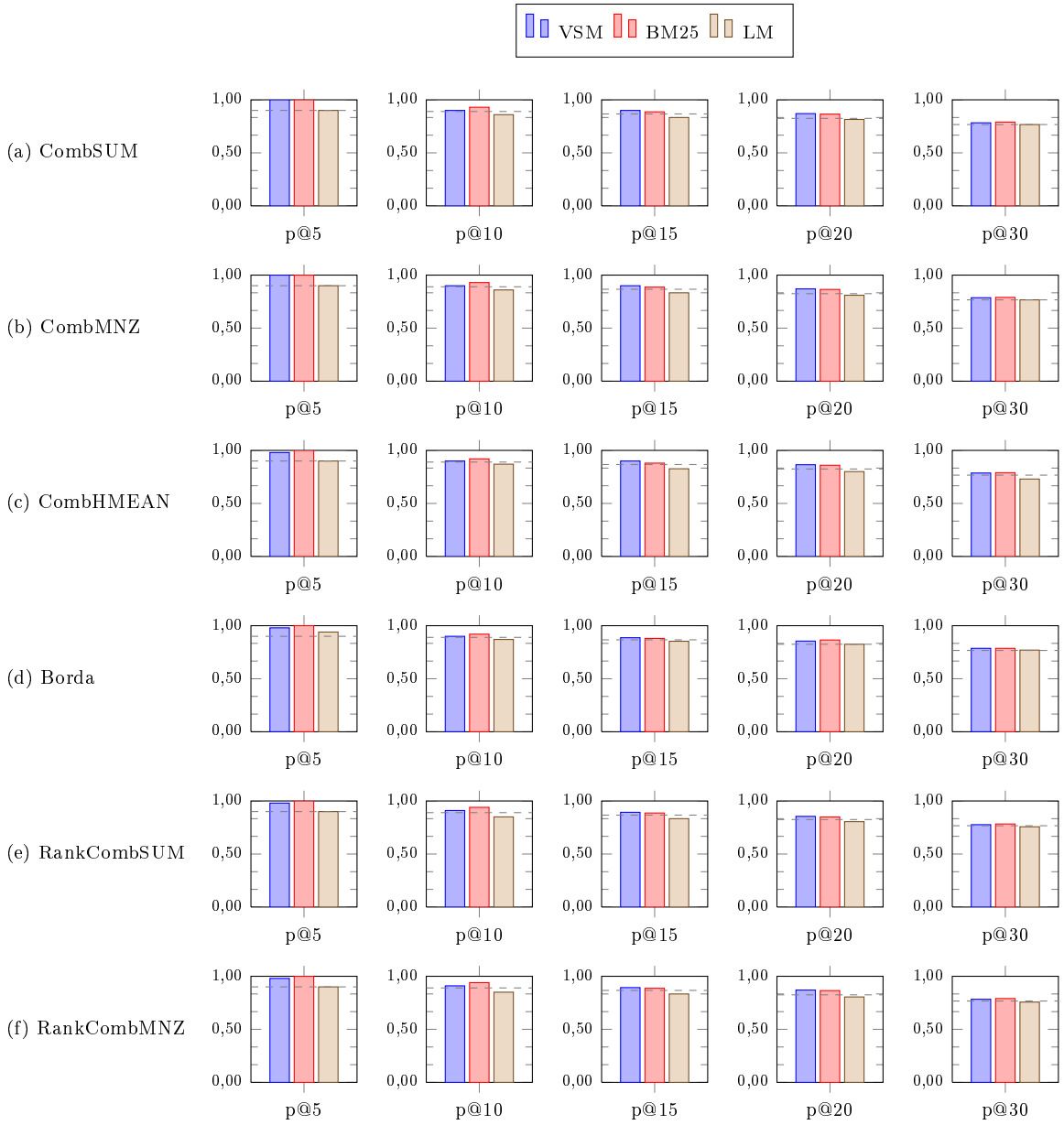


FIGURE 7.5 : Précision à n documents après fusion avec les documents de `lk-text`. La ligne pointillée indique les performances pour IS.

stratégies de RI, VSM et BM25, avec MyScript® InkSearch® permettent d'améliorer la p@n, à tous les niveaux considérés, avec les documents de `lk-text` et `lk-slex`. Lorsque la version `lk-free` est utilisée avec les mêmes systèmes, la fusion améliore seulement la précision à 5 documents. Au delà de 5 documents les précisions relevées sont toujours inférieures ou égales à celles de IS.

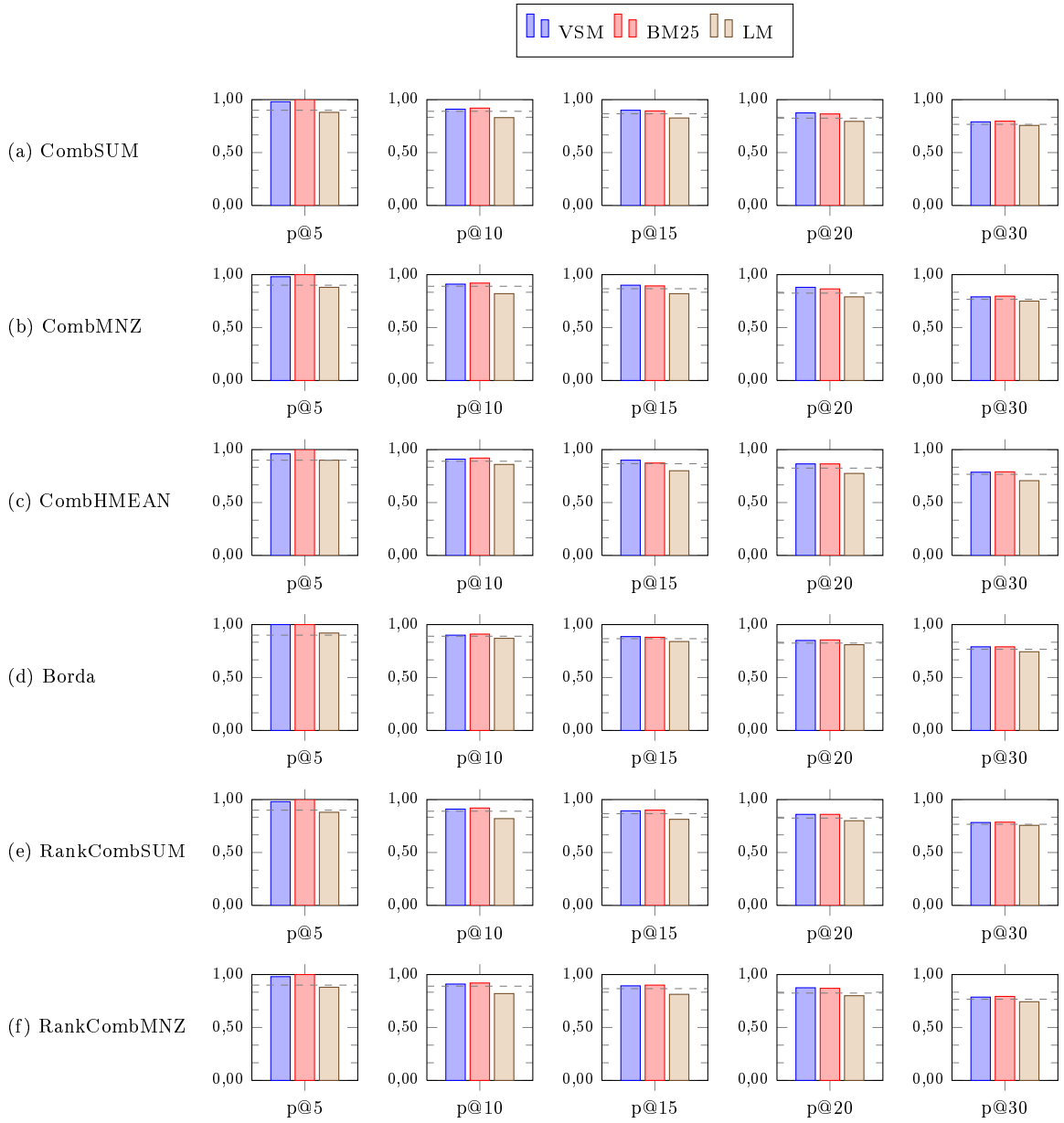


FIGURE 7.6 : Précision à n documents après fusion avec les documents de `lk-slex`. La ligne pointillée indique les performances pour IS.

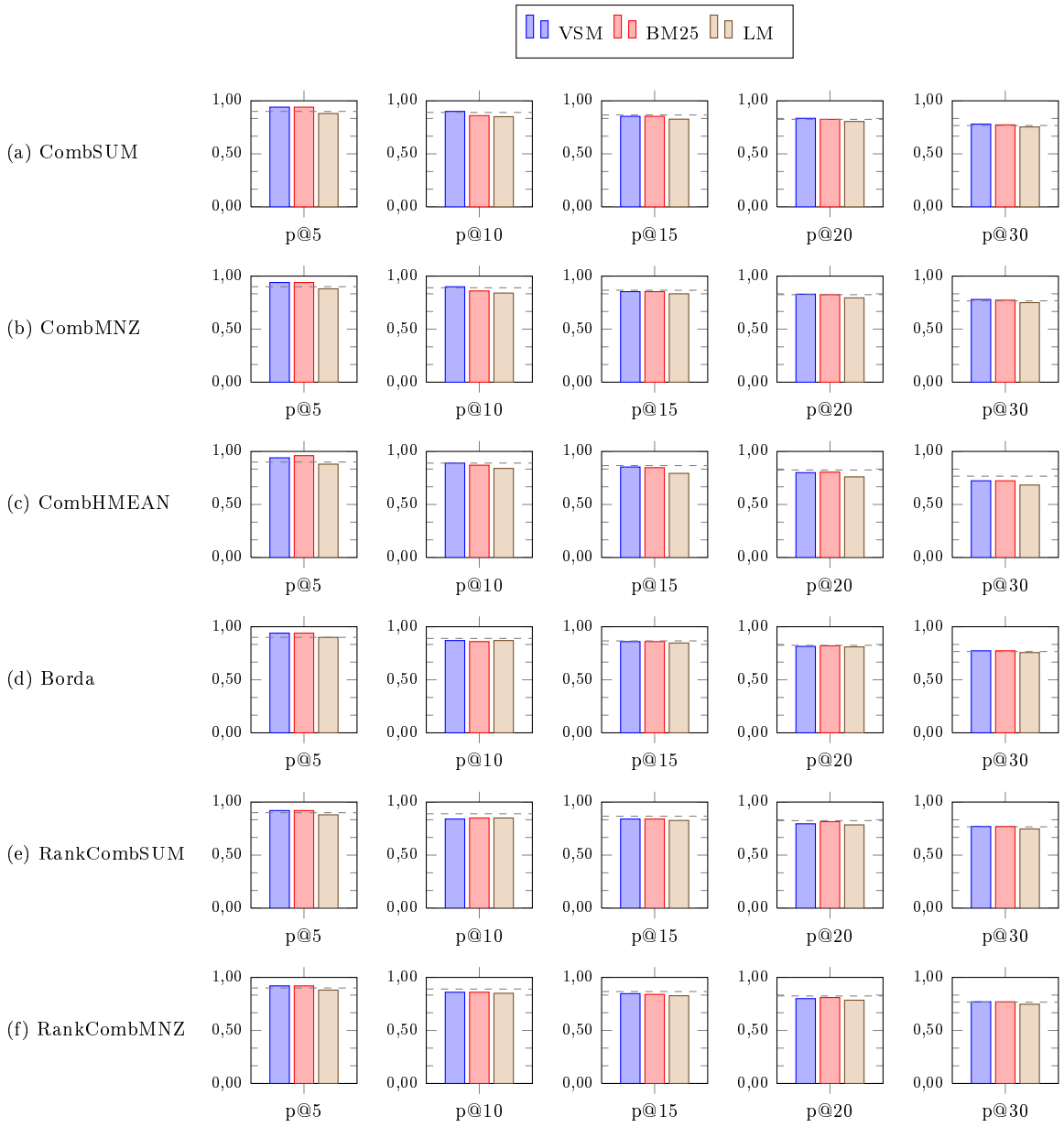


FIGURE 7.7 : Précision à n documents après fusion avec les documents de `lk-free`. La ligne pointillée indique les performances pour IS.

7.3.3 Analyse de la réussite ou de l'échec de la fusion

Les résultats qui viennent d'être présentés sont très variables. Afin de mieux comprendre cette variabilité, ll reste à analyser les conditions nécessaires à une fusion fructueuse. Pour les taux de chevauchement suggérés par Lee (1997), la première hypothèse se vérifie pour toutes les configurations. En effet, $r_{\text{pertin.}} \gg r_{\text{npertin.}}$. La différence entre les deux taux (Δ_r) ne semble pas particulièrement prédictive. En mettant en perspective les tableaux 7.4(a) et 7.4(c), les différences sont sensiblement les mêmes mais les résultats de la fusion ne sont systématiquement positifs que pour les documents de `lk-text`.

TABLE 7.3 : Taux de chevauchement et ρ de Spearman pour les différentes méthodes de référence par rapport à IS.

(a) `lk-text`

	VSM	BM25	LM
$r_{\text{pertin.}}$	90,45 %	91,26 %	72,47 %
$r_{\text{npertin.}}$	26,71 %	26,03 %	16,89 %
Δr	63,74 %	65,23 %	55,58 %
ρ	0,6606	0,6792	0,6375

(b) `lk-slex`

	VSM	BM25	LM
$r_{\text{pertin.}}$	90,91 %	92,06 %	69,98 %
$r_{\text{npertin.}}$	22,95 %	23,12 %	14,16 %
Δr	67,96 %	68,93 %	55,82 %
ρ	0,6089	0,6332	0,5735

(c) `lk-free`

	VSM	BM25	LM
$r_{\text{pertin.}}$	76,48 %	76,71 %	63,77 %
$r_{\text{npertin.}}$	12,54 %	12,13 %	9,60 %
Δr	63,93 %	64,58 %	54,17 %
ρ	0,5701	0,5757	0,5368

Le coefficient de corrélation par rangs de Spearman (ρ) semble encore moins prédictif. En effet, les résultats de la fusion fluctuent en fonction de sa valeur. Selon Beitzel et collab. (2004), un ρ faible est synonyme d'effets positifs. Or, dans toutes les configurations les corrélations sont positives, plutôt fortement, mais les effets de la fusion ne sont pas majoritairement positifs ou négatifs. D'après ces résultats ce sont plutôt les corrélations importantes qui permettraient d'améliorer les performances, exception faite de la colonne correspondant à LM dans le tableau 7.4(a). En effet, ce tableau montre que la valeur du ρ pour LM/`lk-text` est supérieure à celle observée dans les colonnes VSM/BM25 du tableau 7.4(b), or pour le couple LM/`lk-text` la fusion a un effet négatif, tandis que les couples VSM/`lk-slex` et BM25/`lk-slex` montrent les améliorations les plus importantes après la fusion.

Il a été également suggéré que les différents systèmes doivent avoir des performances similaires pour que la fusion soit efficace (Wu et McClean, 2005). Cependant, les résul-

tats obtenus avec la version `lk-free` dans certaines configurations montrent que cette affirmation n'est pas toujours vérifiée empiriquement.

Les mesures proposées par Lee (1997) semblent être celles qui s'adaptent le mieux à notre corpus. Le problème principal du ρ de Spearman, par rapport aux taux de chevauchement, est qu'il ne prend pas en compte la pertinence des documents dans les calculs. De plus, cette mesure cache l'apport immuable de la fusion, à savoir l'augmentation du nombre de documents pertinents retrouvés (*cf.* figure 7.8).

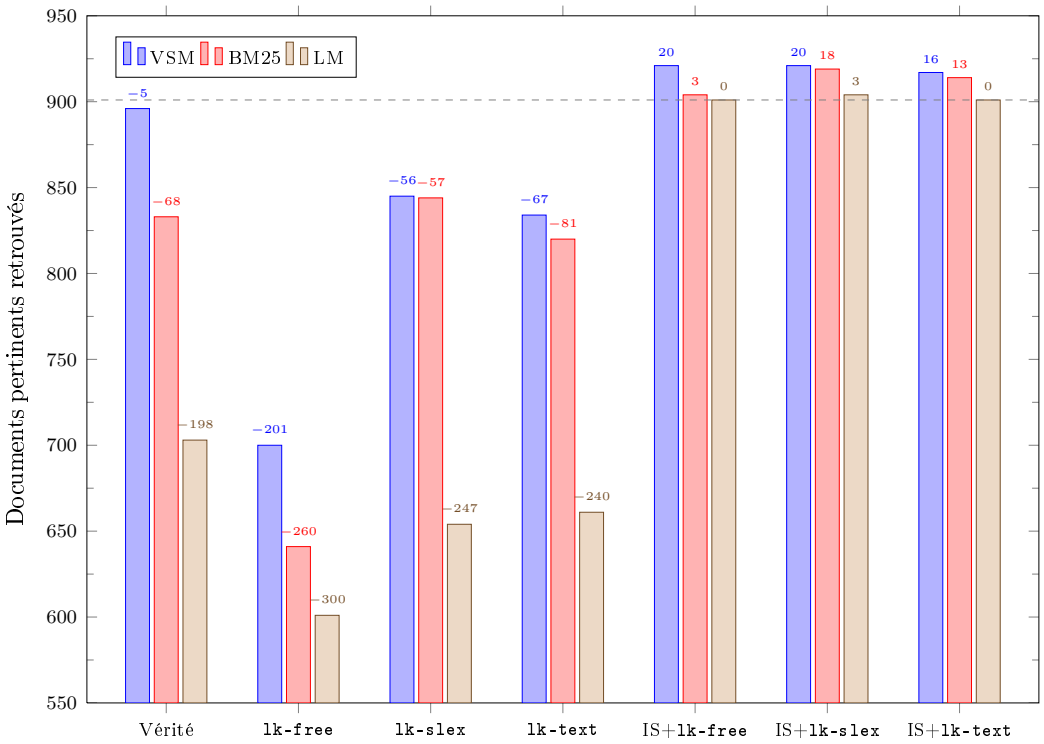


FIGURE 7.8 : Nombre de documents pertinents avant et après fusion. La ligne pointillée correspond au nombre de documents pertinents retrouvés avec IS.

Il faut remarquer que les documents retrouvés par la méthode LM, sont toujours, ou presque, des sous-ensembles de ceux retrouvés par IS. Dans toutes les autres configurations, la fusion introduit des nouveaux documents pertinents, y compris avec la version `lk-free`. Cela permet d'expliquer, dans une certaine mesure, pourquoi avec cette ressource les performances après fusion peuvent être dégradées. En effet, suite aux erreurs de reconnaissance, les documents pertinents introduits dans le résultat final ont un score très inférieur à celui qu'ils auraient dû avoir, ce qui ne permet pas de les reclasser toujours de manière convenable.

7.4 Conclusion

Dans ce chapitre, il a été présenté une étude empirique de la fusion de résultats en recherche d'information appliquée au domaine des documents manuscrits en-ligne.

Les différences systémiques entre les différentes approches de recherche de documents manuscrits sont source de diversité au niveau des documents restitués à l'utilisateur. Cette observation est à l'origine de cette étude et il a été montré expérimentalement qu'il était possible d'améliorer les résultats de la recherche grâce aux méthodes de fusion.

Nos expériences ont été menées sur un ensemble de requêtes générées automatiquement à partir d'un sous-ensemble de notre corpus (*cf.* §3.3.2). Bien que cela constitue une des limites de notre travail, cela permet d'étudier certains aspects liés à la problématique de la recherche de documents manuscrits en-ligne.

Dans le chapitre précédent, l'évaluation des méthodes de référence, se basant sur le résultat d'un moteur de reconnaissance de l'écriture en-ligne, a mis en évidence un impact négatif dans les performances de la recherche d'information. De plus, lorsque les méthodes de recherche sont évaluées individuellement, c'est la méthode de word spotting qui obtient les meilleures performances. Cependant, ces résultats doivent être considérés à la lumière de l'algorithme de génération de requêtes.

D'autre part, lorsque les méthodes de recherche sont combinées, les performances sont améliorées de façon substantielle dans la plupart des configurations étudiées, et ce malgré les nombreuses erreurs de reconnaissance. De plus, dans certains cas, les performances après fusion sont supérieures à celles qui seraient obtenues en n'utilisant que les documents de la vérité terrain. C'est-à-dire une amélioration brute d'environ 5 points en moyenne pour la MAP. Lorsque la version `lk-free` du corpus est utilisée les améliorations sont cependant moins systématiques et moins importantes.

La fusion avec l'opérateur proposé CombHMEAN obtient des performances similaires ou supérieures aux opérateurs classiques dans toutes les configurations où les systèmes VSM et BM25 interviennent. La fusion avec la stratégie de recherche à base de modèles de langage n'est pas concluante. Dans toutes les configurations de fusion où elle intervient, les performances sont dégradées. Il a été constaté que cela était dû principalement au fait que l'ensemble des documents restitués par LM était un sous ensemble de ceux restitués par IS, et ce quelle que soit la version du corpus. A contrario, même avec la ressource `lk-free`, les méthodes VSM et BM25 introduisent systématiquement des nouveaux documents pertinents dans le classement après fusion, même si la MAP peut être dégradée à cause des disparités trop importantes entre les scores des différents systèmes malgré l'étape de normalisation.

Enfin, l'analyse des conditions nécessaires au succès de la fusion suggère qu'il y a une relation entre les erreurs de reconnaissance, la dégradation des performances en RI bruitée et le bénéfice attendu après fusion. Cette relation n'est pas entièrement capturée par les mesures prédictives étudiées, à savoir le taux de chevauchement et le coefficient de corrélation. Des modèles et mesures qui prennent en compte l'influence du bruit dans tout le processus sont souhaitables. Cela constitue une nouvelle voie de recherche pour de futurs travaux.

Chapitre 8

Régularisation de résultats de recherche issus du word spotting

Le word spotting est le modèle le plus couramment employé pour la recherche des documents manuscrits en-ligne. Cependant, la grande majorité des méthodes de word spotting comporte deux limites qui en font un modèle de RI rudimentaire. D'une part, il est incapable de considérer les formes fléchies d'un même mot comme une seule entité. D'autre part, ce modèle se caractérise par l'absence de schémas de pondération, ce qui ne permet pas de favoriser les mots en fonction de leur importance dans le document.

Dans ce chapitre, il sera appliquée une méthode de régularisation de scores, permettant de pallier les limites du word spotting. Il s'agit d'un post-traitement applicable aux résultats d'un algorithme de word spotting tout venant. C'est une méthode issue de la recherche en apprentissage semi-supervisé basé sur les graphes (Kondor et Lafferty, 2002 ; Zhou, Bousquet, Navin Lal, Weston et Schölkopf, 2004a ; Zhou et Schölkopf, 2004 ; Belkin, Matveeva et Niyogi, 2004) et en particulier sur les résultats issus de la théorie spectrale des graphes (Cvetkovic, Doob et Sachs, 1998 ; Chung, 1997). Cette méthode a été appliquée par le passé à l'interrogation par l'exemple d'une base de chiffres manuscrits (Zhou, Weston, Gretton, Bousquet et Schölkopf, 2004b)[1], et plus récemment à la RI (Díaz, 2007, 2008) avec une partie des corpus de la campagne d'évaluation TREC.

La première partie de ce chapitre (§8.1) présente les concepts clés à l'origine de la méthode utilisée, en particulier la matrice laplacienne normalisée d'un graphe. Une deuxième section (§8.2) détaille la méthode proprement dite. L'approche originalement proposée par Zhou et collab. (2004a) s'appuie sur la structure des données, également appelée variété, modélisée par un graphe de similarité entre les documents. Cette approche a de nombreuses connections avec des algorithmes d'analyse des liens (Brin et Page, 1998 ; Kleinberg, 1999 ; Ng, Zheng et Jordan, 2001) et les méthodes spectrales de classification non-supervisée (Shi et Malik, 2000 ; Ng, Jordan et Weiss, 2002 ; von Luxburg, Bousquet et Belkin, 2005 ; von Luxburg, 2007).

1. Il s'agit dans cette étude de la base USPS : `http://www.kernel-machines.org/data/`

Le problème de la RI se pose en termes de régularité, l'hypothèse sous-jacente est qu'une fonction de recherche doit être suffisamment régulière pour que des documents proches se voient affecter des scores de pertinence similaires. La validation expérimentale de cette hypothèse dans le domaine manuscrit en-ligne est détaillée au cours de la section 8.3. Enfin, la dernière section fait la synthèse des contributions présentées dans ce dernier chapitre.

8.1 Théorie spectrale des graphes

La théorie spectrale des graphes est une branche de la théorie algébrique des graphes. Elle s'intéresse principalement à l'utilisation de l'algèbre linéaire pour l'analyse des matrices d'adjacence d'un graphe. Il s'agit de l'étude du rapport entre les caractéristiques structurelles d'un graphe et l'ensemble de ses attributs algébriques : valeurs et vecteurs propres, polynôme caractéristique, etc. (Chung, 1997 ; Cvetkovic et collab., 1998).

Un graphe peut-être représenté par plusieurs matrices, notamment sa matrice d'adjacence et les matrices laplaciennes. En algèbre linéaire, le spectre d'une matrice est l'ensemble de ses valeurs propres, $(\lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1})$. De même, en théorie spectrale des graphes, le spectre d'un graphe est l'ensemble des valeurs propres des matrices d'un graphe.

Parmi les applications de la théorie spectrale des graphes, il y a la segmentation d'images (Shi et Malik, 2000) et l'apprentissage non-supervisé (von Luxburg, 2007) entre autres. Cette section présente les concepts nécessaires à la définition de la méthode de régularisation utilisée. La première sous-section s'intéresse à la construction des graphes de similarité pour la représentation de données. Enfin le concept clé de notre algorithme de régularisation, à savoir la matrice laplacienne normalisée, sera présenté. Cette matrice vérifie de nombreuses propriétés, par exemple, son spectre peut donner une idée du nombre de composantes connexes du graphe et son déterminant correspond au nombre d'arbres couvrants minimaux.

8.1.1 Graphes de similarité

L'idée de caractériser des objets en fonction des rapports de similarité qu'ils entretiennent entre eux est largement appliquée dans le domaine de l'apprentissage automatique et particulièrement en ce qui concerne les méthodes à noyaux (Schölkopf et Smola, 2002).

Étant donné un ensemble de documents $\mathcal{D} = \{j | 1 \leq j \leq n\}$, ou plus généralement d'objets, définis dans un espace quelconque, est appelée noyau la fonction de similarité $\mathcal{K} : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$ qui vérifie les propriétés suivantes

$$\mathcal{K}(i,j) = \mathcal{K}(j,i) \qquad \text{symétrie}$$

et

$$\mathcal{K}(i,j) \geq 0, \quad \forall i,j \in \mathcal{D} \qquad \text{non-négativité}$$

Grâce à $\mathcal{K}$, les éléments de $\mathcal{D}$ pourront être représentés par un graphe pondéré non dirigé $G = (\mathcal{V}, E)$ où chaque sommet $v_i \in \mathcal{V}$ correspond à un élément de $\mathcal{D}$. Il existe une arête $e_{ij} = v_i v_j \in E$ entre deux sommets v_i et v_j si $\mathcal{K}(i,j) > 0$.

Le caractère non dirigé de G découle directement de la symétrie du noyau $\mathcal{K}$. La matrice d'adjacence $\mathbf{W}$ du graphe G est définie par :

$$\mathbf{W}_{ij} = \begin{cases} \mathcal{K}(i,j) & \text{si } i \neq j\,; \\ 0 & \text{sinon.} \end{cases} \tag{8.1}$$

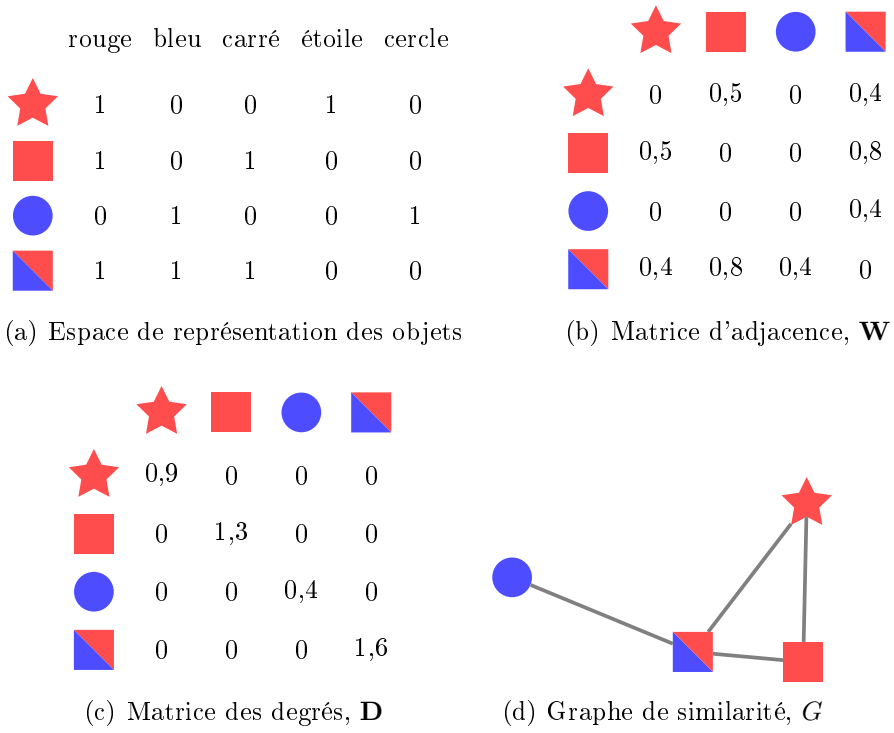
Cette matrice peut être rapprochée de la matrice de Gram utilisée en apprentissage statistique (Schölkopf et Smola, 2002), à l'exception que la diagonale de $\mathbf{W}$ est nulle, c'est-à-dire que G ne possède pas de boucle.

La matrice des degrés $\mathbf{D}$ est une matrice diagonale où l'élément $\mathbf{D}_{ii}$ est défini par le degré du sommet v_i :

$$\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{W}_{ij} \tag{8.2}$$

La figure 8.1 montre la construction de G, $\mathbf{W}$ et $\mathbf{D}$ à partir de l'espace de représentation des objets.

Le graphe G, construit grâce à la formule 8.1, est connexe. Dans la figure 8.1 une arête a été créée pour tous les points qui ont une similarité positive, cela peut conduire à



(a) Espace de représentation des objets

(b) Matrice d'adjacence, $\mathbf{W}$

(c) Matrice des degrés, $\mathbf{D}$

(d) Graphe de similarité, G

FIGURE 8.1 : Représentation par graphe d'un ensemble d'objets.

la construction de graphes où tous les sommets sont reliés entre eux. Toutefois, avec un nombre important d'objets dans des espaces à très grande dimension, le but d'une telle construction est de modéliser le voisinage local d'un objet donné. Dans la suite de ce chapitre, G sera considéré comme étant le graphe des k-plus proches voisins, c'est-à-dire où il existe une arête $e_{ij} = v_i v_j \in E$ entre deux sommets v_i et v_j si et seulement si i fait partie des k-plus proches voisins de j ou vice versa. En plus de modéliser le voisinage local des documents, l'exploitation d'un tel graphe présente beaucoup d'avantages en ce qui concerne la complexité calculatoire car la matrice d'adjacence associée est souvent creuse (Stoer et Bulirsch, 2002, chap. 8).

8.1.2 Matrice laplacienne

Un graphe peut-être représenté par plusieurs matrices. La théorie algébrique des graphes s'intéresse principalement aux matrices d'adjacence, la théorie spectrale des graphes s'intéresse, quant à elle, aux matrices laplaciennes. Dans le cadre de cette thèse, c'est la matrice laplacienne normalisée, dénommée également laplacien dans la suite de ce document, qui sera utilisée.

Définition 8.1. *La matrice laplacienne, normalisée de G, $\mathcal{L}$ est définie par :*

$$\mathcal{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} \tag{8.3}$$

Le laplacien peut être vu comme un opérateur qui induit une forme quadratique de la fonction $f : \mathcal{V} \to \mathbb{R}$, et par extension, du vecteur $\mathbf{f}$ des images de f (Chung, 1997 ; von Luxburg, 2007) :

$$f'\mathcal{L}f = \sum_{i,j=1}^{n} \mathbf{W}_{ij} \left(\frac{f_i}{\sqrt{\mathbf{D}_{ii}}} - \frac{f_j}{\sqrt{\mathbf{D}_{jj}}} \right)^2 \tag{8.4}$$

De plus, le laplacien peut être interprété comme un opérateur qui mesure la différence entre la valeur d'une fonction f en un point i quelconque et la valeur moyenne de f au voisinage i. Il a également été défini comme la dérivée des graphes (Zhou et Schölkopf, 2004).

8.2 Régularisation des scores

La « cluster hypothesis » en RI a été formulée par Jardine et van Rijsbergen (1971) de la façon suivante :

> « les relations qu'entretiennent les documents entre eux véhiculent des renseignements utiles sur la pertinence des documents vis-à-vis des requêtes

(...) des ensembles de documents [similaires] tendent à être pertinents vis-à-vis des mêmes requêtes » [2]

Un écho à cette hypothèse, venu du domaine de l'apprentissage semi-supervisé, est la « cluster assumption » (Chapelle, Weston et Schölkopf, 2003). Il s'agit du postulat selon lequel deux objets proches ont tendance à appartenir à la même classe (Chapelle et collab., 2003 ; Zhou et collab., 2004a). Ce postulat a été exprimé par Zhou et collab. (2004a) en termes de régularité : une fonction de classification doit être suffisamment régulière pour que des objets proches soit classés à l'identique.

La même idée peut être transposée à la recherche d'information de la façon suivante : si des documents proches tendent à être pertinents pour une même requête, alors la fonction f_q devrait être suffisamment régulière pour que des scores similaires soient associés à des documents proches.

8.2.1 Formulation du problème

Étant donné le vecteur initial de scores τ pour un ensemble de documents, nous cherchons à calculer l'ensemble $\mathbf{f}$ des scores réguliers pour ce même ensemble de documents. La régularisation des scores peut se formuler comme un problème d'optimisation continu. Elle consiste à trouver une solution au problème suivant :

$$\mathbf{f}^* = \arg\min_{\mathbf{f}\in\mathbb{R}^n} \mathcal{Q}(\tau, \mathbf{f}) \tag{8.5}$$

avec

$$\mathcal{Q}(\tau, \mathbf{f}) = \mathcal{S}(\mathbf{f}) + \mu\mathcal{E}(\tau, \mathbf{f}) \tag{8.6}$$

Dans l'équation 8.6, $\mathcal{S}$ et $\mathcal{E}$ sont deux fonctions objectif opposées et μ est le paramètre qui contrôle leur importance relative. La première fonction $\mathcal{S}$ pénalise l'irrégularité des scores entre documents voisins. La seconde, quant à elle, pénalise les écarts entre τ et $\mathbf{f}$.

8.2.1.1 Régularité des scores entre documents voisins

La relation de voisinage entre les documents est modélisée par le graphe G, cette relation est représentée par la matrice $\mathbf{W}$. L'ensemble des scores $\mathbf{f}$ est considéré comme régulier si les documents voisins possèdent des scores similaires. Une façon de mesurer cette régularité est d'utiliser le laplacien normalisé du graphe (Díaz, 2008, p. 54) :

$$\mathcal{S}(\mathbf{f}) = \sum_{i,j=1}^{n} \mathbf{W}_{ij} \left(\frac{\mathbf{f}_i}{\sqrt{\mathbf{D}_{ii}}} - \frac{\mathbf{f}_j}{\sqrt{\mathbf{D}_{jj}}} \right)^2 \tag{8.7}$$

La valeur de cette fonction est haute pour des fonctions peu régulières et basse pour des fonctions régulières. Une autre mesure de la régularité des scores est également

2. « the associations between documents convey information about the relevance of documents to requests (...) closely associated documents tend both to belong to the same clusters and to be relevant to the same requests. »

donnée par le laplacien combinatoire (Mohar, 1988 ; Shi et Malik, 2000). Cependant, seul le cas normalisé sera étudié ici. En effet, des recherches théoriques montrent que le laplacien normalisé doit être préféré au combinatoire dans les méthodes spectrales de classification non-supervisée (von Luxburg et collab., 2005).

8.2.1.2 Écart des scores par rapport aux scores initiaux

La seconde fonction objectif, $\mathcal{E}(\tau, \mathbf{f})$, doit pénaliser les écarts importants entre l'ensemble des scores initial, τ, et celui des scores régularisés, $\mathbf{f}$. Elle est définie par la somme des écarts au carré :

$$\mathcal{E}(\tau, \mathbf{f}) = \sum_{i=1}^{n} (\mathbf{f}_i - \tau_i)^2 \tag{8.8}$$

Si seule la fonction $\mathcal{E}$ était minimisée, une solution triviale au problème d'optimisation serait $\mathbf{f} = \tau$. Dans l'autre sens, minimiser uniquement $\mathcal{S}$ reviendrait à trouver un vecteur constant $\mathbf{c}$ tel que $\mathcal{S}(\mathbf{c}) = 0$. C'est pour cette raison que les deux fonctions sont combinées de façon linéaire avec un paramètre qui mesure l'importance donnée à l'un ou l'autre critère.

8.2.2 Minimisation de la fonction objectif

La solution au problème d'optimisation de l'équation 8.5 proposée par Zhou et collab. (2004a) exploite les propriétés du laplacien. La figure 8.2 est une approximation du graphe de la fonction objectif qui montre clairement que le minimum de la fonction se trouve en 0. De par sa définition, la fonction objectif ne peut être négative dans $\mathbb{R}$. Ainsi, une solution peut être trouvée en cherchant un point où $\mathcal{Q}$ accepte des tangentes horizontales, c'est-à-dire un point où la dérivée s'annule.

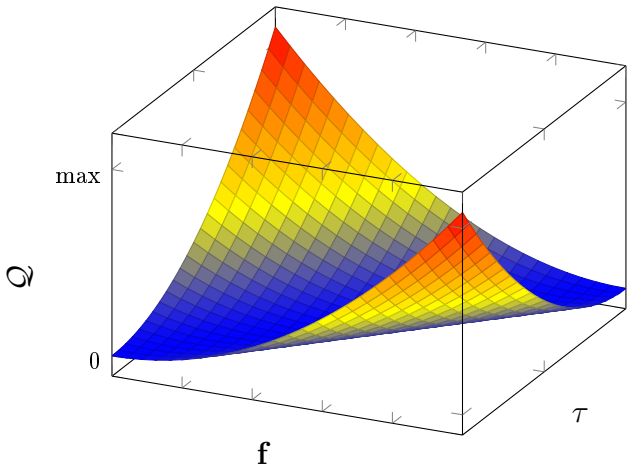


FIGURE 8.2 : Graphique approximatif de la fonction objectif.

Afin de faciliter le calcul, deux variables α et β sont introduites :

$$\alpha = \frac{1}{1+\mu} \tag{8.9}$$

$$\beta = 1 - \alpha \tag{8.10}$$

En considérant la définition du laplacien donnée précédemment, et puisque l'objectif est de chercher le point où la dérivée par rapport à $\mathbf{f}$ s'annule (Zhou et collab., 2004a ; Díaz, 2008),

$$\begin{aligned} \mathcal{L}\mathbf{f}' + \mu(\mathbf{f}' - \tau) &= 0 \\ \alpha\mathcal{L}\mathbf{f}' + \beta\mathbf{f}' - \beta\tau &= 0 \\ (\alpha\mathcal{L} + \beta\mathbf{I})\mathbf{f}' &= \beta\tau \\ \mathbf{f}' &= \beta(\alpha\mathcal{L} + \beta\mathbf{I})^{-1}\tau \\ \mathbf{f}' &= (1-\alpha)(\alpha\mathcal{L} + (1-\alpha)\mathbf{I})^{-1}\tau \end{aligned} \tag{8.11}$$

Une solution itérative peut également être trouvée, elle est formulée de la façon suivante (Zhou et collab., 2004a,b) :

$$\mathbf{f}^{t+1} = (1-\alpha)\tau + \alpha(\mathcal{L} - \mathbf{I})\mathbf{f}^t \tag{8.12}$$

avec $\mathbf{f}^0 = \tau$. Il faut noter que la solution itérative converge vers la solution donnée par l'équation 8.11 (Zhou et collab., 2004a ; Zhou et Schölkopf, 2004). Elle peut s'avérer utile pour des collections de documents de taille très importante. En effet, cette solution permet d'obtenir de bons résultats en peu d'itérations tout en évitant l'inversion coûteuse de la matrice $\alpha\mathcal{L} + \beta\mathbf{I}$. Eu égard à la taille de notre corpus, l'équation 8.11 sera utilisée dans le cadre de nos expériences.

8.3 Expériences

La méthode qui vient d'être présentée a été appliquée aux résultats de recherche obtenus avec MyScript® InkSearch®. Dans cette section seront présentés les résultats des différentes expériences menées afin d'évaluer l'efficacité de la stratégie de régularisation présentée. Contrairement aux résultats présentés précédemment, ces expérimentations n'utilisent pas un ensemble de requêtes générées automatiquement, mais utilisent le nom des catégories comme des requêtes. Dans un premier temps, il s'agira de présenter l'ensemble des requêtes utilisées dans ces expériences (§8.3.1). Enfin, les paramètres expérimentaux (§8.3.2) et les résultats obtenus (§8.3.3) seront détaillés. Les mesures choisies pour évaluer les améliorations obtenues avec la méthode proposée sont la précision moyenne, les courbes de précision-rappel ainsi que les mesures dites de haute précision.

8.3.1 Requêtes

Les expérimentations présentées dans les chapitres 6 et 7 utilisent un ensemble de requêtes générées automatiquement. Outre le fait de ne pas avoir été produites par un

humain, ces requêtes avaient également le défaut de pouvoir favoriser les approches de word spotting de par la façon dont elles sont générées (§3.3.2). Dans les expériences présentées dans ce chapitre, il a été choisi d'utiliser les noms des catégories en tant que requêtes. Le tableau 8.1 montre les requêtes soumises au système de word spotting de référence dans les expériences sur la régularisation.

TABLE 8.1 : Requêtes soumises à MyScript® InkSearch® pour les expériences sur la régularisation.

Identifiant	Requête
1	Earnings
2	Acquisition
3	Grain
4	Money
5	Crude
6	Interest
7	Trade
8	Ship
9	Sugar
10	Coffee

Ces requêtes correspondent à chacune des catégories présentées précédemment (§3.2.2). Comparativement aux requêtes générées, la recherche à partir de ces nouvelles requêtes est plus difficile. En effet, le nom de la catégorie n'est pas forcément un bon indicateur des documents qu'elle contient. Par exemple, *earnings* n'est pas un mot qui apparaît plus dans les documents de la classe `earn` que dans les autres. De plus, les requêtes ont été soumises sans racinisation, alors que les systèmes de word spotting sont connus pour ne pas considérer les formes fléchies d'un mot comme un seul terme (Vinciarelli, 2005a). Cela veut dire, par exemple, que pour la requête *acquisition* le système ne peux pas restituer les documents qui contiennent le mot *acquire*.

8.3.2 Paramètres expérimentaux

L'algorithme de régularisation présenté ici nécessite la spécification de trois paramètres clés : la mesure de similarité, $\mathcal{K}$, le nombre de voisins considérés pour la construction du graphe de similarité, k, et le paramètre $\alpha \in [0, 1]$ qui contrôle l'importance donnée à chacune des fonctions composant la fonction objectif. Lorsque α vaut 0, les scores ne sont pas régularisés. Plus la valeur de α s'approche de 1, plus le score d'un document est déterminé par celui de ses proches voisins.

En ce qui concerne $\mathcal{K}$, il a été choisi de mesurer l'affinité entre deux documents i et j par la mesure de similarité vectorielle traditionnelle donnée par le cosinus de l'angle entre deux vecteurs :

$$\mathcal{K}(i,j) = \frac{\mathbf{d}_i \mathbf{d}_j}{\|\mathbf{d}_i\| \, \|\mathbf{d}_j\|} \tag{8.13}$$

Le tableau 8.2 montre les intervalles de recherche des paramètres de l'algorithme. En ce qui concerne la construction du graphe de similarité, il a été considéré entre 5 et 25 voisins par pas de 5. En ce qui concerne le paramètre de régularisation, α, toutes les valeurs entre 0,1 et 0,9 avec un pas de 0,1 ont été inspectées.

TABLE 8.2 : Intervalles de recherche des paramètres de l'algorithme de régularisation.

Paramètre	Intervalle	Pas
α	[0,1, 0,9]	$0,1$
k	[5, 25]	5

8.3.3 Résultats

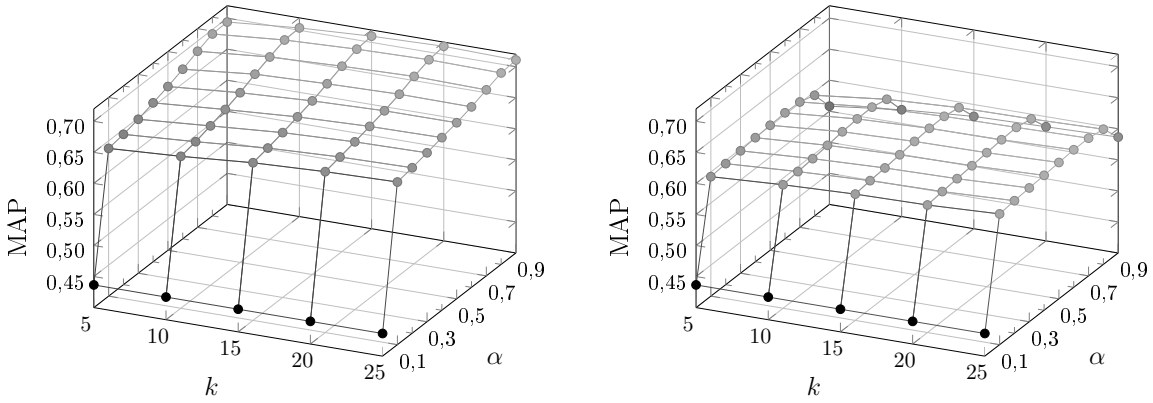
Dans un premier temps, ce sont les effets de la régularisation sur la précision moyenne qui ont été examinés (cf. §6.3). La figure 8.3 montre les améliorations en termes de précision moyenne obtenues après régularisation des scores initiaux restitués par MyScript® InkSearch®.

La figure 8.3 montre que l'application de l'algorithme de régularisation permet d'améliorer les résultats dans toutes les configurations. Pour une valeur fixe du paramètre α, les variations au niveau de la MAP en fonction du nombre de voisins sont très faibles. En effet, l'algorithme est très robuste vis-à-vis du paramètre k dans l'intervalle considéré.

Réciproquement, pour un nombre fixe de voisins, les performances varient de façon plus importante en fonction de la valeur du paramètre de régularisation. Les meilleurs résultats sont obtenus pour des valeurs de α élevées. Ces résultats concordent avec ceux présentés par Zhou et collab. (2004b,a), où de très bons résultats sont obtenus en fixant tout simplement α à 0,99.

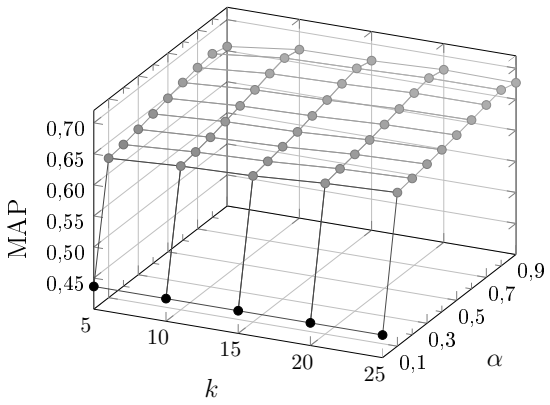
Indépendamment de la valeur des paramètres, l'application de l'algorithme de régularisation permet d'améliorer les résultats de la recherche. La figure 8.4 montre l'importance de ces améliorations. Dans la pire des configurations, la MAP est augmentée d'environ 37 % alors que dans la configuration idéale, l'amélioration atteint plus de 60 % des performances initiales.

Ces résultats montrent également que la qualité de la reconnaissance a une influence sur les améliorations attendues. Comme il a été vu dans le chapitre 5, elle a un impact sur la structure des données, structure qui est exploitée par l'algorithme de régularisation. Lorsque les documents de `lk-free` sont utilisés, la MAP commence à décroître lorsque la valeur de α est supérieure à 0,7. Avec les documents de `lk-text` et `lk-slex` le même phénomène se produit à partir de $\alpha > 0{,}8$. A contrario, cela n'est pas observé avec les documents de la vérité terrain.

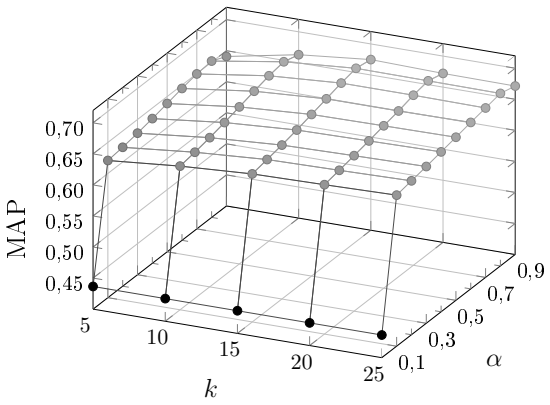


(a) Réf.

(b) `lk-free`



(c) `lk-slex`



(d) `lk-text`

FIGURE 8.3 : Précision des précisions moyennes (MAP) en fonction du nombre de voisins considérés et de la valeur de α. La première ligne de chaque graphique correspond à $\alpha = 0$, c'est-à-dire aux scores non-régularisés.

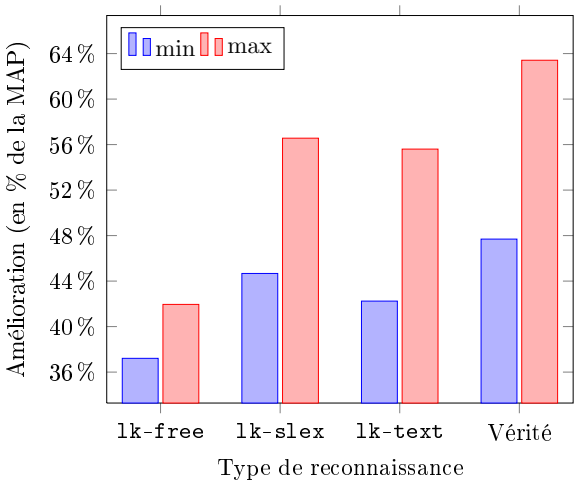


FIGURE 8.4 : Valeurs minimum et maximum des améliorations en % de la MAP initiale.

Ces résultats tendent à montrer que lorsque le travail est effectué avec des documents bruités, la valeur de α doit refléter la confiance portée à la structure des données *malgré* l'impact des erreurs de reconnaissance. Le tableau 8.3 récapitule les paramètres optimaux selon les différentes versions du corpus.

TABLE 8.3 : Paramètres optimaux selon la ressource utilisée pour la reconnaissance.

Reco.	k	α	MAP
`lk-free`	25	0,7	0,6218
`lk-slex`	25	0,8	0,6834
`lk-text`	15	0,8	0,6792
Réf.	20	0,9	0,7133

Afin de savoir quelle est la nature des améliorations, il a été étudié d'une part la précision moyenne au niveau requête, et d'autre part, les mesures de haute précision ainsi que les courbes de précision-rappel pour les trois configurations optimales utilisant le résultat de la reconnaissance pour la construction du graphe de similarité. La figure 8.5 montre la précision moyenne par catégorie pour MyScript® InkSearch® avec ou sans l'étape de régularisation.

Il faut noter que pour 3 catégories (*trade*, *sugar* et *coffee*), les noms sont de très bons indicateurs des documents appartenant à la classe. Pour ces 3 requêtes, les améliorations au niveau de la MAP sont en moyenne très faibles et sont dues à quelques reclassements de documents. Pour les requêtes restantes, les améliorations sont plus nettes en particulier pour les deux catégories comptabilisant le plus d'effectifs.

Lorsque les améliorations sont évaluées en termes de haute précision, les résultats sont moins évidents. La figure 8.6 montre la précision à n documents pour les trois configurations basées sur le résultat de la reconnaissance du tableau 8.3. La précision est calculée par `trec_eval` à des points de coupure allant de 5 à 1 000 documents. Pour

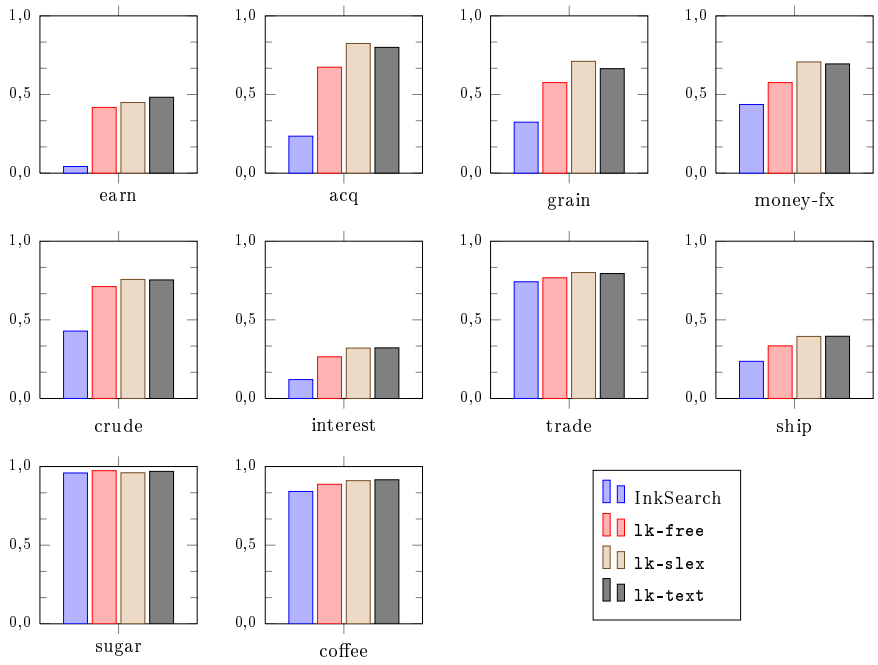


FIGURE 8.5 : Précision moyenne par requête avec ou sans régularisation et selon la ressource utilisée pour la reconnaissance et la construction du graphe de similarité.

un utilisateur qui interrogerait notre base de documents, cela reviendrait à inspecter la moitié du corpus, pour cette raison 100 documents maximum sont considérés dans les résultats présentés en figure 8.6.

Jusqu'à 15 documents, les améliorations au niveau de la p@n ne concernent que très peu de catégories. La plupart du temps la précision ne change pas, voire se dégrade dans quelques cas. En moyenne, l'amélioration de la p@n en fonction de la valeur de n varie entre 1 % et 5 % en fonction de la ressource utilisée pour la reconnaissance. En comparaison des améliorations en termes de précision moyenne, les améliorations en termes de p@n sont faibles. Lorsque plus de 15 documents sont considérés pour le calcul de la précision, le nombre et l'intensité des améliorations tend à augmenter. En réalité, les améliorations ne deviennent durables et uniformes qu'à partir de 100 documents. Cela montre que l'algorithme proposé est plus cohérent pour des taux de rappel au delà de 30 documents. Il ne peut pas remonter brusquement des documents pertinents dans les toutes premières positions, en revanche il peut introduire de manière uniforme des documents pertinents dans la partie centrale du classement.

La figure 8.7 confirme cette première analyse. En effet, en se référant à la courbe de précision-rappel de base de IS, il peut être observé que la partie bombée est tournée vers le bas. Cette convexité indique que la recherche avec InkSearch n'est pas très robuste pour les valeurs du rappel entre 20 % et 80 %. Il faut également remarquer dans cette figure que les courbes correspondant à la recherche avec régularisation sont plutôt concaves. Cela veut dire que la régularisation améliore de manière uniforme et importante la précision dans les mêmes intervalles où l'algorithme de word spotting est défaillant.

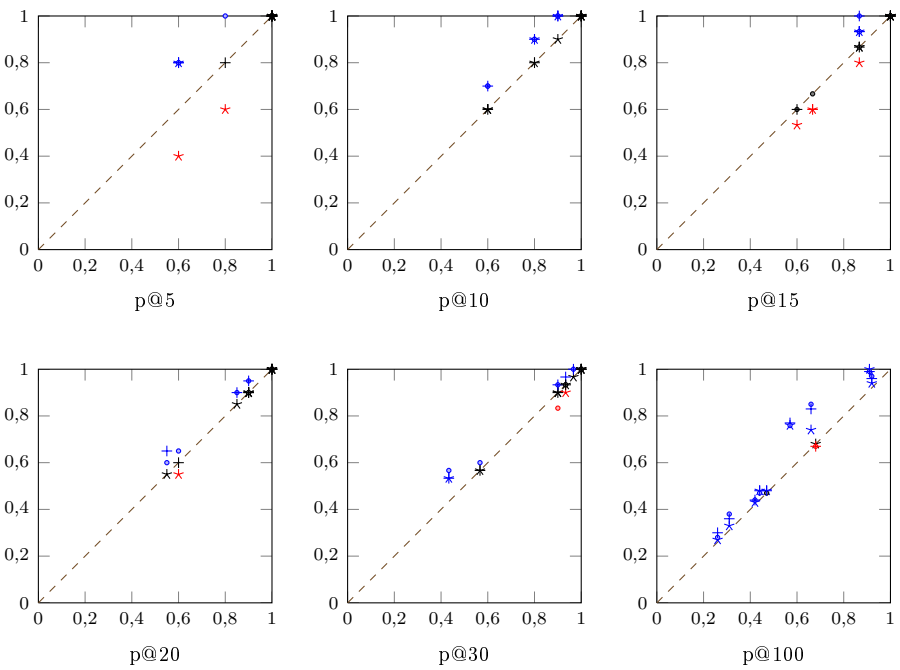


FIGURE 8.6 : Précision à n documents pour les configurations optimales de l'algorithme de régularisation. L'axe horizontal correspond à la précision de IS, l'axe vertical correspond à la précision après l'étape de régularisation. Chaque point correspond à une requête, un point bleu indique une amélioration, un point rouge une dégradation, un point noir indique des performances égales. Légende : `lk-free` ⋆, `lk-slex` ●, `lk-text` +.

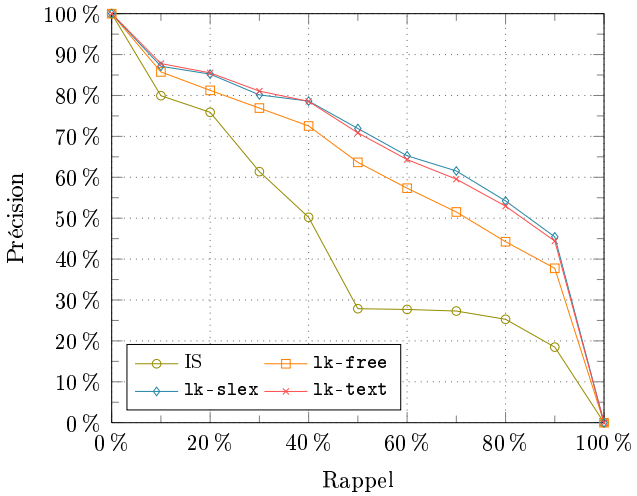


FIGURE 8.7 : Courbes de précision vs rappel pour MyScript® InkSearch® avec ou sans l'étape de régularisation.

8.4 Conclusion

Il a été présenté dans ce chapitre une méthode de régularisation des scores de recherche et son application aux résultats d'un algorithme de word spotting. L'application de cet algorithme se justifie ici par les limites de l'approche dominante de la recherche de documents manuscrits, à savoir le word spotting. En effet, la détection de mots-clés dans une collection de documents n'est pas un mécanisme suffisant pour répondre à un besoin d'information (Baeza-Yates et Ribeiro-Neto, 1999, p. 1). L'approche qui vient d'être proposée permet de faire face aux limites du word spotting, ainsi, son application devrait permettre d'obtenir des améliorations significatives en ce qui concerne le nombre de documents pertinents restitués en réponse à un besoin d'information.

La méthode de régularisation proposée exploite la structure des données, modélisée par un graphe de similarité. L'algorithme de régularisation est issu de la recherche en apprentissage semi-supervisé. Il a de nombreuses connexions avec des méthodes d'analyse des liens (Brin et Page, 1998 ; Kleinberg, 1999 ; Ng et collab., 2001) et des méthodes spectrales de classification non-supervisée (Shi et Malik, 2000 ; Ng et collab., 2002 ; von Luxburg et collab., 2005 ; von Luxburg, 2007).

Ensuite ont été présentés les résultats expérimentaux de l'application de la méthode de régularisation sur le corpus manuscrit qui a servi de base à cette étude. Comme cette base ne possède pas un ensemble de requêtes prédéfini avec des ensembles de documents pertinents correspondants, ce sont les noms des catégories du corpus qui ont été utilisés comme requêtes.

Les résultats montrent une amélioration très importante des performances en termes de précision moyenne. En effet, selon les configurations, cette amélioration se situe entre 37 % et 64 % de la précision moyenne observée avant l'étape de régularisation. Lorsque le système est évalué grâce aux mesures de haute précision, les améliorations sont beaucoup moins importantes : une amélioration absolue de seulement 1 % à 5 %. Cependant, même si l'algorithme ne peut pas améliorer de manière significative les résultats dans les toutes premières positions du classement, il peut introduire de manière uniforme, constante et significative, des documents pertinents dans la partie centrale du classement.

La répercussion des erreurs de reconnaissance sur l'amélioration des résultats de recherche a également été observée. Puisque les erreurs de reconnaissance ont un impact sur la structure des données, elles en ont aussi un sur l'algorithme de régularisation. Naturellement, les versions du corpus affichant les taux d'erreur les plus bas permettent d'améliorer les résultats de façon plus importante. Cependant, la présence de nombreuses erreurs de reconnaissance n'est pas rédhibitoire puisqu'en construisant le graphe de similarité avec les documents de `lk-free`, la précision moyenne est améliorée de 42 % grâce à la méthode proposée, par rapport aux performances de base, dans la configuration optimale.

Un des points clés de la méthode qui vient d'être proposée est la mesure de similarité $\mathcal{K}$. Les expériences présentées ici utilisent une mesure textuelle de similarité, mais cela n'interdit pas d'envisager une mesure similarité en fonction de critères différents comme la structure des documents, l'identité du scripteur ou les croquis, tableaux, équations, etc. À partir du moment où il est possible de détecter les différents blocs d'un document,

il devient possible de concevoir des systèmes de recherche particuliers en exploitant les caractéristiques des blocs non-textuels présents dans les documents. Cela veut dire que ces systèmes serait capables de retrouver des documents qui contiennent les mots-clés de la requête mais qui partagent en plus des caractéristiques communes comme des structures similaires, ou des styles d'écriture similaires.

Chapitre 9

Conclusion

L'étude qui vient d'être présentée porte sur la question de l'accès à l'information textuelle contenue dans des bases documentaires manuscrites en-ligne. Les documents en-ligne sont un nouveau média issu de l'émergence de dispositifs de saisie comme les stylos digitaux, tableaux blancs interactifs, etc. L'objectif principal de cette thèse était de proposer des modèles et méthodes permettant d'avoir accès à l'information contenue dans ce nouveau type de données.

Un bilan de chacune des différentes parties abordées dans ce document, ainsi que les perspectives de ce travail, seront présentés dans la suite de ce chapitre.

9.1 Bilan

Notre intérêt a porté sur deux questions particulières : la catégorisation et la RI ad-hoc. L'objectif de la catégorisation est de classer de façon automatique les documents dans des catégories définies au préalable par un humain. Les résultats de la catégorisation peuvent être utiles aussi bien pour la recherche d'information que pour l'extraction de connaissances. En ce qui concerne la RI ad-hoc, le but est de pouvoir interroger une base de documents manuscrits afin de sélectionner les documents qui correspondent aux critères de recherche propres à un utilisateur.

9.1.1 Collecte de données et reconnaissance de l'écriture

La première partie de ce mémoire, composée des chapitres 2 et 3, a introduit les deux éléments transversaux à cette étude : la reconnaissance de l'écriture et les données expérimentales. Le premier chapitre s'est concentré sur la question de la reconnaissance et son évaluation. En effet, la reconnaissance est un processus faillible et sa qualité varie en fonction des stratégies mises en œuvre. De ce fait, les transcriptions qui résultent de ce processus sont *bruitées*, c'est-à-dire qu'elles contiennent des erreurs. L'évaluation consiste alors à quantifier ce bruit à l'aide de différentes mesures.

Paradoxalement, les recherches visant l'accès à l'information contenue dans des données en-ligne se heurtent à l'absence de données expérimentales. La première étape de cette étude a donc été la collecte des données nécessaires à sa validation expérimentale. Cette collecte a permis de combler un manque flagrant de données disponibles pour la conduite de telles recherches. C'est un sous-ensemble du corpus Reuters-21578 qui a été utilisé comme base pour la collecte qui elle, a mobilisé environ 1 500 scripteurs. Au total, nous avons obtenu un corpus de 2 310 documents manuscrits, soit environ 24 484 lignes d'écriture ou 166 619 mots.

9.1.2 Catégorisation de documents manuscrits en-ligne

La seconde partie a porté sur la catégorisation de textes. Dans un premier temps, nous avons effectué la reconnaissance des données du corpus, puis étudié la perturbation induite par les erreurs de reconnaissance. Nous avons montré que les mesures classiques du bruit sont à considérer avec précaution. En effet, le fait de calculer la moyenne d'une mesure sur l'ensemble des documents peut dissimuler des effets liés à certains échantillons, et fausser l'interprétation des résultats.

Nous nous sommes ensuite intéressés à l'effet du bruit sur chacune des étapes de la catégorisation. D'abord au niveau de la tokenisation, où nous avons vu que le bruit se traduit par l'introduction importante d'hapax. Par ailleurs, nous avons exploré les effets du bruit sur la représentation et la structure des données. Nous avons cherché à donner une représentation graphique intuitive du phénomène, permettant de mieux saisir le processus en jeu. Enfin, nous avons appliqué des algorithmes d'apprentissage à ces données. Nous avons étudié le biais introduit dans la phase d'apprentissage, ainsi que l'impact du bruit sur les performances pendant la validation. Les résultats montrent que lorsque 75 % des termes sont bien reconnus, l'impact sur la catégorisation est négligeable. En revanche, lorsque plus de la moitié des termes sont perdus, les performances sont dégradées de façon considérable.

9.1.3 Recherche d'information et documents en-ligne

Enfin, la troisième et dernière partie, composée des chapitres 6 à 8, s'est intéressée à la recherche d'information dans des bases documentaires manuscrites en-ligne. Par rapport à l'approche classique du domaine, qui consiste à détecter les mots-clés dans les documents, nous avons placé notre regard à un niveau d'abstraction supérieur. La question n'était pas « comment peut-on retrouver des documents ? », mais « comment tirer le meilleur parti des mécanismes de recherche existants ? ». Pour répondre à cette question nous avons utilisé des méthodes de fusion de résultats ainsi que de régularisation du word spotting.

Fusion de résultats

L'idée de la fusion s'appuie sur la diversité des résultats que peuvent retourner différents systèmes de recherche. Nous avons dressé une typologie des approches existantes pour la recherche de documents manuscrits et montré pourquoi leurs différences systé-

miques sont source de diversité. Ainsi, nous avons fait l'hypothèse qu'en combinant les résultats de ces différentes approches, nous pouvions améliorer les résultats restitués à l'utilisateur.

Nous avons vérifié cette hypothèse sur un ensemble de requêtes générées automatiquement à partir des documents issus de la vérité terrain. Les résultats montrent que la fusion des résultats permet d'améliorer quasi-systématiquement les performances et ce, malgré les nombreuses erreurs de reconnaissance. De plus, dans certains cas, les performances après fusion sont supérieures à celles qui seraient obtenues en n'utilisant que les documents de la vérité terrain.

Régularisation

Nous avons montré que le word spotting, par sa nature, n'est pas une stratégie suffisante pour répondre à une demande d'information. L'approche par régularisation a été conçue de façon à pallier les défauts des approches de word spotting. Cette approche est une étape de post-traitement de résultats, applicable à des approches arbitraires de word spotting. Elle s'appuie sur les relations thématiques entre les documents, modélisées par un graphe de similarité, afin de régulariser les résultats initiaux. L'objectif est alors de recalculer les scores initiaux afin que les documents aux thématiques proches obtiennent des scores de pertinence proches.

Les expériences sur la régularisation utilisent le nom des catégories en tant que requêtes. Les résultats montrent une amélioration très importante des performances en termes de précision moyenne. Les améliorations apportées par cette méthode sont plus visibles au niveau du rappel. De plus, notre approche est très robuste face aux erreurs de reconnaissance car, même avec des documents fortement dégradés, nous avons observé des améliorations dans toutes les configurations.

9.2 Perspectives

Bien que les documents en-ligne soient apparus depuis plusieurs années, ils ne sont l'objet d'étude de la RI que depuis peu de temps. Par ailleurs, le domaine de l'analyse de textes bruités est également un domaine émergeant. Cela ouvre un large éventail de perspectives.

9.2.1 Modéliser le bruit

Jusqu'ici nous avons étudié le bruit expérimentalement. Cela nous a permis d'affirmer qu'il était possible d'effectuer la catégorisation sur les résultats de la reconnaissance sans perte significative de performances. Cependant, les méthodes expérimentales peinent à prédire de manière précise cet impact du bruit. Cela conduit à penser que la question du bruit devrait être abordée d'un point de vue fondamental.

Une façon d'envisager le problème est de modéliser le bruit par une variable aléatoire. La compréhension des effets du bruit sur la représentation des données peut passer par une telle modélisation, comme le montrent les travaux de Mittendorf (1998), peu connus malgré leur intérêt théorique.

Une autre façon de procéder est de s'intéresser aux perturbations de la fonction de similarité. Soit $\mathbf{A}$ la matrice des documents électroniques et $\tilde{\mathbf{A}}$ celles des documents bruités, et $\mathcal{K}$ et $\tilde{\mathcal{K}}$ les fonctions noyaux associées. L'idéal serait que les variations de $\mathbf{A}$ n'induisent qu'une faible variation de la mesure de similarité. En posant $\mathbf{K}$ la matrice de similarité de la collection définie par :

$$\mathbf{K}_{ij} = \mathcal{K}(i, j) \tag{9.1}$$

Cette idée peut être formalisée de la façon suivante :

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \leq \frac{\|\mathbf{A} - \tilde{\mathbf{A}}\|}{\|\mathbf{A}\|} \tag{9.2}$$

S'il peut être prouvé que cette inégalité est vraie, cela veut dire que l'impact du bruit est moins important sur la matrice de similarité $\mathbf{K}$, celle qui détermine le processus de catégorisation, que sur la représentation des données. De plus, cette inégalité peut permettre de trouver des bornes ou des relations de proportionnalité permettant de mieux comprendre le comportement des méthodes de classification face aux erreurs de reconnaissance.

9.2.2 Effectuer la RI à partir de requêtes produites par un humain

En ce qui concerne la RI, le principal inconvénient de nos expériences est le besoin soit de générer les requêtes, soit d'utiliser les noms de catégories en tant que requêtes. D'une part, parce que ces requêtes sont difficilement interprétables d'un point de vue humain et d'autre part parce qu'elles ne sont pas à même de couvrir la richesse du spectre lexical que pourrait utiliser un être humain. Un autre inconvénient est celui du faible nombre de requêtes, celles-ci étant limitées par le nombre de catégories.

Pour faire face à ces inconvénients, la priorité sera donnée à la construction d'un ensemble de requêtes produites par un humain et du jugement de pertinence par un panel d'experts. Une première idée consiste à utiliser les mêmes requêtes que celles de Vinciarelli (2005a). En effet, puisque ce travail et le nôtre se basent sur le même corpus de référence (Reuters-21578), ces requêtes sont également appropriées pour notre travail. De plus, en mettant en parallèle les deux corpus, nous pouvons exploiter les jugements de pertinence existants. Ce travail est actuellement en cours de réalisation et doit conduire à la création d'un corpus standardisé et à sa mise à disposition pour la communauté.

Une seconde idée consisterait à définir nos propres requêtes, et avoir recours au *crowdsourcing*[1] pour l'obtention de jugements de pertinence pour chacune des requêtes (Alonso, Rose et Stewart, 2008).

1. Il s'agit d'externaliser la tâche du jugement de pertinence à un grand nombre d'internautes. Voir http://fr.wikipedia.org/wiki/Crowdsourcing

9.2.3 Affiner les stratégies de recherche en fonction de critères applicatifs

La méthode de régularisation proposée est suffisamment souple et adaptable. Les expériences que nous avons présentées se basent sur une similarité cosinus et des vecteurs de termes issus de la reconnaissance. Mais il suffit de mesurer cette similarité en fonction d'autres critères pour changer le *sens* des résultats restitués. Nous pouvons par exemple imaginer une mesure de similarité basée sur l'agencement des différents blocs de contenu, ainsi, les documents restitués à l'utilisateur auront non seulement tendance à contenir les mots-clés de la requête mais aussi à avoir un agencement similaire.

Tout en restant dans le cadre des vecteurs de termes, nous pourrions également envisager l'utilisation d'autres mesures de similarité comme le noyau polynomial ou gaussien. De plus, nous ne sommes pas obligés de cantonner la régularisation au word spotting. En effet, il est possible d'étendre l'application de la régularisation aux résultats de la RI bruitée.

9.2.4 Définir une similarité textuelle sans reconnaissance

L'ensemble des problèmes traités ici ont un point en commun : la notion de similarité. Le point clé du word spotting est de trouver une mesure de similarité entre le motif requête et les motifs de la base de données. De même, le point clé de la RI est la mesure de similarité entre la requête et les documents de la base. Dans les méthodes de classification c'est également la fonction noyau qui est au cœur des algorithmes.

Les travaux que nous avons présentés s'appuient toujours sur un système de reconnaissance. Cependant, dans certains cas il serait souhaitable d'éviter le processus de reconnaissance. Une autre perspective de ce travail consisterait à définir des mesures de similarité entre documents sans reconnaissance. Nous resterons toujours dans le cadre de la textualité car la similarité sera toujours définie en fonction des mots.

Si le word spotting est capable de déterminer quels motifs d'une base de données correspondent à un motif requête, cela veut dire qu'il est également capable de déterminer quels sont les motifs communs à deux documents. L'idée est de créer des classes d'équivalence de mots qui serviront de descripteurs dans les vecteurs correspondant aux deux documents (cf. figure 9.1).

Dans l'idéal, comme le montre la figure 9.1, les classes d'équivalence contiendraient des mots manuscrits qui partagent le plus long *préfixe encre* commun. Les distances élastiques ouvertes peuvent se prêter à ce type d'exercice (Tormene, Giorgino, Quaglini et Stefanelli, 2008). Ces classes d'équivalence permettent de construire des représentations vectorielles des documents. Le calcul de la similarité peut alors s'effectuer de façon classique.

La définition d'une telle mesure permettrait d'ouvrir les applications de gestion de contenu de haut niveau, basées sur l'apprentissage automatique, aux documents manuscrits tout en évitant l'étape de reconnaissance. En particulier, les méthodes à noyaux deviendraient directement utilisables.

Cependant, dans un premier temps, la possibilité d'une mesure de similarité textuelle au niveau encre devrait être explorée dans un cadre monoscripteur. En effet, les

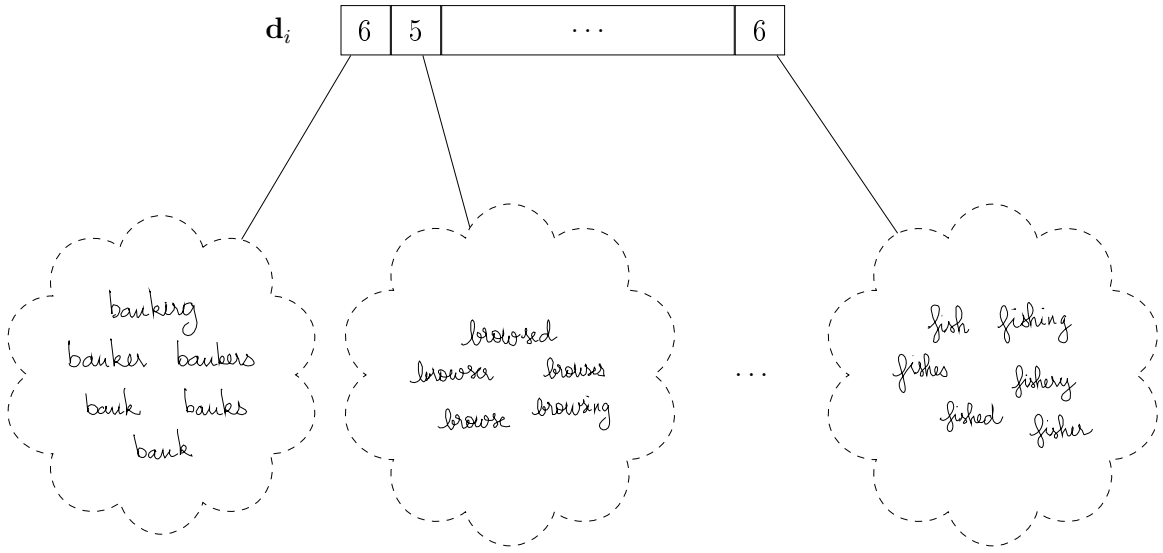


FIGURE 9.1 : Classes d'équivalence (clusters) pour les termes *bank*, *browse* et *fish* et leur utilisation en tant que descripteurs pour la création du vecteur d'un document.

distances élastiques sont plus adaptées à ce cadre. Ces techniques peuvent également rencontrer des difficultés d'une part pour gérer indistinctement les majuscules et les minuscules, et pour faire face à des langues fortement flexionnelles d'autre part.

De nombreuses perspectives d'extension de nos travaux ont été évoquées, mais dans l'immédiat nous envisageons de concentrer nos efforts sur la création et la mise à disposition d'un corpus standardisé à partir des requêtes proposées par Vinciarelli (2005a).

Annexe A

Sélection de termes

La sélection de termes est un problème crucial pour la catégorisation et l'apprentissage automatique en général. D'une part, certains algorithmes ont des difficultés à gérer un espace de caractérisation de documents trop important. D'autre part, les termes qui apportent peu ou pas d'information peuvent induire en erreur les algorithmes d'apprentissage. Notre stratégie de sélection de termes se base sur la statistique du χ^2 et l'algorithme de Forman (2004). Ces deux procédures sont détaillées ci-dessous.

A.1 Le test du χ^2

Le test statistique du χ^2 mesure l'écart entre un ensemble théorique et un ensemble réel d'informations réparties en classes. En fonction de cet écart, il possible de rejeter ou de valider une hypothèse de départ. En médecine et sciences humaines, il est souvent utilisé comme test d'indépendance.

A.1.1 Le test d'indépendance

À la base d'un test statistique il y a la formulation d'une hypothèse appelée hypothèse nulle (h_0). Dans le cas d'un test d'indépendance, h_0 est l'indépendance de deux variables aléatoires. Par exemple, l'appartenance d'un document à la classe `earn` (X) est indépendante de l'apparition du mot `profit` dans le document (Y).

Les étapes du calcul du test, à partir de l'ensemble de données réparti en classes, se déroule en 4 étapes :

1. Déterminer le nombre de degrés de liberté, noté ν, du problème. Il s'agit du nombre de variables aléatoires moins 1.
2. Calculer la distance entre l'ensemble théorique et l'ensemble de données observé.
3. Choisir a priori une probabilité d'erreur, r. La valeur 0,05 est souvent choisie.
4. À l'aide d'une table de distribution du χ^2 (tableau A.1), déduire la distance critique à partir de ν et la probabilité d'erreur.

Si la distance calculée est supérieure à la distance critique, le résultat n'est pas dû seulement aux fluctuations d'échantillonnage et l'hypothèse nulle doit donc être rejetée. Dans le cas de l'exemple précédent, la conclusion serait que l'appartenance d'un document à la classe `earn` n'est pas indépendante de l'apparition du mot `profit` dans le document.

TABLE A.1 : Table des valeurs critiques du χ^2 jusqu'à 20 degrés de liberté.

$\nu \backslash r$	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01
1	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0,2110	0,4460	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210
3	0,5840	1,0050	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,341
4	1,0640	1,6490	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277
5	1,6100	2,3430	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086
6	2,2040	3,0700	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812
7	2,8330	3,8220	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475
8	3,4900	4,5940	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090
9	4,1680	5,3800	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666
10	4,8650	6,1790	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209
11	5,5780	6,9890	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725
12	6,3040	7,8070	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217
13	7,0420	8,6340	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688
14	7,7900	9,4670	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141
15	8,5470	10,3070	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578
16	9,3120	11,1520	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000
17	10,0850	12,0020	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409
18	10,8650	12,8570	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805
19	11,6510	13,7160	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191
20	12,4430	14,5780	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566
21	13,2400	15,4450	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932
22	14,0410	16,3140	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289
23	14,8480	17,1870	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638
24	15,6590	18,0620	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980
25	16,4730	18,9400	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314
26	17,2920	19,8200	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642
27	18,1140	20,7030	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963
28	18,9390	21,5880	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278
29	19,7680	22,4750	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588
30	20,5990	23,3640	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892

A.1.2 Un exemple

Soit les deux variables aléatoires suivantes :

X un document appartient à la classe `earn`

Y le mot `profit` apparaît dans un document

Ces variables peuvent prendre deux valeurs : vrai en cas de succès, lorsque la proposition est vérifiée pour un document et faux dans le cas contraire. Nous cherchons à

savoir si X et Y sont indépendantes. L'hypothèse nulle h_0 assume l'indépendance de X et Y. L'hypothèse alternative h_1 dit que l'appartenance à `earn` n'est pas indépendante de l'apparition de `profit` dans le document. Les données extraites du corpus Reuters-21758 pour X et Y sont les suivantes (en nombre de documents) :

	$X =$ vrai	$X =$ faux	Total
$Y =$ vrai	976	2 321	3 297
$Y =$ faux	1 733	3 776	5 509
Total	2 709	6 097	8 806

TABLE A.2 : Effectifs observés.

Ensuite, l'échantillon théorique est calculé à partir des données observées, en multipliant les effectifs totaux du tableau A.2 par ligne et par colonne divisés par le nombre total d'effectifs :

	$X =$ vrai	$X =$ faux	Total
$Y =$ vrai	$\frac{2\,709 \times 3\,297}{8\,806} = 1\,014{,}26$	$\frac{6\,097 \times 3\,297}{8\,806} = 2\,282{,}74$	3 297
$Y =$ faux	$\frac{2\,709 \times 5\,509}{8\,806} = 1\,694{,}74$	$\frac{6\,097 \times 5\,509}{8\,806} = 3\,814{,}26$	5 509
Total	2 709	6 097	8 806

TABLE A.3 : Effectifs théoriques.

Le degré de liberté est simplement $\nu = 1$. L'écart entre les deux ensembles est donné par les différences au carré entre les tableaux A.2 et A.3 :

$$\begin{aligned} d &= \frac{(976 - 1\,014,24)^2}{1\,014,24} + \frac{(1\,733 - 1\,694,74)^2}{1\,694,74} \\ &\quad + \frac{(2\,321 - 2\,282,74)^2}{2\,282,74} + \frac{(3\,776 - 3\,814,26)^2}{3\,814,26} \\ &= 3,330551062 \end{aligned}$$

D'après le tableau A.1, pour $\nu = 1$ et $r = 0.05$, le seuil critique du χ^2 est de 3,841. Il est alors impossible de rejeter l'hypothèse nulle. La dépendance de X et Y ne peut pas être prouvée.

Indépendamment du seuil critique, la valeur du χ^2 est utilisée comme mesure de pertinence d'un mot vis-à-vis d'une catégorie. Ainsi, pour chaque catégorie de notre corpus, il est possible d'obtenir un classement des mots en fonction de leur dépendance estimée à la catégorie. Le but de la sélection de termes est d'obtenir un espace de représentation intercatégories. Or, les classements obtenus correspondent à une seule catégorie. C'est là qu'intervient l'algorithme de Forman (2004).

A.2 L'algorithme de Forman

Le but l'algorithme de Forman (2004) est de choisir les m meilleurs termes à partir des n meilleurs termes pour chacune des catégories. L'idée sous-jacente est celle du tourniquet, chaque catégorie propose un terme à son tour, si le terme existe déjà dans la liste en cours de construction alors la catégorie propose le terme suivant, et ainsi de suite jusqu'à trouver un terme qui ne soit pas dans la liste.

L'algorithme se base sur un ensemble de termes classés par ordre de pertinence décroissante, ici par ordre décroissant de la valeur du χ^2, pour chacune des catégories $i \in \mathcal{C}$ avec $|\mathcal{C}|$ le nombre de catégories :

Algorithme 1 Algorithme de Forman

Entrées: $\mathcal{C}$, listes
Sorties: l'ensemble des m meilleurs termes

```
j ← 1
i ← 1
tantque j ≤ |C| faire
    t ← terme suivant de listes_i
    si t ∈ termes alors
        termes[j] ← t
        j ← j + 1
        i ← (i + 1) mod |C|
    finsi
fin tantque
Retourner termes.
```

La taille de l'espace de représentation, m, a été définie pendant l'étape de validation pour chacun des corpus utilisés dans nos expériences. Le détail des paramètres expérimentaux utilisés est donné dans les chapitres 4 et 5.

Publications

Revues internationales

Peña Saldarriaga, S., C. Viard-Gaudin et E. Morin. 2010, « Impact of on-line handwriting recognition performance on text categorization », *International Journal on Document Analysis & Recognition*, vol. À paraître, p. yy–zz.

Conférences internationales

Peña Saldarriaga, S., E. Morin et C. Viard-Gaudin. 2008, « Categorization of on-line handwritten documents », dans *DAS 2008, proceedings of the 8th International Workshop on Document Analysis Systems*, p. 95–102.

Peña Saldarriaga, S., E. Morin et C. Viard-Gaudin. 2009a, « Using top n recognition candidates to categorize on-line handwritten documents », dans *ICDAR 2009, proceedings of the 10th International Conference on Document Analysis & Recognition*, p. 881–885.

Peña Saldarriaga, S., E. Morin et C. Viard-Gaudin. 2010a, « Ranking fusion methods applied to on-line handwriting information retrieval », dans *ECIR 2010, proceedings of the 32nd Annual European Conference on Information Retrieval*, *Lecture Notes in Computer Science*, vol. 5993, édité par C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger et K. van Rijsbergen, p. 253–264.

Peña Saldarriaga, S., C. Viard-Gaudin et E. Morin. 2009b, « On-line handwritten text categorization », dans *DRR2009, Document Recognition & Retrieval XVI, proceedings of the SPIE-IS&T Electronic Imaging*, vol. 7247, Best Student Paper Award, p. 724709.

Peña Saldarriaga, S., C. Viard-Gaudin et E. Morin. 2010b, « Combining approaches to on-line handwriting information retrieval », dans *DRR 2010, Document Recognition*

& Retrieval XVII, proceedings of the SPIE-IS&T Electronic Imaging, vol. 7534, p. 753403.

Conférences nationales

Peña Saldarriaga, S., E. Morin et C. Viard-Gaudin. 2009a, « Impact de la reconnaissance de l'écriture en-ligne sur une tâche de catégorisation », dans *CORIA 2009, Actes de la 6ème Conférence en Recherche d'Information et Applications*, p. 219–234.

Peña Saldarriaga, S., E. Morin et C. Viard-Gaudin. 2009b, « Un nouveau schéma de pondération pour la catégorisation de documents manuscrits », dans *TALN 2009, Actes de la Conférence sur le Traitement Automatique des Langues Naturelles.*

Peña Saldarriaga, S., E. Morin et C. Viard-Gaudin. 2010, « Fusion de résultats en recherche d'information : application aux documents manuscrits en-ligne », dans *CIFED 2010, Actes du Colloque Francophone sur l'Écrit et le Document*, p. 3–18.

Bibliographie

Agarwal, S., S. Godbole, D. Punjani et S. Roy. 2007, « How much noise is too much : a study in automatic text classification », dans *ICDM 2007, proceedings of the 7th IEEE International Conference on Data Mining*, p. 3–12.

Aho, A. V. et M. J. Corasick. 1975, « Efficient string matching : an aid to bibliographic search », *Communications of the ACM*, vol. 18, n° 6, p. 333–340.

Alonso, O., D. E. Rose et B. Stewart. 2008, « Crowdsourcing for relevance evaluation », *ACM SIGIR Forum*, vol. 42, n° 2, p. 9–15.

Apté, C., F. Damerau et S. M. Weiss. 1994, « Towards language independent automated learning of text categorization models », dans *SIGIR 1994, proceedings of the 17th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 23–30.

Aref, W. G., D. Barabará et D. Lopresti. 1995a, *Multimedia Database Systems : Issues and Research Directions*, chap. Ink as a first-class datatype in multimedia databases, Artificial Intelligence, Springer-Verlag, Berlin, p. 113–163.

Aref, W. G., D. Barabará et P. Vallabhaneni. 1995b, « The handwritten trie : indexing electronic ink », dans *SIGMOD 1995, proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, p. 151–162.

Aref, W. G., I. Kamel et D. Lopresti. 1995c, « On handling electronic ink », *ACM Computing Surveys*, vol. 27, n° 4, p. 564–567.

Aslam, J. A. et M. Montague. 2001, « Models for metasearch », dans *SIGIR 2001, proceedings of the 24th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 276–284.

Baeza-Yates, R. et B. Ribeiro-Neto. 1999, *Modern Information Retrieval*, Addison Wesley-ACM Press, New York.

Beitzel, S. M., E. C. Jensen, A. Chowdury, D. Grossman, O. Frieder et N. Goharian. 2004, « On fusion of effective retrieval strategies in the same information retrieval system », *Journal of the American Society of Information Science & Technology*, vol. 50, n° 10, p. 859–868.

Belaïd, A. et H. Cecotti. 2006, *Les documents écrits – de la numérisation à l'indexation par le contenu*, chap. Reconnaissance de caractères : évaluation des performances, Hermès Science Publications-Lavoisier, p. 311–361.

Belkin, M., I. Matveeva et P. Niyogi. 2004, « Regularization and semi-supervised learning on large graphs », dans *COLT 2004, proceedings of the 17th Annual Conference on Learning Theory, Lecture Notes in Computer Science*, vol. 3120, p. 624–638.

Belkin, M. et P. Niyogi. 2002, « Laplacian eigenmaps and spectral techniques for embedding and clustering », dans *Advances in Neural Information Processing Systems*, vol. 14, édité par T. G. Dietterich, S. Becker et Z. Ghahramani, MIT Press, p. 585–591.

Belkin, N. J., C. Cool, W. B. Croft et J. P. Callan. 1993, « The effect of multiple query representations on information retrieval system performance », dans *SIGIR 1993, proceedings of the 16th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 339–346.

Beney, J. 2008, *Classification supervisée de documents*, Hermès Science Publications-Lavoisier, Paris.

Borda, J. C. 1781, *Mémoire sur les élections au scrutin*, Histoire de l'Académie Royale des Sciences, Paris.

Boyer, R. S. et J. S. Moore. 1977, « A fast string searching algorithm », *Communications of the ACM*, vol. 20, n° 10, p. 762–772.

Brin, S. et L. Page. 1998, « The anatomy of a large-scale hypertextual web search engine », dans *WWW 1998, proceedings of the 7th International Conference on World Wide Web*, p. 107–117.

Buckley, C. et E. M. Voorhees. 2000, « Evaluating evaluation measure stability », dans *SIGIR 2000, proceedings of the 23rd Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 33–40.

Burges, C. J. C. 1998, « A tutorial on support vector machines for pattern recognition », *Data Mining and Knowledge Discovery*, vol. 2, n° 2, p. 121–167.

Caillault, E., C. Viard-Gaudin et A. R. Ahmad. 2005, « MS-TDNN with global discriminant trainings », dans *ICDAR 2005, proceedings of 8th International Conference on Document Analysis & Recognition*, p. 856–860.

Cao, H. et V. Govindaraju. 2007, « Vector model based indexing and retrieval of handwritten medical forms », dans *ICDAR 2007, proceedings of 9th International Conference on Document Analysis & Recognition*, p. 88–92.

Cavnar, W. B. et J. M. Trenkle. 1994, « N-gram-based text categorization », dans *SDAIR 1994, proceedings of the 3rd Annual Symposium on Document Analysis & Information Retrieval*, p. 161–175.

Chapelle, O., J. Weston et B. Schölkopf. 2003, « Cluster kernels for semi-supervised learning », dans *Advances in Neural Information Processing Systems*, vol. 14, p. 585–592.

Cheng, C., B. Zhu, X. Chen et M. Nakagawa. 2009, « Improvements in keyword search japanese characters within handwritten digital ink », dans *ICDAR 2009, proceedings of 10th International Conference on Document Analysis & Recognition*, p. 863–866.

Chung, F. R. K. 1997, *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics, American Mathematical Society, Providence, Rhode Island.

Church, K. W. et P. Hanks. 1989, « Word association norms, mutual information, and lexicography », dans *ACL 1989, proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, p. 76–83.

Condorcet, M. J. A. N. 1785, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Imprimerie Royale, Paris.

Cortes, C. et V. Vapnik. 1995, « Support-vector networks », *Machine Learning*, vol. 20, n° 3, p. 273–297.

Craven, M., D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam et S. Slattery. 1998, « Learning to extract symbolic knowledge from the world wide web », dans *AAAI 1998, proceedings of the 15th National Conference on Artificial Intelligence*, p. 509–516.

Croft, W. B., S. M. Harding, K. Taghva et J. Borsack. 1994, « An evaluation of information retrieval accuracy with simulated OCR output », dans *SDAIR 1994, proceedings of the 3rd Annual Symposium on Document Analysis & Information Retrieval*, p. 115–126.

Cvetkovic, D. M., M. Doob et H. Sachs. 1998, *Spectra of Graphs : Theory and Applications*, 3e éd., Wiley, New York.

Damerau, F. J. 1964, « A technique for computer detection and correction of spelling errors », *Communications of the ACM*, vol. 7, n° 3, p. 171–176.

Davis, M. R. et T. O. Ellis. 1964, « The RAND tablet : a man-machine graphical communication device », dans *Proceedings of the 26th AFIPS Fall Joint Computer Conference*, p. 325–331.

Debole, F. et F. Sebastiani. 2005, « An analysis of the relative hardness of Reuters-21578 subsets », *Journal of the American Society for Information Science and Technology*, vol. 56, n° 6, p. 584–596.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer et R. Harshman. 1990, « Indexing by latent semantic analysis », *Journal of the American Society for Information Science*, vol. 41, n° 6, p. 391–407.

Díaz, F. 2007, « Regularizing query-based retrieval scores », *Information Retrieval*, vol. 10, n° 6, p. 531–562.

Díaz, F. 2008, *Autocorrelation and Regularization of Query-based Information Retrieval Scores*, thèse de doctorat, University of Massachussets, Amherst.

Dubois, J., M. Giacomo, L. Gespin, C. Marcellesi, J.-B. Marcellesi et J.-P. Mével. 1999, *Dictionnaire de linguistique et des sciences du langage*, Larousse, Paris.

Dumais, S. T., G. W. Furnas, T. K. Landauer, S. Deerwester et R. Harshman. 1988, « Using latent semantic analysis to improve access to textual information », dans *CHI 1988, proceedings of the 5th Conference on Human Factors in Computing Systems*, p. 281–285.

Dwork, C., R. Kumar, M. Naor et D. Sivakumar. 2001, « Rank aggregation methods for the web », dans *WWW 2001, proceedings of the 10th International Conference on World Wide Web*, p. 613–622.

van Erp, M. et L. Schomaker. 2000, « Variants of the Borda count method for combining ranked classifier hypotheses », dans *IWFHR 2000, proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition*, p. 443–452.

Farah, M. et D. Vanderpooten. 2007, « An outranking approach for rank aggregation in information retrieval », dans *SIGIR 2007, proceedings of the 30th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 591–598.

Forman, G. 2003, « An extensive empirical study of feature selection metrics for text classification », *Journal of Machine Learning Research*, vol. 3, p. 1289–1305.

Forman, G. 2004, « A pitfall and solution in multi-class feature selection for text classification », dans *ICML 2004, proceedings of the 21st International Conference on Machine Learning*, p. 38–46.

Galavotti, L., F. Sebastiani et M. Simi. 2000, « Experiments on the use of feature selection and negative evidence in automated text categorization », dans *ECDL 2000, proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, *Lecture Notes in Computer Science*, vol. 1923, p. 59–68.

Garfield, E. 1997, « A tribute to Calvin N. Mooers, a pioneer of information retrieval », *The Scientist*, vol. 11, n° 6, p. 9.

Guyon, I., L. Schomaker, R. Plamondon, M. Liberman et S. Janet. 1994, « UNIPEN project of on-line data exchange and recognizer benchmarks », dans *ICPR 1994, proceedings of the 12th International Conference on Pattern Recognition*, vol. 2, p. 29–33.

Hastie, T., R. Tibshirani et J. Friedman. 2008, *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*, 2e éd., Springer.

Hayes, P. J. et S. P. Weinstein. 1990, « Construe-TIS : a system for content-based indexing of a database of news stories », dans *IAAI 1990, proceedings of the 2nd Annual Conference on Innovative Applications of Artificial Intelligence*, p. 49–64.

Hersh, W., C. Buckley, T. J. Leone et D. Hickam. 1994, « OHSUMED : an interactive retrieval evaluation and new large test collection for research », dans *SIGIR 1994, proceedings of the 17th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 192–201.

Hirschberg, D. S. 1977, « A linear space algorithm for computing maximal common subsequences », *Communications of the ACM*, vol. 18, no 6, p. 341–343.

Ide, E. 1971, *The SMART Retrieval System – Experiments in Automatic Document Processing*, chap. New Experiments in Relevance Feedback, Prentice-Hall, Inc., p. 337–354.

Ittner, D. J., D. D. Lewis et D. D. Ahn. 1995, « Text categorization of low quality images », dans *SDAIR 1995, proceedings of the 4th Annual Symposium on Document Analysis & Information Retrieval*, p. 301–315.

Jaeger, S., S. Manke, J. Reichert et A. Waibel. 2001, « On-line handwriting recognition : the NPen++ recognizer », *International Journal on Document Analysis & Recognition*, vol. 3, no 3, p. 169–180.

Jain, A. K. et A. M. Namboodiri. 2003, « Indexing and retrieval of on-line handwritten documents », dans *ICDAR 2003, proceedings of the 7th International Conference on Document Analysis & Recognition*, p. 655–659.

Jardine, N. et C. J. van Rijsbergen. 1971, « The use of hierarchic clustering in information retrieval », *Information Storage & Retrieval*, vol. 7, no 5, p. 217–240.

Jawahar, C. V., A. Balasubramanian, M. Meshesha et A. M. Namboodiri. 2009, « Retrieval of online handwriting by synthesis and matching », *Pattern Recognition*, vol. 42, no 7, p. 1445–1457.

Joachims, T. 2002, *Learning to Classify Text Using Support Vector Machines : Methods, Theory and Algorithms*, Kluwer Academic Publishers, Norwell, MA.

Junker, M. et R. Hoch. 1998, « An experimental evaluation of OCR text representations for learning document classifiers », *International Journal on Document Analysis & Recognition*, vol. 1, no 2, p. 116–122.

Kantor, P. B. et E. M. Voorhees. 2000, « The TREC-5 confusion track : comparing retrieval methods for scanned text », *Information Retrieval*, vol. 2, no 2-3, p. 165–176.

Kleinberg, J. M. 1999, « Authoritative sources in a hyperlinked environment », *Journal of the ACM*, vol. 46, n° 6, p. 604–632.

Knoblock, C., D. Lopresti, S. Roy et L. V. Subramaniam, éd.. 2007, *AND 2007, proceedings of the 1st Workshop on Analytics for Noisy Unstructured Text Data*.

Koch, G. 2006, *Catégorisation Automatique de Documents Manuscrits : Application aux Courriers Entrants*, thèse de doctorat, Université de Rouen.

Kondor, R. I. et J. Lafferty. 2002, « Diffusion kernels on graphs and other discrete structures », dans *ICML '02, proceedings of the 19th International Conference on Machine Learning*, p. 315–322.

Kullback, S. et R. A. Leibler. 1951, « On information and sufficiency », *Annals of Mathematical Statistics*, vol. 22, n° 1, p. 79–86.

Kwok, T., M. P. Perrone et G. Russell. 2000, « Ink retrieval from handwritten documents », dans *IDEAL 2000, proceedings of the 2nd International Conference on Intelligent Data Engineering and Automated Learning.*, *Lecture Notes in Computer Science*, vol. 1983, p. 461–466.

Lafferty, J. et C. Zhai. 2001, « Document language models, query models, and risk minimization for information retrieval », dans *SIGIR 2001, proceedings of the 24th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 111–119.

Lang, K. 1995, « NewsWeeder : learning to filter netnews », dans *ICML 1995, proceedings of the 12th International Conference on Machine Learning*, p. 331–339.

Lee, J. H. 1997, « Analysis of multiple evidence combination », dans *SIGIR 1997, proceedings of the 20th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 267–276.

Levenshtein, V. I. 1966, « Binary codes capable of correcting deletions, insertions, and reversals », *Soviet Physics Doklady*, vol. 10, n° 8, p. 707–710. Traduit du russe. 1965, titre original : « Dvoichnyye kody s ispravleniyem vypadeniy, vstavok i zameshcheniy simvolov ». *Doklady Akademii Nauk SSSR*, vol. 163, n° 4, p. 845-848, 1965.

Lewis, D. D. et M. Ringuette. 1994, « A comparison of two learning algorithms for text categorization », dans *SDAIR 1994, proceedings of the 3rd Annual Symposium on Document Analysis & Information Retrieval*, p. 81–93.

Lewis, D. D., Y. Yang, T. G. Rose et F. Li. 2004, « RCV1 : a new benchmark collection for text categorization research », *Journal of Machine Learning Research*, vol. 5, p. 361–397.

Liwicki, M. et H. Bunke. 2005, « IAM-OnDB – an on-line english sentence database acquired from handwritten text on a whiteboard », dans *ICDAR 2005, proceedings of the 8th International Conference on Document Analysis & Recognition*, p. 956–961.

Lopresti, D., S. Roy, K. Schulz et L. V. Subramaniam, éd.. 2008, *AND 2008, proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data*.

Lopresti, D., S. Roy, K. Schulz et L. V. Subramaniam, éd.. 2009, *AND 2009, proceedings of the 3rd Workshop on Analytics for Noisy Unstructured Text Data*.

Lopresti, D. et A. Tomkins. 1994, « On the searchability of electronic ink », dans *IWFHR 1994, proceedings of the 4th International Workshop on Frontiers of Handwriting Recognition*, p. 156–165.

Lopresti, D. et J. Zhou. 1996, « Retrieval strategies for noisy text », dans *SDAIR 1996, proceedings of the 5th Annual Symposium on Document Analysis & Information Retrieval*, p. 255–259.

Lopresti, D. a. 1996, « Robust retrieval of noisy text », dans *ADL 1996, proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries*, p. 76–85.

Lorette, G. et T. Paquet. 2006, *Les documents écrits – de la numérisation à l'indexation par le contenu*, chap. Reconnaissance de l'écriture manuscrite, Hermès Science Publications-Lavoisier, p. 37–86.

Luhn, H. P. 1958, « Auto-encoding of documents for information retrieval systems », dans *Modern Trends in Documentation*, p. 45–58.

von Luxburg, U. 2007, « A tutorial on spectral clustering », *Statistics and Computing*, vol. 17, n° 4, p. 395–416.

von Luxburg, U., O. Bousquet et M. Belkin. 2005, « Limits of spectral clustering », dans *Advances in Neural Information Processing Systems*, vol. 17, p. 857–864.

van der Maaten, L. et G. Hinton. 2008, « Visualizing data using t-SNE », *Journal of Machine Learning Research*, vol. 9, p. 2579–2605.

Madhvanath, S. et V. Govindaraju. 2001, « The role of holistic paradigms in handwritten word recognition », *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 23, p. 149–164.

Manmatha, R., H. Chengfeng et E. M. Riseman. 1996a, « Word spotting : a new approach to indexing handwriting », dans *CVPR 1996, proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, p. 631–637.

Manmatha, R., H. Chengfeng, E. M. Riseman et W. B. Croft. 1996b, « Indexing handwriting using word matching », dans *ICDL 1996, proceedings of the 1st International Conference on Digital Libraries*, p. 151–159.

Manmatha, R., T. M. Rath et F. Feng. 2001, « Modeling score distributions for combining the outputs of search engines », dans *SIGIR 2001, proceedings of the 24th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 267–275.

Manning, C. D., P. Raghavan et H. Schütze. 2008, *Introduction to Information Retrieval*, Cambridge University Press.

Maron, M. E. 1961, « Automatic indexing : an experimental inquiry », *Journal of the ACM*, vol. 8, n° 3, p. 404–417.

Milewski, R. J., V. Govindaraju et A. Bhardwaj. 2009, « Automatic recognition of handwritten medical forms for search engines », *International Journal on Document Analysis & Recognition*, vol. 11, n° 4, p. 203–218.

Mitchell, T. M. 1997, *Machine Learning*, McGraw-Hill.

Mittendorf, E. 1998, *Data Corruption and Information Retrieval*, thèse de doctorat, ETH Zürich.

Mittendorf, E. et P. Schäuble. 1996, « Measuring the effects of data corruption on information retrieval », dans *SDAIR 1996, proceedings of the 5th Annual Symposium on Document Analysis & Information Retrieval*, p. 179–189.

Mittendorf, E. et P. Schäuble. 2000, « Information retrieval can cope with many errors », *Information Retrieval*, vol. 3, n° 3, p. 189–216.

Mohar, B. 1988, « The laplacian spectrum of graphs », dans *proceedings of the 6th International Conference on the Theory and Applications of Graphs*, p. 871–898.

Montague, M. et J. A. Aslam. 2001, « Relevance score normalization for metasearch », dans *CIKM 2001, proceedings of the 10th International Conference on Information & Knowledge Management*, p. 427–433.

Montague, M. et J. A. Aslam. 2002, « Condorcet fusion for improved retrieval », dans *CIKM 2002, proceedings of the 11th International Conference on Information & Knowledge Management*, p. 538–548.

Murata, M., L. S. P. Busagala, W. Ohyama, T. Wakabayashi et F. Kimura. 2006, « The impact of OCR accuracy and feature transformation on automatic text classification », dans *DAS 2006, proceedings of the 7th IAPR International Workshop on Document Analysis Systems*, p. 506–517.

Myers, J. L. et A. D. Well. 2003, *Research Design and Statistical Analysis*, 2e éd., Lawrence Erlbaum Associates.

Myers, K., M. Kearns, S. Singh et M. A. Walker. 2000, « A boosting approach to topic spotting on subdialogues », dans *ICML 2000, proceedings of the 17th International Conference on Machine Learning*, p. 655–662.

Namboodiri, A. M. 2004, *On-line handwritten document understanding*, thèse de doctorat, Michigan State University, East Lansing, MI.

Namer, F. 2000, « FLEMM : un analyseur flexionnel du français à base de règles », *Traitement Automatique des Langues*, vol. 41, n° 2, p. 523–547.

Ng, A. Y., M. I. Jordan et Y. Weiss. 2002, « On spectral clustering : analysis and an algorithm », dans *Advances in Neural Information Processing Systems*, vol. 14, p. 849–856.

Ng, A. Y., A. X. Zheng et M. I. Jordan. 2001, « Stable algorithms for link analysis », dans *SIGIR 2001, proceedings of the 24th International ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 258–266.

Perraud, F. 2005, *Modélisation du Langage Naturel Appliquée à la Reconnaissance de l'Écriture Manuscrite En-Ligne*, thèse de doctorat, Université de Nantes.

Perraud, F., C. Viard-Gaudin, E. Morin et P.-M. Lallican. 2003, « N-gram and n-class models for on line handwriting recognition », dans *ICDAR 2003, proceedings of the 7th International Conference on Document Analysis & Recognition*, p. 1053–1059.

Perronnin, F. et J. A. Rodriguez-Serrano. 2009, « Fisher kernels for handwritten word-spotting », dans *ICDAR 2009, proceedings of 10th International Conference on Document Analysis & Recognition*, p. 106–110.

Pfeifer, U., N. Fuhr et T. Huynh. 1995, « Searching structured documents with the enhanced retrieval functionality of freeWAIS-sf and SFgate », dans *WWW 1995, proceedings of the 3rd International World Wide Web Conference*, p. 1027–1036.

Pittman, J. A. 2007, « Handwriting recognition : tablet PC text input », *Computer*, vol. 40, n° 9, p. 49–54.

Ponte, J. M. et W. B. Croft. 1998, « A language modeling approach to information retrieval. », dans *SIGIR 1998, proceedings of the 21st Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 275–281.

Porter, M. F. 1980, « An algorithm for suffix stripping », *Program*, vol. 14, n° 3, p. 130–137.

Prochasson, E., C. Viard-Gaudin et E. Morin. 2007, « Language models for handwritten short message services », dans *ICDAR 2007, proceedings of the 9th International Conference on Document Analysis & Recognition*, p. 83–87.

Quiniou, S. et E. Anquetil. 2006, « A priori and a posteriori integration and combination of language models in an on-line handwritten sentence recognition system », dans *IWFHR 2006, proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition*, p. 403–408.

Quiniou, S., E. Anquetil et S. Carbonnel. 2005, « Statistical language models for on-line handwritten sentence recognition », dans *ICDAR 2005, proceedings of the 8th International Conference on Document Analysis & Recognition*, p. 516–520.

Rath, T. M. et R. Manmatha. 2003a, « Features for word spotting in historical manuscripts », dans *ICDAR 2003, proceedings of the 7th International Conference on Document Analysis & Recognition*, p. 218–222.

Rath, T. M. et R. Manmatha. 2003b, « Word image matching using dynamic time warping », dans *CVPR 2003, proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, p. 521–527.

Rath, T. M., R. Manmatha et V. Lavrenko. 2004, « A search engine for historical manuscript images », dans *SIGIR 2004, proceedings of the 27th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 369–376.

Ratzlaff, E. H. 2000, « Inter-line distance estimation and text line extraction for unconstrained on-line handwriting », dans *IFWHR 2000, proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition*, p. 33–42.

Renda, M. E. et U. Straccia. 2003, « Web metasearch : rank vs. score based rank aggregation methods », dans *SAC 2003, proceedings of the 18th Annual ACM Symposium on Applied Computing*, p. 841–846.

van Rijsbergen, C. J. 1979, *Information Retrieval*, 2e éd., Butterworths, London.

Robertson, S. E. et K. Spärck Jones. 1976, « Relevance weighting of search terms », *Journal of the American Society for Information Science*, vol. 27, n° 3, p. 129–146.

Rocchio, J. J. 1971, *The SMART Retrieval System – Experiments in Automatic Document Processing*, chap. Relevance feedback in information retrieval, Prentice-Hall, Inc., p. 313–323.

Rossi, J.-P. 2005, *Psychologie de la mémoire. De la mémoire épisodique à la mémoire sémantique.*, 1re éd., Éditions De Boeck Université, Bruxelles.

Russell, G., M. P. Perrone et Y. M. Chee. 2002, « Handwritten document retrieval », dans *IWFHR 2002, proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition*, p. 233–238.

Sakoe, H. et S. Chiba. 1978, « Dynamic programming algorithm optimization for spoken word recognition », *IEEE Transactions on Acoustics, Speech & Signal Processing*, vol. 26, n° 1, p. 43–49.

Salton, G. 1968, *Automatic Information Organization and Retrieval*, McGraw Hill Text, New York.

Salton, G. et C. Buckley. 1990, « Improving retrieval performance by relevance feedback », *Journal of the American Society for Information Science*, vol. 41, n° 4, p. 288–297.

Salton, G. et M. E. Lesk. 1965, « The SMART automatic document retrieval systems – an illustration », *Communications of the ACM*, vol. 8, n° 6, p. 391–398.

Salton, G., A. Wong et C. S. Yang. 1975, « A vector space model for automatic indexing », *Communications of the ACM*, vol. 18, n° 11, p. 613–620.

Sanderson, M. 1994, « Word sense disambiguation and information retrieval », dans *SIGIR 1994, proceedings of the 17th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 142–151.

Savoy, J. 1999, « A stemming procedure and stopword list for general french corpora », *Journal of the American Society for Information Science*, vol. 50, n° 10, p. 944–952.

Savoy, J., A. Le Calvé et D. Vrajitoru. 1997, « Report on the TREC-5 experiment : data fusion and collection fusion », dans *TREC-5, proceedings of the 5nd Text REtrieval Conference*, p. 489–502.

Sayre, K. M. 1973, « Machine recognition of handwritten words : a project report », *Pattern Recognition*, vol. 5, n° 3, p. 213–228.

Schäuble, P. et U. Glavitsch. 1994, « Assessing the retrieval effectiveness of a speech retrieval system by simulating recognition errors », dans *HLT 1994, proceedings of the Workshop on Human Language Technology*, p. 370–372.

Schimke, S. et C. Vielhauer. 2006, « Document retrieval in pen-based media data », dans *Proceedings of the 2nd International Conference on Automated Production of Cross Media Content for Multi-channel Distribution*, p. 186–190.

Schmid, H. 1994, « Probabilistic part-of-speech tagging using decision trees », dans *NEMLP 1994, proceedings of the 1st International Conference on New Methods in Language Processing*, p. 44–49.

Schölkopf, B. et A. J. Smola. 2002, *Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA.

Sebastiani, F. 2002, « Machine learning in automated text categorization », *ACM Computing Surveys*, vol. 34, n° 1, p. 1–47.

Shaw, J. A. et E. A. Fox. 1994, « Combination of multiple searches », dans *TREC-2, proceedings of the 2nd Text REtrieval Conference*, p. 243–252.

Shi, J. et J. Malik. 2000, « Normalized cuts and image segmentation », *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 22, n° 8, p. 888–905.

Shridhar, M., G. F. Houle et K. Fumitaka. 1997, « Handwritten word recognition using lexicon free and lexicon directed word recognition algorithms », dans *ICDAR 1997, proceedings of the 4th International Conference Document Analysis & Recognition*, p. 861–865.

Spärck Jones, K. 1979, « Experiments in relevance weighting of search terms », *Information Processing & Management*, vol. 15, n° 3, p. 133–144.

Spärck Jones, K., S. Walker et S. E. Robertson. 2000a, « A probabilistic model of information retrieval : development and comparative experiments, part 1 », *Information Processing & Management*, vol. 36, n° 6, p. 779–808.

Spärck Jones, K., S. Walker et S. E. Robertson. 2000b, « A probabilistic model of information retrieval : development and comparative experiments, part 2 », *Information Processing & Management*, vol. 36, n° 6, p. 809–840.

Srihari, S. N., C. Huang et H. Srinivasan. 2005, « A search engine for handwritten documents », dans *DRR2005, Document Recognition & Retrieval XII, proceedings of the SPIE-IS&T Electronic Imaging*, vol. 5676, p. 66–75.

Stoer, J. et R. Bulirsch. 2002, *Introduction to Numerical Analysis*, 3e éd., Springer-Verlag, New York.

Subrahmonia, J., K. Nathan et M. P. Perrone. 1996, « Writer dependent recognition of on-line unconstrained handwriting », dans *ICASSP 1996, IEEE International Conference on Acoustics, Speech & Signal Processing*, p. 3478–3481.

Sutherland, I. E. 1963, *Sketchpad, a man-machine graphical communication system*, thèse de doctorat, Massachusetts Institute of Technology.

Taghva, K., J. Borsack, A. Condit et S. Erva. 1994, « The effects of noisy data on text retrieval », *Journal of the American Society for Information Science*, vol. 45, n° 1, p. 50–58.

Taghva, K., T. A. Nartker, J. Borsack, S. Lumos, A. Condit et R. Young. 2000, « Evaluating text categorization in the presence of OCR errors », dans *DRR 2000, proceedings of Document Recognition & Retrieval VIII*, vol. 4307, p. 68–74.

Tan, G. X., C. Viard-Gaudin et A. C. Kot. 2009a, « Impact of alphabet knowledge on online writer identification », dans *ICDAR 2009, proceedings of the 10th International Conference on Document Analysis & Recognition*, p. 56–60.

Tan, G. X., C. Viard-Gaudin et A. C. Kot. 2009b, « Online writer identification using alphabetic information clustering », dans *DRR2009, Document Recognition & Retrieval XVI, proceedings of the SPIE-IS&T Electronic Imaging*, vol. 7247, p. 72470F.

Terasawa, K. et Y. Tanaka. 2009, « Slit style HOG feature for document image word spotting », dans *ICDAR 2009, proceedings of 10th International Conference on Document Analysis & Recognition*, p. 116–120.

Tormene, P., T. Giorgino, S. Quaglini et M. Stefanelli. 2008, « Matching incomplete time series with dynamic time warping : an algorithm and an application to post-stroke rehabilitation », *Artificial Intelligence in Medicine*, vol. 45, n° 1, p. 11–34.

Vapnik, V. N. 2000, *The Nature of Statistical Learning Theory*, 2e éd., Springer-Verlag, New York.

Viard-Gaudin, C., P.-M. Lallican, S. Knerr et P. Binter. 1999, « The IRESTE on/off (IRONOFF) dual handwriting database », dans *ICDAR 1999, proceedings of the 5th International Conference on Document Analysis & Recognition*, p. 455–458.

Vinciarelli, A. 2004, « Effect of recognition errors on information retrieval performance », dans *IWFHR 2004, proceedings of 9th International Workshop on Frontiers in Handwriting Recognition*, p. 275–279.

Vinciarelli, A. 2005a, « Application of information retrieval techniques to single writer documents », *Pattern Recognition Letters*, vol. 26, n° 14, p. 2262–2271.

Vinciarelli, A. 2005b, « Noisy text categorization », *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, n° 12, p. 1882–1895.

Vinciarelli, A. 2006, « Indexation de documents manuscrits », dans *CIFED 2006, Actes du Colloque International Francophone sur l'Ecrit et le Document*, p. 49–54.

Vogt, C. C. et G. W. Cottrell. 1999, « Fusion via a linear combination of scores », *Information Retrieval*, vol. 1, n° 3, p. 151–173.

Voorhees, E. M. et D. K. Harman, éd.. 2005, *TREC : Experiment and Evaluation in Information Retrieval*, MIT Press, Cambridge, MA.

Waibel, A., H. Toshiyuki, G. Hinton, S. Kiyohiro et K. J. Lang. 1989, « Phoneme recognition using time-delay neural networks », *IEEE Transactions on Neural Networks*, vol. 37, n° 3, p. 328–339.

Wu, S., Y. Bi, X. Zeng et L. Han. 2009, « Assigning appropriate weights for the linear combination data fusion method in information retrieval », *Information Processing & Management*, vol. 45, n° 4, p. 413–426.

Wu, S. et S. McClean. 2005, « Data fusion with correlation weights », dans *ECIR 2005, proceedings of the 27th Annual European Conference on Information Retrieval, Lecture Notes in Computer Science*, vol. 3408, p. 275–286.

Xue, H. et V. Govindaraju. 2002, « On the dependence of handwritten word recognizers on lexicons », *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 24, n° 12, p. 1553–1564.

Yang, Y. 2001, « A study of thresholding strategies for text categorization », dans *SIGIR 2001, proceedings of the 24th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 137–145.

Yang, Y. et X. Liu. 1999, « A re-examination of text categorization methods », dans *SIGIR 1999, proceedings of the 22th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 42–49.

Yang, Y. et J. O. Pedersen. 1997, « A comparative study on feature selection in text categorization », dans *ICML 1997, proceedings of the 14th International Conference on Machine Learning*, p. 412–420.

Zhai, C. et J. Lafferty. 2001, « A study of smoothing methods for language models applied to ad hoc information retrieval », dans *SIGIR 2001, proceedings of the 24th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 334–342.

Zhou, D., O. Bousquet, T. Navin Lal, J. Weston et B. Schölkopf. 2004a, « Learning with local and global consistency », dans *Advances in Neural Processing Systems*, vol. 16, p. 321–328.

Zhou, D. et B. Schölkopf. 2004, « A regularization framework for learning from graph data », dans *SRL 2004, proceedings of the Workshop on Statistical Relational Learning & its Connections to Other Fields*, p. 132–137.

Zhou, D., J. Weston, A. Gretton, O. Bousquet et B. Schölkopf. 2004b, « Ranking on data manifolds », dans *Advances in Neural Information Processing Systems*, vol. 16, p. 169–176.

Approches textuelles pour la catégorisation et la recherche de documents manuscrits en-ligne

L'évolution technologique des dispositifs électroniques de capture de l'écriture manuscrite se traduit par l'apparition d'une grande quantité de documents manuscrits en-ligne. Cela pose la question de l'accès à l'information contenue dans ces données. Ce travail s'intéresse à l'accès à l'information textuelle contenue dans des documents qui se présentent sous la forme d'une séquence temporelle de points (x, y). Deux tâches principales ont été étudiées : la première concerne le développement d'un système de catégorisation de documents, tandis que la seconde s'intéresse à la recherche d'information dans des bases documentaires manuscrites. En amont, une première étape importante a consisté à collecter un corpus manuscrit de référence pour la validation expérimentale de cette étude. L'utilisation d'un système de reconnaissance de l'écriture étant l'élément transversal des approches proposées, une partie de notre travail a consisté à analyser le comportement de ces approches face aux erreurs de reconnaissance. La catégorisation est effectuée en enchaînant un système de reconnaissance à un système de catégorisation basé sur des méthodes d'apprentissage statistique. Pour la recherche d'information, deux approches ont été proposées. La première tire parti de la diversité des résultats restitués par des algorithmes de recherche différents, l'idée étant que la combinaison des résultats peut pallier leurs faiblesses respectives. La seconde approche exploite les relations de proximité thématique entre les documents. Si deux documents proches ont tendance à répondre au même besoin d'information, alors ces mêmes documents doivent avoir des scores de pertinence proches.

Mots-clés : documents manuscrits en ligne, reconnaissance de l'écriture, catégorisation, recherche d'information, fusion de résultats, régularisation

Text-based approaches to on-line handwritten document categorization and retrieval

With recent technical evolutions, pen-based input devices have become very popular. As a result, large amounts of on-line handwritten data are being created. Consequently, algorithms for efficient storage and retrieval of on-line data, represented as a temporal sequence of (x, y) coordinates, are being increasingly demanded. This thesis addresses the problem of accessing textual information in on-line handwritten documents. The overall goal of this work is the design of a system for text categorization and retrieval. In order to validate the methods proposed in this study, we collected a benchmark collection of handwritten documents. The use of an on-line handwriting recognition engine, as the common component of our approaches, leads us to focus part of our work on the impact of handwriting recognition errors. We address the problem of document categorization by pipelining the output of a handwriting recognition system into the input of a text categorization engine based on machine learning algorithms. We also develop two retrieval algorithms. First, we propose combining different approaches for retrieving handwritten documents. Our hypothesis is that different retrieval algorithms should retrieve different sets of documents for the same query. Therefore, improvements in retrieval performances can be expected. The second proposed algorithm is based on the topical relationships between documents. If closely associated documents tend to be relevant to the same requests, then topically-related documents should be assigned close retrieval scores.

Keywords : on-line handwriting, handwriting recognition, text categorization, information retrieval, data fusion, regularization