Deux stratégies pour l'extraction automatique de champs numériques dans des documents manuscrits

Clément Chatelain, Laurent Heutte, Thierry Paquet

Laboratoire LITIS, Université de Rouen
76800 Saint Etienne du Rouvray, FRANCE
clement.chatelain@univ-rouen.fr

Résumé. Dans cet article, nous abordons le problème de l'extraction automatique de champs numériques dans des courriers entrants manuscrits. Nous montrons que deux stratégies opposées peuvent être considérées : la première est basée sur une approche issue des techniques classiques de reconnaissance de séquences numériques, et la seconde est inspirée des méthodes d'extraction d'information dans les documents électroniques. Les résultats montrent que cette seconde approche est plus pertinente.

Problématique

Actuellement, les systèmes industriels de lecture automatique de documents se limitent aux applications bien connues de lecture automatique de chèques bancaires, d'adresses postales ou de formulaires. Ces applications industrielles sont parfaitement opérationnelles et traitent plusieurs millions de documents par jour. Cependant, hormis ces applications très spécifiques, la reconnaissance de l'écriture manuscrite reste toujours un problème délicat en l'absence de connaissances *a priori* sur les documents traités. Depuis quelques années, un nouveau tournant a été amorcé avec une orientation des recherches vers la lecture automatique de documents aux contenus moins contraints tels que les textes libres. Initialement dénués de motivations applicatives industrielles, ces travaux ont cherché à effectuer une lecture intégrale de textes. Plus récemment, les besoins industriels se sont précisés, dans l'optique d'effectuer un traitement automatique des masses de courriers manuscrits reçus quotidiennement en très grand nombre par les grandes entreprises ou administrations. Afin de traiter cette masse de documents appelée *courrier entrant*, des méthodes d'extraction d'information sont mises en œuvre, visant à résumer un document par un ensemble de champs pertinents tels que l'objet du courrier, le nom de l'expéditeur, la date de l'envoi du courrier, etc.

Le sujet traité dans ces travaux concerne l'extraction de séquences numériques dans des documents manuscrits quelconques, et se situe donc pleinement dans cette nouvelle problématique. Il s'agit d'extraire des séquences numériques qui constituent une information pertinente pour la tâche de traitement automatique du courrier. Les numéros de téléphone, les codes postaux ou les numéros de contrat permettent par exemple d'effectuer un tri du courrier vers le service compétent dans l'entreprise. La problématique se situe donc au croisement de deux domaines de recherches : la *reconnaissance de l'écriture manuscrite* et l'*extraction d'information*. Si ces deux disciplines ont été largement étudiées indépendamment, les travaux concernant l'extraction d'information dans les documents manuscrits sont beaucoup plus rares. Dans ces travaux, nous proposons deux stratégies opposées pour l'extraction d'information numérique dans des images de documents.

Première stratégie

La première stratégie est une extension des méthodes traditionnelles de reconnaissance de séquences numériques isolées : montant de chèque, code postal, etc. Elle repose sur l'idée relativement évidente qu'une ligne de texte est composée d'une séquence numérique entourée d'information non pertinentes. D'où la nécessité d'ajouter un processus de rejet au processus classique de "segmentation/reconnaissance" classiquement utilisé pour la reconnaissance de séquences isolées. On obtient alors une stratégie de segmentation/reconnaissance/rejet qui permet de localiser et de reconnaître les chiffres dans les lignes de texte du cocument. Il ne reste plus qu'à détecter les séquences de chiffres qui nous intéressent : par exemple une séquence de 10 chiffres qui se suivent pour un numéro de téléphone, 5 chiffres pour un code postal, etc. Celà est réalisé par une recherche de meilleur chemin dans le treillis des hypothèses de segmentation/reconnaissance/rejet. Cette méthode peut être vue comme la plus évidente. L'implémentation de cette stratégie a donné des résultats acceptables, permettant d'extraire environ 50% des champs recherchés en première proposition Chatelain et al. (2006a,c).

Seconde stratégie

La seconde stratégie cherche à s'inspirer des méthodes d'extraction d'information dans les documents électroniques. Elle repose sur l'idée que la reconnaissance des entités numériques n'apporte rien pour la localisation des champs. Par conséquent, les phases de localisation et de reconnaissance sont disjointes. Cette méthode est plus originale dans la mesure où, en vertu du paradoxe de Sayre, ces deux opérations sont généralement liées afin de se fiabiliser mutuellement. L'approche proposée ici pour la localisation des champs est basée sur une modélisation markovienne d'une ligne de texte. Ce modèle exploite la syntaxe spécifique des champs numériques que l'on souhaite extraire (nombre de chiffres, présence et position de séparateurs) pour parvenir à localiser les séquences numériques, sans toutefois procéder à la reconnaissance des chiffres. C'est en effet par une étape de pré-reconnaissance ńsyntaxiqueż que l'on va chercher à interpréter globalement les lignes de texte. Des classes syntaxiques sont ainsi définies afin de décrire la nature alphabétique ou numérique des composantes, sans pour autant préciser leur valeur numérique : Chiffre, Séparateur, Chiffres liés, Rejet. Une fois les champs localisés dans le treillis de reconnaissance syntaxique, les séquences ainsi localisées sont reconnues par une méthode classique de reconnaissance de séquences isolées. Cette méthode a permis d'obtenir des résultats supérieurs à la première stratégie Chatelain et al. (2004, 2006b).

Application et résultats

Afin de comparer les résultats, deux chaînes de traitement complètes ont été mise en œuvre, nécessitant l'implémentation des différents modules : segmentation du document en lignes, classifieur et segmenteur chiffre, méthode de rejet des éléments non numériques, classification syntaxique, apprentissage des modèles de ligne. Les résultats montrent que cette seconde méthode semble être le meilleur moyen d'aborder le problème puisque ses performances en rappel-précision dépassent celles de la première. Près de 60% des champs sont ainsi correctement extraits en première proposition.

Il en résulte un système complet, générique et industrialisable permettant d'effectuer l'extraction de séquences numériques dans des documents manuscrits faiblement contraints. La suite de ces travaux concerne la fusion de ce système avec l'approche développés par Koch pour l'extraction de mots clefs dans ces mêmes courriers manuscrits, en vue de fournir un système d'extraction d'information de haut niveau dans des documents complexes.

Références

Chatelain, C., L. Heutte, et T. Paquet (2004). A syntax-directed method for numerical field extraction using classifier combination. *9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan*, 93–98.

Chatelain, C., L. Heutte, et T. Paquet (2006a). Discrimination between digits and outliers in handwritten documents applied to the extraction of numerical fields. *International Workshop on Frontiers in Handwriting Recognition, La Baule, France*, 475–480.

Chatelain, C., L. Heutte, et T. Paquet (2006b). Segmentation-driven recognition applied to numerical field extraction from handwritten incoming mail documents. *Document Analysis System, Nelson, New Zealand*, 564–575.

Chatelain, C., L. Heutte, et T. Paquet (2006c). A two-stage outlier rejection strategy for numerical field extraction in handwritten documents. *International Conference on Pattern Recognition 3*, 224–227.

Summary

Donner la traduction anglaise du résumé dans le préambule avec la commande `\summary{Your abstract ...}`