Post-traitement de la transcription automatique des manuscrits latins médiévaux grâce à un tokeniseur spécialisé et à des modèles de transformers

Svetlana Yatsyk*[1]

[1]CIHAM – CIHAM (UMR 5648) – France

Résumé

Ce projet établit un pont entre l'informatique et les sciences humaines pour aborder les défis complexes associés à la transcription automatique des manuscrits latins médiévaux. Les modèles de transcription automatique atteignent généralement un plateau de précision, entre 94% et 97% au niveau des caractères. L'affinement de ces modèles exige alors des améliorations incrémentielles, où chaque fraction de pourcentage supplémentaire nécessite des efforts disproportionnés. Les complexités sont accrues par les sauts de ligne inhérents aux manuscrits médiévaux, où l'absence de césure pose des problèmes. Déterminer précisément si un mot doit être coupé entre deux lignes introduit une ambiguïté considérable. Une solution à ce problème est le post-traitement.
Bien que les méthodes algorithmiques et l'apprentissage automatique aient été utilisés avec succès pour cette tâche, chaque approche présente des limitations. Les méthodes algorithmiques (telles que Pie) sont limitées dans leur adaptabilité, tandis que les méthodologies d'apprentissage automatique exigent des ensembles de données étendus et des ressources computationnelles pour l'entraînement. Compte tenu de ces facteurs, nous proposons une solution alternative offrant des avantages économiques par rapport à l'affinement de modèles comme RoBERTa pour la correction d'erreurs.

La solution proposée consiste à entraîner le transformer sur un vocabulaire assez restreint mais spécifique de trigrammes, que le tokeniseur spécialisé est formé à extraire. Plus précisément :

- Développement d'un tokeniseur spécialisé.

- Entraînement du modèle de transformer : Notre modèle est spécifiquement conçu pour reconnaître et corriger les erreurs couramment trouvées dans les transcriptions automatiques de manuscrits latins médiévaux datant du 13ème au 16ème siècle.
- Analyse et simulation des erreurs : En incorporant des erreurs simulées reflétant les erreurs de transcription courantes, nous facilitons une formation robuste du modèle.

Mots-Clés: Post, traitement, ATR, RoBERTa

*Intervenant