Extraction automatique de champs numériques dans des documents manuscrits

Clément Chatelain, Laurent Heutte, Thierry Paquet

Laboratoire PSI, CNRS FRE 2645,
Université de Rouen, 76800 Saint Etienne du Rouvray, FRANCE
clement.chatelain@univ-rouen.fr

Résumé. Nous décrivons dans cet article une chaine de traitement complète et générique permettant d'extraire automatiquement les champs numériques (numéros de téléphone, codes clients, codes postaux) dans des documents manuscrits libres. Notre chaïne de traitement est constituée des trois étapes suivantes: localisation des champs numériques potentiels selon une approche markovienne sans reconnaissance chiffre ni segmentation, reconnaissance des séquences extraites, et vérification des hypothèses de localisation / reconnaissance en vue de limiter la fausse alarme génerée lors de l'étape de localisation. L'évaluation de notre système sur une base de 300 courriers manuscrits montre des performances en rappel-précision intéressantes.

1 Introduction

Aujourd'hui, la lecture automatique des documents manuscrits se limite à quelques cas applicatifs particuliers : lecture automatique de chèques ou d'adresses postales, reconnaissance des champs d'un formulaire. Cette lecture est possible car le contenu de ces documents est très largement contraint : structure du document stable, position des informations connue, redondance de l'information, lexique limité, etc. Lors de la lecture, le système bénéficie ainsi d'informations *a priori* importantes permettant de limiter ou de vérifier les hypothèses de reconnaissance, autorisant une lecture fiable des documents.

Peu de travaux abordent des problèmes de reconnaissance moins contraints car il est alors plus difficile de bénéficier de moyens automatiques de vérification des hypothèses de reconnaissance. C'est le contexte de nos travaux portant sur la lecture automatique des courriers entrants manuscrits. Il s'agit de courriers manuscrits tels que des lettres de réclamation, de changement d'adresse, de modification de contrat, etc., reçus en très grand nombre quotidiennement par des grandes organisations. Contrairement aux applications précédemment citées, aucune information a priori n'est disponible : le contenu, la structure, l'expéditeur ou encore l'objet du document sont totalement inconnus du système de lecture, ce qui rend la lecture intégrale du document extrêmement délicate. Il est cependant possible de considérer des problèmes de lecture partielle du document, visant à en extraire l'information pertinente. C'est ce que nous envisageons dans cet article en proposant une méthode de localisation et de reconnaissance de champs numériques (numéros de téléphones, codes clients, etc.) dans des courriers entrants manuscrits (voir figure 1). La reconnaissance de ces champs permettra par