Reconnaissance optique des caractères (ROC ou OCR)

Édition au 26-mars-14 ATTENTION CETTE VERSION EST EN TRAVAUX !!!!!!!!!!!!!!!!!!!!!

Les documents écrits sont une forme de mémoire de l'humanité. Leur diffusion électronique est difficile : il est évidemment possible de faire des images de chaque page d'un roman mais ces images occupent une place importante et l'on ne peut apporter de corrections dans le texte.

Des techniques logicielles ont été développées pour transformer ces pages de texte en fichier utilisables comme s'ils avaient été frappés sur l'ordinateur. C'est la Reconnaissance optique des caractères (ROC ou OCR en anglais).

Le logiciel examine l'image et la traite pour en extraire les caractères. Cette opération est bien sûr d'une grande complexité. Elle se heurte aux différentes polices utilisées pour les caractères, aux colonnes éventuelles, aux saletés scannées, aux images et aux tableaux. Certains logiciels arrivent à gérer tous ces éléments et restituer un texte « relativement propre ». Il ne faut toutefois pas se leurrer : des corrections sont nécessaires après l'OCR. L'exemple du l illustre bien les problèmes : l est-il un L minuscule ou un I majuscule ?

L'ajout d'une « intelligence artificielle » pourra peut-être améliorer l'efficacité de l'opération grâce à une certaine compréhension par la machine de ce qu'elle lit. Le même problème est rencontré avec l'orthographe et la grammaire.

Quant à l'écriture manuscrite, son décodage s'avère particulièrement difficile.

On peut toutefois constater, depuis les débuts de l'OCR, combien les progrès ont été importants.

Corrélats :