Numériser des fiches de retour d'expériences sur le développement de lanceurs spatiaux

Elvis MBoning [1], Nadège Lechevrel [1], Michal Kurela [2], Damien Nouvel [1]
(1) ERTIM, INALCO, 2 rue de Lille, 75007 Paris, France
(2) CNES, 52, rue Jacques Hillairet 75612 Paris CEDEX, France

Mots-clés : OCR, LSTM, REX, Extraction d'information, Sûreté de fonctionnement
Type de soumission : sénior

Résumé de la présentation

Le travail présenté a été réalisé dans le cadre d'un projet initié par le CNES avec léquipe ERTIM de l'INALCO. Le CNES et ses sous-traitants industriels produisent depuis plusieurs décennies un nombre considérable de documents constitués de « retours dexpérience » (REX). Parmi ceux-ci, des « fiches de point critique » (FPC), qui renseignent les risques majeurs encourus durant les phases de développement des lanceurs ARIANE. L'enjeu du projet concerne l'exploitation de ces fiches scannées par le CNES sous forme d'images encapsulées dans des fichiers au format PDF. Un bon nombre de ces documents a fait l'objet de campagnes de numérisation et de traitements, pour que les experts puissent les exploiter plus facilement : recherches textuelles, extraction d'information, ingénierie des connaissances, tendances et signaux faibles.

Dans le cadre de ce projet, nous avons utilisé le logiciel OCR openSource *ocropy* (`https://github.com/tmbdev/ocropy`), basé sur des réseaux de neurones (LSTM), pour adapter l'OCR à ces données du domaine spatial faisant appel à une terminologie (vocabulaire spatial et technique, mesures). Une surcouche dédiée a été développée, appelée *ScanRexs*, afin d'adapter la numérisation au format PDF fourni et de faciliter les opérations d'entraînement et d'évaluation. Cette adaptation, réalisée par transcription manuelle d'un échantillon réduit de 124 fiches puis par apprentissage automatique de modèles, a permis d'améliorer les performances de l'OCR, depuis un (WER) de 37% (modèle par défaut), jusque 22% (ou un CER de 10% jusque 4%). les modèles entraînés ont ensuite été utilisés sur un lot de 3000 fiches.

Ce projet a montré l'efficacité de l'apprentissage automatique neuronal récurent (RNN) sur ces documents, malgré un volume de données d'apprentissage limité. Cependant, des difficultés ont été rencontrées, tant au niveau des formats d'encodage des images encapsulées dans les PDF, que par l'instabilité de l'apprentissage automatique sur des contenus de documents qualitativement et quantitativement hétérogènes. Parmi les perspectives, nous souhaitons améliorer le traitement des différents formats, stabiliser l'apprentissage et avoir une vision plus précise des points forts et points faibles des modèles appris, en particulier pour la terminologie du domaine spatial.

Références

[1] *Improving OCR Accuracy on Early Printed Books by utilizing Cross Fold Training and Voting.* Christian Reul, Uwe Springmann, Christoph Wick, Frank Puppe. 2017

[2] *OCR and post-correction of historical Finnish texts.* Senka Drobac, Pekka Kauppinen, Krister Lindén. 2017

[3] *PoCoTo - an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts.* Vobl, Thorsten and Gotscharek, Annette and Reffle, Uli and Ringlstetter, Christoph and U. Schulz, Klaus. In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage. Vol. 6815. P.57-61. 2014

[4] *High Performance OCR for Printed English and Fraktur using LSTM Networks.* Thomas M. Breuel, Adnan Ul-Hasan, Mayce Al Azawi. Faisal Shafait. 2013

[5] *An Overview of the Tesseract OCR Engine.* R. Smith. Ninth International Conference on Document Analysis and Recognition (ICDAR). Vol. 2. P.629-633. Sept. 2007