Intégration de modèles seq2seq pour la correction et la normalisation de textes issus de l'HTR

Dans le cadre de la deuxième édition des journées annuelles de Biblissima+, je souhaiterais proposer une communication sur l'utilisation des technologies d'intelligence artificielle, en particulier des modèles d'apprentissage automatique seq2seq, pour la correction des transcriptions normalisées issues de l'HTR des manuscrits médiévaux latins des XIIIe-XVe siècles, notamment des collections des distinctions.

Le projet repose sur l'adaptation du modèle roberta-base-latin-cased en tant qu'encodeur et l'exploitation du modèle T5 multilingue comme décodeur, afin d'optimiser la précision des transcriptions. Il vise à surmonter les défis posés par les variantes textuelles, les abréviations et les spécificités paléographiques des manuscrits médiévaux latins, en utilisant un cadre seq2seq pour améliorer la qualité des données textuelles. La présentation abordera les aspects techniques de la combinaison des modèles seq2seq, avec une attention particulière sur la résolution des problèmes liés à l'incompatibilité des tokenizers de roberta et T5.

Nous discuterons également des techniques d'entraînement et de mise au point (fine-tuning) des modèles sur un corpus spécialement annoté, composé de textes divisés en fragments sémantiques de la taille d'un paragraphe, séparés par des pieds-de-mouche. L'entraînement est effectué sur la vérité de terrain obtenue dans le cadre du projet Distinguo. À ce jour, l'entraînement a été réalisé sur 556 000 tokens, mais nous prévoyons de doubler cette quantité d'ici le début du mois de mars.

Svetlana Yatsyk, Chargée de recherche, CIHAM (UMR 5648), projet DISTINGUO

Svetlana Yatsyk
https://ciham.cnrs.fr/annuaire/membres_statutaires/svetlana-yatsyk/

Formation

2012 – 2015
Doctorat d'histoire médiévale (Candidat des sciences), École des hautes études en sciences économiques (HSE), Russie

2007 – 2012
Spécialiste d'histoire médiévale, Université d'État Lomonossov de Moscou, Faculté d'histoire, Département d'histoire médiévale

Expérience professionnelle

2022 – ajd
CNRS, CIHAM, Projet Distinguo, Chargée de recherche

2021 – 2022
École Normale Supérieure, Projet Translitteræ, Post-doctorante

2020 – 2021
École Nationale des Chartes, Chercheuse invitée

2016 – ajd
Revue « Vox medii aevi », Rédactrice en chef

2013 – 2022
HSE, Centre d'histoire médiévale, Chercheuse

Formation complémentaire

2022
PSL-Week « Intelligence artificielle pour les SHS : initiation à la transcription automatique des documents », PSL, Paris, France

2021–2023
Écoles d'été : European Summer University in Digital Humanities « Culture & Technology », Université de Leipzig, Allemagne
Cours intensifs du TEI/XML, de la stylométrie, de l'analyse textuelle et visualisation des données avec R.

2021
École d'été : « Le livre médiéval au regard des méthodes quantitatives », IRHT, Paris, France

Présentations et Conférences Choises

2023
International Medieval Congress, University of Leeds, Royaume-Uni
Applying Handwritten Text Recognition to the Distinctiones Collections: Building a Broad Model for Fine-Tuning with eScriptorium.

2023
The 2nd International Medieval Sermon Studies Society Doctoral and Postdoctoral Online Forum
A diachronic analysis of circulation of « Breviloquium de virtutibus » by John of Wales.

2022
International Medieval Congress, University of Leeds, Royaume-Uni
From a manual for slow reading to the reference text: on the circulation of « Breviloquium de virtutibus ».