Courte évaluation de la Reconnaissance de caractères (OCR) du modèle français de Tesseract par le WER

Alice Breton - alice.breton@lip6.fr

Septembre 2024

1 Description

Le but de cette expérimentation est d'évaluer les performances du modèle OCR de Tesseract, version 5.3.4, pour la langue française. L'objectif principal est de mesurer la qualité de la transcription des textes d'un corpus littéraire français, en analysant le Taux d'erreur de mots (Word Error Rate ou WER) et en les catégorisant : la substitution, l'insertion ainsi que la suppression.

2 Corpus

Le corpus utilisé pour l'évaluation est composé de 105 textes issus du livre littéraire "Le style, mode d'emploi" de Stéphane Tufféry. L'ensemble des textes contient un total de 20 198 tokens. Ce corpus littéraire représente une variété de styles d'écriture diverses.

3 Méthodologie

Pour mener cette expérience, plusieurs scripts automatisés ont été utilisés pour gérer le traitement des images et l'évaluation du modèle de Tesseract. Le processus global de l'expérimentation se déroulait en plusieurs étapes :

- **Capture des images** : Prise de photos avec un téléphone Android. Ce procédé était plus rapide que le scan et il n'y avait pas de grande différence significative d'OCR entre les images scannées ou capturées par un téléphone.
- **Prétraitement des images** : À partir des images du texte, un script Python appliquait diverses transformations d'amélioration d'image (`./ocr_tesseract/scripts/process_image.py`), telles que la rotation pour corriger l'orientation et un traitement de binarisation afin d'augmenter le contraste entre le texte et le fond.
- **Reconnaissance de texte** : Le modèle Tesseract 5.3.4 avec le modèle linguistique français intégré (`-l fra`) a été utilisé pour extraire le texte des images traitées.

— **Évaluation de la qualité de l'OCR** : Un autre script a permis de comparer les résultats obtenus par Tesseract aux fichiers de référence via le calcul du WER (Word Error Rate) (`./ocr_tesseract/scripts/eval.py`). Cette métrique mesure le nombre de mots incorrectement reconnus par rapport aux textes originaux. Les erreurs ont été classées en trois catégories : substitutions, insertions et suppressions.

4 Résultats

Les résultats de l'expérimentation montrent un taux d'erreur global de **16,31%** pour le WER, ce qui reflète la qualité moyenne de la transcription effectuée par Tesseract sur ce corpus spécifique. Le tableau suivant résume les types d'erreurs observés :

Type d'Erreur	**Nombre**	**Pourcentage (%)**
Substitutions	1 646	52,12
Insertions	1 246	39,46
Suppressions	266	8,42
Total des Erreurs	**3 158**	**100,00**

TABLE 1 – Types d'erreurs de reconnaissance de Tesseract 5.3.4

Référence	**Hypothèse**	**Nombre**
je	Je	48
-	—	46
la	Ja	25
jeune	Jeune	22
Il	I]	18
-	__	9
le	Île	8
jamais	Jamais	6
!		6
jamais	Jamais	6
l'autobus	autobus	5
lui	Jui	5
Il	[1	5
le	Je	5
...	...	1432
Total		1646

TABLE 2 – Substitutions les plus fréquentes de Tesseract 5.3.4

Les substitutions représentent plus de la moitié des erreurs totales (52,12 %), suivies par les insertions (39,46 %) et enfin les suppressions (8,42 %). Les erreurs de substitution

les plus fréquentes concernaient principalement les caractères : j, J, i, I, l, L, - et —. Par exemple, la substitution la plus courante était « je » transformé en « Je ».

5 Discussion des résultats

Les erreurs observées sont typiques des systèmes OCR lorsqu'ils rencontrent des variations typographiques ou des caractères visuellement proches. Tesseract a montré des difficultés particulières avec :

- **Les majuscules et minuscules** : Des mots comme « je » et « Je » ont souvent été confondus.
- **Les caractères spéciaux** : Les ponctuations et traits d'union ont été mal reconnus à certains moments.
- **Les confusions avec des caractères similaires** : La reconnaissance incorrecte de caractères tels que « I », « l » et le chiffre « 1 » a été récurrente.

Il est important de mentionner que les textes comportent des *mots inventés* pour respecter un certain style littéraire. Sans ces mots inventés, un meilleur WER aurait été obtenu.

6 Conclusion

L'expérimentation a montré que le modèle OCR de Tesseract pour le français présente des limites, mais produit des résultats acceptables avec un WER de 16,31 % sur un corpus littéraire. En somme, la majorité des erreurs sont des substitutions. Il serait préférable de refaire une même analyse en éliminant les textes qui contiennent des mots inventés.