Proposition de stage recherche en laboratoire 2019-2020

Titre : Transcription de documents anciens par apprentissage profond

Description du sujet :

Dans le cadre de leurs recherches, les historiens sont confrontés à la relecture de nombreux documents manuscrits anciens. Dès lors qu'il s'agit d'archives volumineuses (recensement, actes notariés, comptabilités, etc.), ce travail est fastidieux et souvent rendu complexe du fait de la présence d'écritures différentes.

Un premier projet exploratoire (2017/2018) a permis d'étudier la possibilité d'automatiser en partie ce travail de déchiffrage, en utilisant les acquis de la reconnaissance de caractères, après numérisation (photographie) des documents concernés.

Ce projet a permis de (i) photographier les documents correspondant à un corpus de taille relativement modeste (recensement de 1790 de la ville d'Arras), (ii) extraire et indexer manuellement quelques éléments pertinents d'informations manuscrites et (iii) étudier les méthodes d'apprentissage automatique susceptibles de fournir la meilleure réponse au problème.

Les premiers résultats de cette étude ont montré la faisabilité de la reconnaissance automatique de caractères anciens manuscrits, avec toutes les problématiques qui peuvent être associées à cette tâche : caractères disparus, abréviations, supports de mauvaise qualité, scribes multiples, etc.

En fonction des premiers résultats obtenus, il est maintenant envisagé une poursuite à ce travail, en visant à la fois des documents plus anciens, mais également plus volumineux, fournis par les historiens (cf exemple ci-après). Cette poursuite permettra d'étudier la mise au point d'une méthode d'analyse générique compatible avec des corpus d'origine et d'époques variées, ainsi que la définition d'outils efficaces pour l'historien, lors de la phase d'exploitation.

Ce module « end to end » sera basé sur une mise en place et une comparaison de structures d'apprentissage profond (Alexnet, ResNet, Autoencoder, codés en langage Python) pour obtenir la meilleure efficacité de transcription.

Objectifs :

Le projet consiste donc à appliquer la représentation d'attributs d'images de structures liées au « Deep Learning » à la transcription de documents anciens. Dans ce sens, les choix des structures deviennent prépondérantes et ces structures seront évaluées suivant les caractéristiques classiques (précision, recall, ...). Pendant ce stage, une étude bibliographique poussée sera mise en œuvre (reconnaissance de caractères,

particularités des documents anciens des 17 et 18 éme siècles). Les programmes seront développés et intégrés dans le logiciel initial. L'étude sera étendue aux problèmes de sélection et extraction d'attributs, de classifications supervisées et semi-supervisées.

L'objectif final est de bâtir une méthode générique permettant d'appliquer ces méthodes à d'autres textes manuscrits de grand volume (Moyen-Age, …), voir d'autres applications des SHS.

Références

Yahya Abbass, rapport de stage master TSI/STIP, ULCO-UL, 2017/18

Documents de la journée de restitution A2U du 17/10/2018.

Encadrant(s) : A.Bigand (MC, HdR)

email :bigand@univ-littoral.fr