Reconnaissance automatique de caractères avec l'OCR et l'HTR

30 janvier 2025

Denisa-Florina BUMBA (IGE, projet ANR EcoLe, Sorbonne Nouvelle)
denisa-florina.bumba@sorbonne-nouvelle.fr

Marine TIGER (Cellf, CERES)
marine.tiger@sorbonne-universite.fr

Définitions

- *Optical Character Recognition* (OCR) = reconnaissance de caractères pour les imprimés

- *Handwritten Text Recognition* (HTR) = reconnaissance de texte manuscrit

- *Automatic Text Recognition (*ATR) = englobe les outils de reconnaissance automatique de texte

- Exemple d'un projet HTR: FoNDUE à l'université de Genève

Définitions

- Gain de temps
- Traiter de plus large et vaste ensemble de document
- Transcrire différents types de documents de périodes différentes
- Ajouter de la valeur aux documents historiques

Définitions

Différents défis dans la reconnaissance de documents historiques:

- Constituer un corpus suffisamment vaste pour entraîner un modèle: https://htr-united.github.io/catalog.html

- Nettoyer le corpus: tâches d'encre, encre qui traverse la page, plusieurs mains, etc…

Préparer son corpus

📸 Collecte des images

Qualité des images :

- formats images recommandés : JPG, JPEG, PNG
- 300 DPI (Dot per inch) : qualité optimale
- < 200 DPI : qualité insuffisante pour une bonne reconnaissance
- > 400 DPI : l'augmentation de la qualité devient négligeable par rapport au poids des fichiers

Manifeste IIIF et la BnF :

- Entièreté du document : eScriptorium
- Sélection d'une plage de téléchargement d'images :

scripts Python pour la récupération des images au format IIIF :
https://github.com/Jean-Baptiste-Camps/IIIF-Crawler (liens Gallica, e-codices, bvmm, etc)
https://gitlab.com/eman8/Ecolired/travail_sur_les_images/-/tree/main/extract_IIIF (uniquement pour les liens Gallica)

🔧 Prétraitement des images

- Amélioration du contraste du texte dans l'image
- Réduction du bruit (tâches d'encre, autre)
- Découpage des images

Précautions concernant la binarisation : dans certains cas ce traitement peut empirer les résultats particulièrement pour les documents avec une écriture pâle ou à un éclairage inégale.

scripts pour le prétraitement et le découpage des images :
https://gitlab.com/eman8/Ecolired/travail_sur_les_images/-/tree/main/image_preprocessing

https://gitlab.com/eman8/Ecolired/travail_sur_les_images/-/tree/main/decoupage_images

Quel outil OCR/HTR choisir ?

Outil	Interface utilisateur	Prétraitement d'image	Prise en main	Formats d'entrée	Formats de sortie	Modèles OCR/HTR intégrés	Possibilité de réentraînement modèle layout	Possibilité de réentraînement modèle HTR/OCR
eScriptorium	X	-	+	PDF, JPG, JPEG, PNG, TIFF	XML PAGE, XML ALTO, TXT	HTR United	X	X
Transkribus	X	-	+	PDF, JPG, JPEG, PNG	XML PAGE, TXT, DOCX, Transkribus PDF	X	X	X
ocr4all	X	X	+	PDF, JPG, JPEG, PNG	XML PAGE, TXT	X	-	X
Tesseract	-	-	++	JPG, JPEG, PNG, TIFF	TXT, hOCR, XML ALTO, PDF (interrogeable)	X	X	X
Kraken	-	-	++	JPG, JPEG, PNG	TXT, hOCR, XML ALTO, XML PAGE, JSON	X	X	X
docTr	-	-	+++	PDF, JPG, JPEG, PNG	JSON ou XML	X	X	X

Étape 1 : Segmentation & mise en page

- Deux niveaux de segmentation : zones de texte & lignes
- Il est important de choisir le bon modèle ou réentraîner à partir d'un modèle de segmentation de base.

  ex. : https://github.com/Gallicorpora/Segmentation-and-HTR-Models/tree/main/Segmentation_model

- Standard recommandé pour l'annotation des segments de texte : SegmOnto

Note : Une annotation au préalable des zones de textes (savoir qu'il s'agit d'une zone de titre, d'une colonne, de la zone principale de texte) peut beaucoup faciliter l'extraction des informations concernant la mise en page en post-traitement.

Cependant, si le focus pour vous est le texte uniquement et que la segmentation est majoritairement correcte, l'annotation des zones de textes n'est pas nécessaire.

Étape 2 : Reconnaissance automatique du texte (OCR/HTR)

Une fois que les images ont été segmentées, la reconnaissance du texte peut commencer.

Facteurs à considérer dans le choix de modèle OCR/HTR :

- support : manuscrit ou imprimé
- état de langue : certains modèles couvrent plusieurs siècles ; d'autres sont spécialisés sur une écriture en particulier d'une période précise.

Catalogue en ligne des modèles OCR/HTR : HTR United

Pour aller plus loin :

Pipeline complète à lancer dans le terminal : https://github.com/Gallicorpora/application?tab=readme-ov-file
Vidéo démo pipeline complète : https://www.youtube.com/watch?v=iSpGQuKMvlY

Pratique avec *eScriptorium*

Lien vers le modèle à télécharger: https://zenodo.org/records/10886224

Documentation eScriptorium pour la prise en main :
https://escriptorium.readthedocs.io/en/latest/

Récapitulatif

Etapes importantes:

- Collecte du corpus
- Pré-traitement des images
- Choisir l'outil de transcription et un modèle adapté à ses besoins
- Segmentation du texte
- Transcription automatique du texte
- Récupération des fichiers transcrits

Liens pratiques

- pour convertir du format txt/docx vers XML TEI:

  https://obtic.huma-num.fr/teinte/

- dépôt Zenodo avec plusieurs modèles d'HTR/OCR:

  https://zenodo.org/communities/ocr_models/records?q=&l=list&p=1&s=10&sort=newest