Un systeme generique d'extraction d'information dans des documents manuscrits non-contraints

S. Thomas, T. Paquet, L. Heutte, C. Chatelain

Un système générique d'extraction d'information dans des documents manuscrits non contraints

Simon Thomas, Thierry Paquet, Laurent Heutte, Clément Chatelain

Université de Rouen, LITIS - EA4108
Avenue de l'université
76800 Saint-Etienne du Rouvray
simon.thomas@etu.univ-rouen.fr

RÉSUMÉ. *Dans cet article, un système d'extraction d'information par analyseur statistique de surface dans des documents manuscrits faiblement contraints est introduit. Contrairement aux principales approches de la littérature que sont le keyword spotting et la reconnaissance complète de documents, l'originalité du système mis en oeuvre réside dans l'attention portée à la modélisation globale de l'écriture. En effet, la ligne de texte est considérée comme une entité indivisible et est modélisée de manière duale à l'aide de modèles de Markov cachés. Ainsi, une analyse surfacique de l'écriture permet d'isoler rapidement l'information pertinente recherchée dans un texte quelconque et ce, en une seule passe. Les premiers résultats sont encourageants et illustrent le potentiel de l'approche en terme d'extraction d'information.*

ABSTRACT. *In this paper, a novel information extraction system by statistical shallow parsing in unconstrained handwritten documents is introduced. Contrary to common approaches found in the literature as keyword spotting or full document recognition, our approch relies on a stronger and powerful global writing modelization. An entire text line is considered as an indivisible entity and is modeled in a dual way thanks to Hidden Markov Models. In this way, text line shallow parsing allows fast extraction of the relevant information in any documents and in only one pass. First results are promising and show the interest of the approch for information extraction.*

MOTS-CLÉS : *Reconnaissance de l'écriture, Extraction d'information, Documents manuscrits, Modèles de Markov Cachés, Analyse de surface*

KEYWORDS: *Handwriting recognition, Information extraction, Unconstrained handwritten documents, Hidden Markov Models, Shallow parsing*

1. Introduction

Malgré l'évolution des technologies numériques, la quantité de documents échangés dans les transactions administratives et économiques ne diminuent pas. De nombreuses applications très spécifiques et déjà industrialisées concernent souvent des documents très contraints dans leur mise en forme et leur syntaxe. Nous pouvons citer la lecture automatique de chèques bancaires (Heutte *et al.,* 1997), d'adresses postales sur des enveloppes (El-Yacoubi *et al.,* 2002) ou encore de formulaires (Niyogi *et al.,* 1996). Mises à part ces applications bien définies, le problème de la reconnaissance de l'écriture manuscrite reste délicat en l'absence de connaissances sur les documents traités. Des solutions de reconnaissance complète de documents faiblement contraints comme les textes libres ont été mis au point (Vinciarelli *et al.,* 2004, Marti *et al.,* 2000) mais, à cause de la complexité de la tâche (lexique ouvert, peu de connaissances a priori ...), les résultats sont insuffisants pour permettre leur industrialisation.

Depuis le début des années 2000, de nouvelles approches correspondant à des besoins industriels mieux définis ont été imaginées. Ainsi, un intérêt croissant est actuellement porté sur le traitement automatique des courriers entrants (cf. figure 1) correspondant aux courriers manuscrits arrivant en entrée de grands groupes et administrations. La base de données Rimes (Grosicki *et al.,* 2009) illustre bien cette ambition. Plutôt que de chercher à reconnaître ces documents dans leur intégralité avec des méthodes de reconnaissance de texte (Nosary *et al.,* 2004, Vinciarelli *et al.,* 2004), les informations pertinentes du courrier comme son objet, le nom de l'expéditeur, la date de l'envoi, éventuellement des numéros de client et de commande peuvent être recherchés. Un système de *keyword spotting*, c'est à dire de détection de mots clés par requête, peut être mis en oeuvre à cette fin. Un système d'*extraction d'information* peut également être envisagé. Nos travaux sur ces courriers entrants s'incrivent dans ce cadre. Ce domaine est pour l'heure assez pauvre en termes de travaux, nous pouvons toutefois citer les travaux (Chatelain, 2006, Koch, 2006) dont le but est d'extraire (localiser et reconnaître) dans des courriers entrants respectivement des séquences numériques et des mots clés. Dans cet article, un système d'extraction d'information innovant est introduit. Celui-ci est caractérisé par un modèle de ligne de texte modélisant l'information pertinente recherchée sous la forme de modèles de mots clés et l'information non pertinente sous la forme d'un modèle d'analyse de surface des mots non pertinents dans les phrases.

Dans la première partie, nous présentons les différentes approches de la littérature permettant d'extraire de l'information depuis des documents manuscrits faiblement contraints. Puis, nous introduisons le modèle générique d'extraction de mots dans des lignes de textes en confrontant nos choix à ces approches. Dans une seconde partie, la chaîne de traitements mise en oeuvre et les algorithmes utilisés sont décrits. Enfin, nous analysons les premiers résultats et concluons en donnant les pistes d'évolution du système et l'orientation de nos travaux à venir.

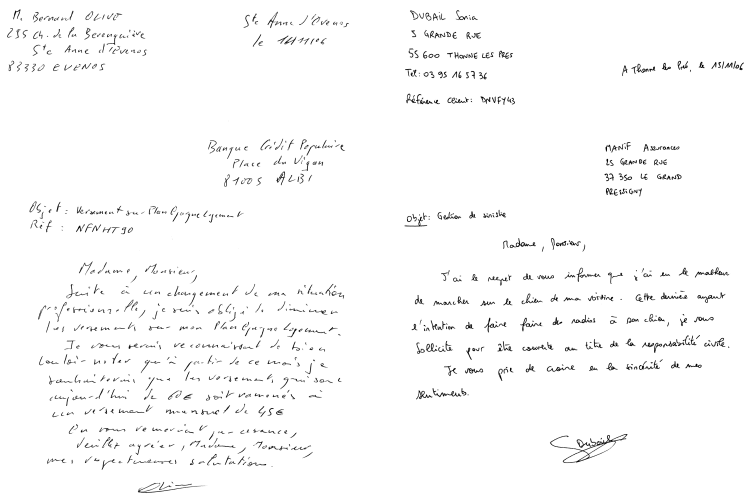

Figure 1. *Exemples de courriers entrants manuscrits extraits de la base de courrier Rimes.*

2. Modélisation

2.1. *Etat de l'art*

Le traitement automatique d'un document manuscrit nécessite l'extraction du contenu de celui-ci. Ce contenu peut être scindé en deux parties : l'une est pertinente et peut être représentée par un lexique, l'autre, qui ne nous intéresse pas, peut être vue comme non pertinente. Isoler l'information pertinente dans un document manuscrit peut se faire de différentes manières, la modélisation étant un point clé de cette problématique. Dans la littérature, nous pouvons regrouper les principales approches d'extraction d'information pertinente en 3 catégories présentées ci-dessous :

– la détection de mots clés (*keyword spotting*) (Cao *et al.*, 2007, Choisy, 2007, Rodríguez-Serrano *et al.*, 2009a)

– la reconnaissance intégrale d'un texte manuscrit (Favata *et al.*, 1998, Marti *et al.*, 2001, Vinciarelli *et al.*, 2004)

– l'extraction d'information dans un texte (Chatelain, 2006, Koch, 2006)

La première stratégie consiste simplement à reconnaître ou rejeter des mots en fonction de ce qui est recherché : c'est le *keyword spotting*. La recherche est souvent formulée sous la forme d'un mot (Rath *et al.*, 2003), d'une image de mot (Cao *et al.*, 2007) ou bien les deux à la fois (Rodríguez-Serrano *et al.*, 2009a). Dans tous les cas, les systèmes mis en oeuvre présentent de nombreuses similitudes dans l'enchainement des traitements.

Généralement, les auteurs se ramènent au cas plus simple de la reconnaissance de mots manuscrits isolés. Pour ce faire, l'image du document est d'abord segmentée en lignes puis ses lignes sont segmentées en mots. Ceci est réalisé par une classification des espaces entre composantes connexes en espaces inter-mot ou intra-mot. Une mesure de distance entre composantes connexes combinée à un seuillage sur ces distances peut être utilisés (Marti *et al.,* 2001). Aussi, différents types de mesures de distance peuvent être combinées (Cao *et al.,* 2007). Dans (Rodríguez-Serrano *et al.,* 2009a), un classifieur *K-means* permettant de discriminer les types d'espaces entre eux est implémenté. Ces méthodes de segmentation présentent de bons résultats mais ne sont pas robustes au changement de base : une différence de résolution par exemple nécessitera un réapprentissage des hyperparamètres de ces algorithmes. De plus, isoler des images de mots dans des documents manuscrits sans procéder à leur reconnaissance est une tâche sujette à beaucoup d'erreurs souvent irreversibles.
La reconnaissance des mots isolés quant à elle est un peu plus dépendante du type d'entrée considérée par le système. Lorsque l'entrée est représentée par une image de mot, une mesure de distance dans l'espace de caractéristiques considéré est couplée à un seuil ce qui permet de prendre une décision (Saon *et al.,* 1997, Rath *et al.,* 2004). Il n'y a pas de reconnaissance à proprement parlé et ces systèmes, du fait de la nature des données d'entrée, sont plutôt mono-scripteur. Dans le cas d'une requête textuelle, l'utilisation d'un score de normalisation peut être envisagé (Choisy, 2007, Rodríguez-Serrano *et al.,* 2009b) pour étalonner au mieux les scores de reconnaissance. L'ajustement d'un tel score est long et fastidieux mais son utilisation est incontournable pour la prise de décision. Celui-ci est également fort dépendant des données et donc de la base utilisée. A contrario, les systèmes basés sur la reconnaissance de mot sont moins sensibles aux variabilités inter-scripteurs.

La seconde stratégie consiste à reconnaître le document dans sa globalité et à vérifier la présence ou non des mots clés reconnus. A l'instar du *keyword spotting*, les documents sont dans un premier temps segmentés en lignes et quelques fois en mots (Favata *et al.,* 1998, Marti *et al.,* 2001), cette phase de segmentation étant pilotée par la reconnaissance. Toutefois, le fait de travailler sur des phrases permet d'ajouter de nouvelles connaissances (Aubert, 2002) sous la forme de modèles statistiques du langage par exemple (Marti *et al.,* 2000) et ainsi contraindre la reconnaissance.
Dans (Vinciarelli *et al.,* 2004), les caractères, mots et phrases sont modélisés par des modèles de Markov (*HMMs*) sur différents niveaux couplés avec un modèle de langage de type bigrammes de mots. Cette approche, largement étudiée au début des années 2000, présente des résultats encourageants. Cependant, le gain en terme de résultats de reconnaissance grâce à l'ajout de connaissances (tri-grammes voire quadri-grammes de mots, lexique ouvert ...) l'est au dépend d'énormes coûts computationnels.

D'un côté, on peut dire que les systèmes de *keyword spotting* sont trop dépendant des données. De l'autre, les systèmes de reconnaissance de texte manuscrit sont bien trop gourmands en mémoire et lourds en calcul (lexique ouvert, modèle de langage immense). Une bonne alternative à ces approches semble être l'extraction d'information, l'information non pertinente étant considérée comme un type de données de base du système au même titre que les mots clés recherchés. Un modèle d'analyse de surface

permet d'isoler les mots clés de cette information non pertinente comme par exemple dans (Chatelain, 2006). Les auteurs mettent en oeuvre un système d'extraction de séquences numériques dont la syntaxe est a priori connue dans des courriers entrants. Le système intègre un modèle d'analyse statistique de surface sous la forme d'un classifieur discriminant de chiffres/non chiffres (Chatelain *et al.,* 2010) permettant d'isoler les séquences numériques dans des lignes de textes complètes. Ceci permet de passer outre le problème de la segmentation mais permet aussi de diminuer la complexité de la tache de reconnaissance. Une approche similaire a été choisie pour le développement de notre système. Nous le décrivons maintenant.

2.2. *Modélisation des données*

Le point clé de notre système réside dans la modèlisation des différentes entités que nous allons prendre en considération. Au regard des quelques systèmes présentés dans la section précédente, il semble peu pertinent de chercher à segmenter à priori les lignes en mots. Ceci présente l'inconvénient de ne pas prendre en compte les informations qui entourent le mot traité et cause des erreurs de mauvaises segmentations irréversibles. D'un autre côté, une approche plus globale comme la reconnaissance complète d'un document dans le but unique d'en extraire un contenu prècis semble trop lourde et rigide.

Nos choix se sont donc orientés vers une modélisation réfléchie de la ligne de texte. Les HMMs (*Hidden Markov Models*) reconnus pour être l'un des plus intéressant outil de modélisation de séquences de symboles (Rabiner, 1990), ont été choisis à cette fin. Ils sont notamment utilisés pour la reconnaissance de mots (Rodríguez-Serrano *et al.,* 2008) ou de phrases manuscrites (Vinciarelli *et al.,* 2004). Etant donnée la nature peu contrainte d'un document manuscrit, nous avons choisi de prendre en considération deux types d'informations à discriminer au sein d'une ligne de texte :

– le ou les mots pertinents formant un lexique de taille raisonnable, que l'on va chercher à extraire du document. Ils sont représentés par la concaténations des modèles HMMs de caractères les composant.

– le reste, identifiable à l'information non pertinente, est constitué des mots n'appartenant pas au lexique mais également des séquences numériques, des signes de ponctuations et de bruit. Il est représenté par un modèle statistique de surface sous la forme d'un modèle de langage de type bigramme. Dans notre cas, le modèle est appris sur les données de la base d'apprentissage.

Le modèle global de la ligne peut donc être représenté par le schéma de la figure 2. Il met en avant la mise en concurrence des deux types d'information modélisés : les mots clés du lexique modélisés par des modèles HMMs de mots et l'analyseur de surface modélisé par les modèles HMMs des symboles (lettres minuscules et majuscules, chiffres, signes de ponctutation). La variable G est un hyperparamètre du système à optimiser et permet de passer d'un type d'information à l'autre. Il peut également être vu comme la proportion d'information pertinente dans une ligne et est donc dépen-

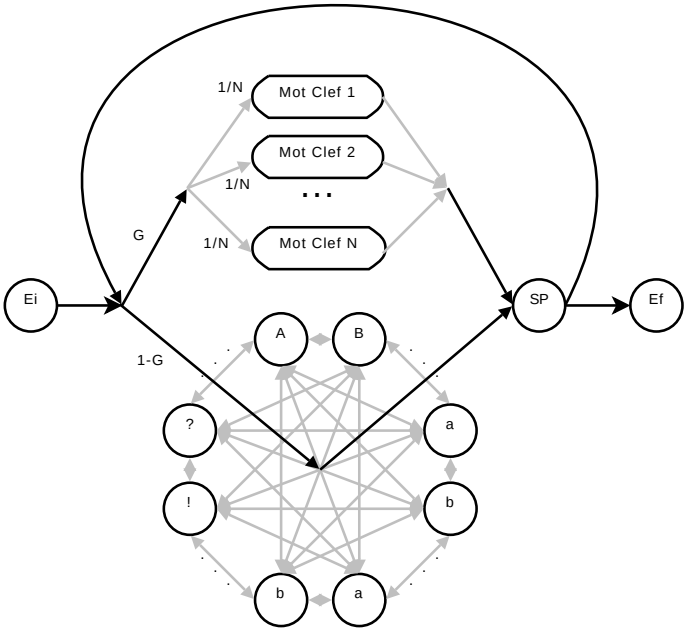

Figure 2. *Modèle de ligne intégrant un modèle surfacique de rejet et les mots du lexique représenté par des HMMs.*

dante de la taille du lexique. Faire varier sa valeur va nous permettre de nous déplacer dans le plan Rappel/Précision. Une ligne peut donc être vue comme une succession de mots clés et de mots non pertinents separés par des espaces. Les transitions entre ces différents éléments et notamment au sein du modèle de surface sont modélisés par un modèle statistique de langage appris sur la base de données d'apprentissage. Par exemple, la transition entre un espace et un caractère du modèle de surface est fixée à $1 - G$ correspondant à la proportion d'information non pertinente dans une ligne et est pondérée par la probabilité de commencer un mot par ce caractère (probabilité extraite du modèle de langage). De même, la transition d'un caractère du modèle de surface vers un espace est fixée à la probabilité que ce caractère termine un mot. A contrario, les transitions entre un espace et un mot clé sont fixées à G et pondérées par $\frac{1}{N}$ où N est le nombre de mots clés (nous supposons que les fréquences d'apparition des mots clés dans les courriers sont indentiques).

3. Chaîne de traitements

Le système complet est composé de deux chaînes de traitements quasi identiques : la première permet d'entrainer des modèles de caractères, la seconde permet de tester

notre approche. Seuls l'apprentissage et le décodage des HMMs diffèrent. Les prétraitements et l'extraction de caractéristiques sont communes, nous allons les décrire brièvement dans un premier temps.

3.1. *Prétraitements*

Ils sont classiques et sont intégrés dans plusieurs modules de manière séquentielle. Dans un premier temps, les images de document sont binarisées. Les lignes de texte sont ensuite extraites puisqu'il s'agit de la brique de base de notre système. Une méthode de regroupement des composantes connexes inspirée de (Likforman-Sulem *et al.,* 1995) est utilisée à cet effet.

Afin de réduire la variabilité entre les scripteurs, il peut être pertinent de chercher à redresser l'écriture. En ce qui concerne la correction de l'inclinaison horizontale, une transformation par rotation est effectuée. Pour la correction de l'inclinaison verticale, une transformation par cisaillement a été réalisée (Koch, 2006).

Dans un dernier temps, les lignes de base de l'écriture sont extraites, elles vont nous être utiles pour calculer certaines des caractéristiques.

3.2. *Extraction de caractéristiques*

Une fois les prétraitements effectués, les caractéristiques sont extraites depuis une fenêtre glissante de largeur d sur une ligne complète avec un recouvrement o entre deux fenêtres consécutives. Elles sont inspirées d'un jeu de caractéristiques tiré des travaux de (Kessentini *et al.,* 2010) qui ont très bien figuré à la compétition de reconnaissance de mots manuscrits d'ICDAR 2009 (Grosicki *et al.,* 2009) et est particulièrement adapté aux HMMs. Ces caractéristiques sont au nombre de $18 + d$:

– $3 + d$ ne dépendent pas des lignes de base (densité de pixels noirs dans la fenêtre, dans chacune des d colonnes de la fenêtre et position du centre de gravité des pixels noirs dans la fenêtre entre autres)

– 17 dépendent des lignes de base (densité de pixels au dessus et en dessous des lignes de base, configurations locales des pixels par rapport aux lignes de base etc.)

Les deux hyperparamètres d et o sont à fixer avant l'apprentissage et le décodage des différents modèles HMMs. Nous en reparlerons dans la section 4.

3.3. *Classification à l'aide de modèles de Markov cachés*

L'apprentissage des modèles de caractères est effectué à l'aide de l'algorithme de Baum-Welch (Baum *et al.,* 1966), algorithme d'apprentissage de modèles de caractères embarqué dans des modèles HMMs de lignes que nous construisons grâce aux annotations des textes de la base. Ceci permet de s'affranchir du délicat problème

de la segmentation des lignes en mots et surtout celui de l'annotation d'une quantité importante d'images au niveau caractère pour l'apprentissage.

Le décodage des modèles de ligne se fait lors de la phase de reconnaissance. Le problème de la reconnaissance de séquences de mots à base de HMMs peut s'écrire suivant l'équation 1, L_{opt} représentant la meilleure séquence de mots par rapport à la séquence d'observations O :

$$L_{opt} = \arg\max_{L}\{P(L|O)\} \quad [1]$$

$$= \arg\max_{L}\left\{\frac{P(O|L)P(L)}{P(O)}\right\} \quad [2]$$

$$= \arg\max_{L}\{P(O|L)P(L)\} \quad [3]$$

où les différents termes sont :

– $P(O|L)$ représente la probabilité d'observer la séquence O sachant la séquence de mots L

– $P(L)$ représente la probabilité de la séquence de mots L

– $P(O)$ représente la probabilité d'observer la séquence O

Le terme $P(O)$ est difficile à estimer. Cependant, il n'affecte pas la recherche de la meilleure séquence d'états par rapport aux observations. Il est logiquement supprimé des calculs.

Dans l'équation 3, l'information $P(L)$ représente la probabilité a priori d'une séquence de mots. Ceci va nous permettre de contraindre la reconnaissance en autorisant ou non certains enchainements de mots. Elle est estimée dans notre cas à l'aide d'un modèle de langage de type bigrammes appris sur la base d'apprentissage et modélise les transitions au sein du modèle statistique de surface mais aussi entre les différentes entités du système comme les mots du lexique ou le modèle d'espace.

Le terme $P(O|L)$ est lui estimé à l'aide de l'algorithme *Time Synchronous Beam Search* présenté dans (Moore, 2002). En notant MC_i, un mot du lexique reconnu dans la ligne, SP_k un espace dans cette ligne et MS_j un non mot, ce terme peut être explicité comme suit :

$$P(O|L) = \prod_{i}^{N} P(O_{M_i}|M_i) \quad [4]$$

$$= \prod_{i}^{N} P(O_i|MC_i) \prod_{j}^{M} P(O_j|MS_j) \prod_{k}^{K} P(O_k|SP_k) \quad [5]$$

$$P(O|L_{opt}) = \max_{i,j,k} P(O_i^*|MC_i)P(O_j^*|MS_j)P(O_k^*|SP_k) \quad [6]$$

L'équation 6 représente le meilleur enchainement d'entités constituant le modèle générique de ligne au regard des observations à savoir les mots clés du lexique (*MC*), les non mots caractérisés par le modèle de surface (*MS*) et le modèle d'espace (*SP*).

4. Expérimentations

4.1. *La base de données courrier entrants de Rimes*

La base de données Rimes (Grosicki *et al.,* 2009) comprenant 1150 courriers de différents scripteurs partiellement annotés au niveau mots a été utilisée pour mener à bien les expérimentations. Seule une partie du corps de texte est annotée, un travail d'étiquetage complet a donc été réalisé sur quelques courriers pour tester le système d'extraction d'information précédemment introduit. Voici le détail d'utilisation de la base Rimes :

– 950 courriers pour la base d'apprentissage soient 36587 mots

– 20 courriers pour la base de test

– 180 courriers non utilisés pour l'instant

Dans un premier temps, seuls 20 courriers ont été annotés. Le travail d'étiquetage va être poursuivi et idéalement permettre la mise à disposition d'une base de test de 200 courriers entrants manuscrits.

4.2. *Hyperparamètres et protocole expérimental*

Les hyperparamètres d et o (largeur de la fenêtre glissante et recouvrement entre deux fenêtres successives) ont été réglés sur la base de données Ironoff chèques (Viard-Gaudin *et al.,* 1999) comprenant des images de mots isolés extraites de chèques bancaires (lexique de 30 mots). Le couple (d, o) ayant donné les meilleurs résultats avec près de 90% de bonnes reconnaissances est le suivant : $(8, 5)$. Avec cette configuration, le nombre d'états optimal par modèle HMM de caractères est 4. Le nombre de gaussienne par états est lui fixé à 5.

Le choix de ce couple a également était validé sur une autre base de données mots afin de vérifier la robustesse du système aux changements de base. Une base de données interne de 6000 images de mots isolées a été utilisée à cet effet. Avec un lexique de 100 mots en test, le système atteint 79% de bonnes reconnaissances, 93% en Top 5 et 97% en Top 10.

Une fois ces hyperparamètres fixés et les modèles appris, le protocole expérimental est défini. Afin de tester le système en extraction d'information, il est commun de déterminer son rappel et sa précision en différents points de fonctionnement. L'hyperparamètre G (défini dans la section 2.2) permet de tester différents points de fonctionnement, favorisant les mots du lexique au profit du modèle de surface pour des valeurs de G proche de 1 (et donc le rappel) et inversement pour des valeurs proches de 0 (et

donc la précision). Il est à noter que sa valeur peut être fixée en fonction des besoins d'une application précise.

Ainsi, pour un G fixé et un courrier donné, un nombre N de mots présents dans ce courrier est recherché. On calcule le nombre de mots bien retrouvé N_{ok} d'un côté, le nombre de fausses alarmes N_{fa} de l'autre, ce qui nous permet de calculer un couple $(Rappel/Precision)$ pour ce courrier :

$$Rappel = \frac{N_{ok}}{N} \quad et \quad Precision = \frac{N_{ok}}{N_{ok} + N_{fa}} \qquad [7]$$

Répéter cette expérience pour l'ensemble des courriers de la base de test permet d'obtenir un point de la courbe mais aussi de calculer la moyenne et l'écart-type du rappel et de la précision en ce point.

4.3. *Résultats*

Le nombre de mots à chercher dans un courrier a été fixé à 10 ce qui permet de tester à la fois la capacité d'extraction de mots clefs mais aussi celle de rejet de l'information non pertinente. L'ensemble des résultats pour une quinzaine de valeurs de G variant entre 0 et 1 est présenté sur la figure 3.

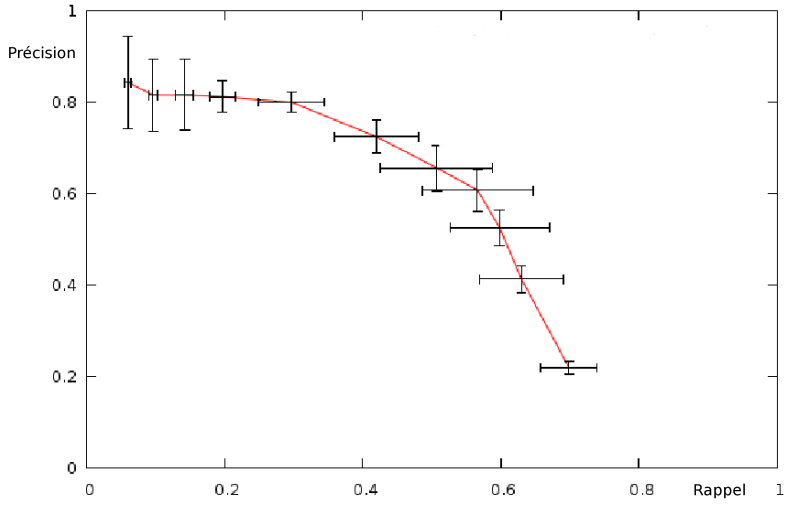

Figure 3. *Résultats obtenus sur la base de test de 20 courriers tirée de Rimes.*

Etant donné la nature des données calculées et la faible quantité de données, la variation des écarts-types de rappel et de précision peut paraître importante. Ceci s'ex-

plique en partie par la faible quantité de données à rechercher dans les courriers. Malgré tout, les résultats sont convainquants, en témoigne un *Break-Even Point* proche de 0.6 correspondant à un rappel et une précision de 60% et une *F-mesure* proche de $0,6$. Enfin, il est à noter que nous sommes les premiers à donner des résultats en extraction d'information sur cette base de données. Par conséquent, nous ne pouvons nous comparer à d'autres travaux.

5. Conclusion et perspectives

Nous venons d'introduire un nouveau système d'extraction d'information dans des documents manuscrits faiblement contraints de type courrier entrant. Après avoir exposé les différentes stratégies de la littérature afin de récupérer le contenu de ces documents, un nouveau modèle générique de ligne de texte est introduit. Celui-ci s'appuie sur les modèles de Markov cachés, particulièrement adaptés à la modélisation de séquences de symboles. Un modèle statistique de surface est utilisé pour modéliser l'information non pertinente et est mis en concurrence avec les modèles de mots clés du lexique dans le modèle global de ligne. Les premiers résultats illustrent le potentiel de l'approche.

A court terme, une base de test plus conséquente de 200 courriers va être mise en place. Ceci nécessite un travail d'étiquetage long et fastidieux mais nécessaire afin de calculer des performances véritablement significatives. Aussi, il va être intéressant de comparer le système présenté avec une méthode de reconnaissance complète de texte.

6. Bibliographie

Aubert X. L., « An overview of decoding techniques for large vocabulary continuous speech recognition », *Computer Speech & Language*, vol. 16, n° 1, p. 89-114, 2002.

Baum L. E., Petrie T., « Statistical Inference for Probabilistic Functions of Finite State Markov Chains », *The Annals of Mathematical Statistics*, vol. 37, n° 6, p. 1554-1563, 1966.

Cao H., Govindaraju V., « Template-Free Word Spotting in Low-Quality Manuscripts », *Proc. ICDAR*p. 392-396, February, 2007.

Chatelain C., « Extraction de séquences numériques dans les documents manuscrits quelconques », *Thèse de Doctorat, Université de Rouen*, 2006.

Chatelain C., Adam S., Lecourtier Y., Heutte L., Paquet T., « A multi-model selection framework for unknown and/or evolutive misclassification cost problems », *Pattern Recognition*, vol. 43, p. 815-823, 2010.

Choisy C., « Dynamic Handwritten Keyword Spotting Based on the NSHP-HMM », *Proc. ICDAR*, vol. 1, p. 242-246, 2007.

El-Yacoubi M. A., Gilloux M., Bertille J.-M., « A Statistical Approach for Phrase Location and Recognition within a Text Line : An Application to Street Name Recognition », *IEEE Trans. on PAMI*, vol. 24, n° 2, p. 172-188, 2002.

Favata J., Srihari S., Govindaraju V., « Off-line handwritten sentence recognition », *Proc. IWFHR*, vol. 1, p. 171-176, 1998.

Grosicki E., El-Abed H., « ICDAR 2009 Handrwriting Recognition Competition », *Proc. ICDAR*, vol. 1, p. 1398-1402, 2009.

Heutte L., Barbosa-Pereira P., Bougeois O., Moreau J.-V., Plessis B., Courtellemont P., Lecourtier Y., « Multi-Bank Check Recognition System : Consideration on The Numeral Amount Recognition Module », *IJPRAI*, vol. 11, n° 4, p. 595-618, 1997.

Kessentini Y., Paquet T., Benhamadou A., « Off-line Handwritten Word Recognition Using Multi-Stream Hidden Markov Models. », *Pattern Recognition Letters*, vol. 31, p. 60-70, 2010.

Koch G., « Catégorisation automatique de documents manuscrits : application aux courriers entrants », *Thèse de Doctorat, Université de Rouen*, 2006.

Likforman-Sulem L., Faure C., « Une methode de resolution des conflits d'alignements pour la segmentation des documents manuscrits », *Traitement du Signal*, vol. 12, p. 541-549, 1995.

Marti U., Bunke H., « Handwritten sentence recognition », *Proc. ICDAR*, vol. 3, p. 467-470, 2000.

Marti U., Bunke H., « Text line segmentation and word recognition in a system for general writer independent handwriting recognition », *Proc. ICDAR*, vol. 1, p. 159-163, 2001.

Moore D., « TODE : A Decoder for Continuous Speech Recognition », *IDIAP Research Report 02-09*, 2002.

Niyogi D., Srihari S., Govindaraju V., « Analysis of printed forms », *Handbook on Optical Character Recognition and Document Image Analysis*p. 330-342, 1996.

Nosary A., Heutte L., Paquet T., « Unsupervised writer adaptation applied to handwritten text recognition. », *Pattern Recognition*, vol. 37, n° 2, p. 385-388, 2004.

Rabiner L. R., « A tutorial on hidden Markov models and selected applications in speech recognition », *Readings in speech recognition*p. 267-296, 1990.

Rath T., Manmatha R., « Features for Word Spotting in Historical Manuscripts », *Proc. ICDAR*p. 218-222, 2003.

Rath T., Manmatha R., Lavrenko V., « A Search Engine for Historical Manuscript Images », *SIGIR 04*p. 369-376, 2004.

Rodríguez-Serrano J. A., Perronnin F., « Handwritten word-spotting using hidden Markov models and universal vocabularies », *Pattern Recognition*p. 2106-2116, February, 2009a.

Rodríguez-Serrano J. A., Perronnin F., Llados J., « Unsupervised writer style adaptation for handwritten word spotting », *ICPR 08*, 2008.

Rodríguez-Serrano J. A., Perronnin F., Llados J., « A similarity measure between vector sequences with application to handwritten word image retrieval », *CVPR 09*, August, 2009b.

Saon G., Belaïd A., « Off-line Handwritten Word Recognition Using A Mixed HMM–MRF Approach », *Fourth International Conference on Document Analysis and Recognition*, vol. 1, p. 118-122, 1997.

Viard-Gaudin C., Lallican P. M., Binter P., Knerr S., « The IRESTE On/Off (IRONOFF) Dual Handwriting Database », p. 455, 1999.

Vinciarelli A., Bengio S., Bunke H., « Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Langage Models », *IEEE Trans. on PAMI*, vol. 26, n° 6, p. 709-720, 2004.