VOUS NE LE SAVEZ PAS, MAIS VOUS AIDEZ À LA NUMÉRISATION DES OUVRAGES ANCIENS !

Témoignage

Grâce au reCapcha, projet créé par l'université américaine Carnegie-Mellon, chaque fois que vous vous crevez les yeux pour décrypter un texte tout tordu afin de valider une inscription ou poster un commentaire, vous participez à l'amélioration d'un programme de numérisation des livres anciens.

ON REPRENDS DEPUIS LE DÉBUT : LES *CAPCHAS* C'EST QUOI ?

Les *capchas* se sont ces mots tous déformés que l'on vous demande de recopier afin de valider une inscription ou une action sur un site internet. Le principe du *capcha* c'est de trouver une opération plus facile à effectuer par un être humain que par un robot, l'objectif étant d'éviter que l'on puisse effectuer automatiquement (*via* des programme-robots) l'action que vous êtes en train de réaliser. Notamment pour éviter que l'on puisse créer automatiquement des milliers de comptes *mail* ou Facebook ou submerger de commentaires un *blog* afin de vendre du faux Viagra. En retapant le texte déformé, vous prouvez ainsi que vous êtes un être humain (uniquement d'un point de vue biologique, hein).

LA PROBLÉMATIQUE DE NUMÉRISATION DES LIVRES ANCIENS

Les livres anciens, tombés dans le domaine public, pourraient facilement être mis à disposition du plus grand monde sur internet mais pour faciliter la recherche parmi ces ouvrages, il faut transformer le *scan* de la page (qui est une photographie) en texte numérisé dans lequel on peut rechercher. Ce sont les logiciels de reconnaissance de caractères (OCR) qui s'en chargent mais ils rencontrent des difficultés particulières avec ces ouvrages. En effet, ceux-ci sont imprimés avec des caractères typographiques particuliers et le temps a souvent abîmé les pages. Pour améliorer leur taux de reconnaissance, les logiciels de reconnaissance de caractère (OCR) ont besoin "d'apprendre". C'est-à-dire qu'ils ont besoin que leurs résultats soient confrontés à des résultats obtenus par des humains pour augmenter

1

peu à peu le nombre de signes qu'ils peuvent reconnaître. Or la transcription par les humains est longue et rébarbative.

ET SI ON JOIGNAIT L'UTILE À ... L'UTILE ?

Luis Van Ham est professeur à l'université Carnegie-Mellon à Pittsburgh, il travaille sur l'*human computation*, c'est-à-dire sur des programmes faisant intervenir la puissance de raisonnement humain et la vitesse de calcul des ordinateurs pour résoudre des problèmes que ni les humains, ni les machines ne pourraient résoudre seuls (le cas des logiciels OCR est un exemple typique). Il a développé le concept de jeux à objectifs, où tout en jouant les êtres humains effectuent des opérations utiles. Même s'il ne s'agit pas d'un jeu, le reCapcha qu'il a développé, reprend ce principe. A chaque fois que vous décodez des mots déformés, issu de la numérisation de livres anciens, pour prouvez au site internet que vous êtes bien un être humain, vous augmentez la base de données utilisée par les logiciels de reconnaissance de caractères et donc leur efficacité à reconnaître les caractères numérisés des livres anciens.

RECAPCHA, ÇA MARCHE COMMENT ?

Les *scans* des livres anciens sont lus par deux logiciels de reconnaissance de caractères différents. Dès qu'un mot est lu différemment par les deux logiciels, il est noté comme suspicieux et ajouté à la base de reCapcha.
Quand on vous demande de prouver le fait que vous êtes un être humain et non une machine *via* un reCapcha, il y a toujours 2 termes, l'un plus déformé que les autres. L'un deux a déjà été identifié avec certitude avec le logiciel OCR (c'est celui qui sert effectivement à vérifier que vous êtes un humain) et l'autre non (c'est celui que vous allez aider à identifier). À partir du moment où un certain nombre d'internautes ont identifié de la même façon un mot suspicieux, il est validé. Il est intégré dans la base de données des mots validés de reCapcha et dans la base de données que le logiciel d'OCR utilise pour reconnaître les caractères des livres numérisés. Actuellement le logiciel de reconnaissance de caractère de reCapcha a atteint un niveau d'erreur semblable à celui de l'être humain.

C'EST UNE BONNE OEUVRE ALORS ?

Google a racheté reCapcha en 2009 et l'a installé sur ses pages demandant ce type de validation. Vu la puissance de Google cela a donné une très grande visibilité au projet et un nombre accru de participants. L'objectif premier de Google est la numérisation des

2

livres de Google Books afin de faciliter leur référencement et la recherche plein texte parmi les pages. Mais il semblerait que Google adapte reCapcha à d'autres projets, on a ainsi vu apparaître sur certains reCapcha des numéros de plaques de rue provenant de Google Street View.

Auteur de la fiche

Hélène Laxenaire - SupAgro Florac

3