FUTURS FANTASTIQUES
Vendredi 10 décembre 2021

Grand Auditorium

Aurélia Rostaing & Alix Chagué, *LectAuRep (Lecture automatique des répertoires de notaires de Paris) – Archives nationales-Inria-Scripta*

Aurélia Rostaing :

Le projet LectAuRep porte sur les répertoires de notaires parisiens de la période 1803-1940 conservés aux Archives nationales. Un échantillon d'images numériques de ce corpus a été traité par reconnaissance de caractères imprimés et manuscrits et, de manière exploratoire, par traitement automatique des langues, reconnaissance des entités nommées et éditorialisation.

Un répertoire est un registre où le notaire consigne par ordre chronologique les actes notariés qu'il a établis. Les contenus de ces colonnes sont autant de métadonnées relatives aux actes décrits ; on peut les catégoriser en types d'actes, dates, noms d'agents, professions, noms géographiques, mots matière. Les instruments de recherche que sont les répertoires constituent de ce fait des corpus de recherche en eux-mêmes, où il est possible d'isoler des lots homogènes de données intéressant l'histoire économique et sociale.

Notre but est de faciliter et de massifier l'accès à ces contenus en offrant à nos usagers un service de lecture enrichi et de fouille de données.

Nous voulons aussi partager le résultat de nos travaux et nos retours d'expériences afin de favoriser la mutualisation et l'interopérabilité des données et métadonnées produites par l'HTR. Il s'agit pour cela de convenir de bonnes pratiques communes, voire de standard entre GLAM, chercheurs, généalogistes, prestataires de logiciels métier ou de services d'HTR.

Enfin, nous avons tenu à prendre en compte quinze ans d'un patrimoine numérique résultant aussi bien d'une numérisation rétrospective de microfilms, que d'une numérisation d'après originaux.

Concrètement, nous avons échantillonné deux lots d'images en noir et blanc et en couleurs, puis nous avons ouvert deux chantiers plus homogènes sur le plan matériel et, par conséquent, moins difficiles à traiter techniquement : un siècle de registre d'enregistrement de contrats de mariage de commerçants, et les répertoires d'un notaire du XVIIIe siècle.

Sur ces quatre lots, représentant 250 mains au moins, environ deux mille pages ont été transcrites (soit quelques dizaines de mains d'écriture), dont un gros quart a été

relu. Des modèles d'HTR satisfaisants ont pu être affinés sur la base d'une vérité terrain de qualité afin de réduire les taux d'erreur par caractère.

Alix Chagué :

Pour la tâche de transcription, l'environnement technologique dans lequel s'inscrit le projet est essentiel. Il convient de distinguer ce qui relève du *software*, la partie la plus visible, et ce qui relève du *hardware*.

Le projet LectAuRep s'inscrit dans une démarche de science ouverte . Tout naturellement, cela se retrouve dans le choix des logiciels :

- Premièrement, Kraken un moteur d'HTR développé par Benjamin Kiessling depuis 2015, désormais sous l'égide de SCRIPTA PSL. Kraken est compatible avec de nombreux systèmes d'écriture, alphabétiques ou non alphabétiques, et plusieurs sens de lecture. Il permet d'entraîner différents types de modèles : pour la transcriptionn, mais aussi pour la segmentation, c'est-à-dire la détection des lignes et/ou des zones de texte sur l'image.

- Deuxièmement, eScriptorium, une application web développée par SCRIPTA PSL depuis 2018, c'est un plan de travail virtuel pour la conduite de projets de transcription. Ilsert de coquille ou d'interface graphique à des moteurs d'HTR, en l'occurrence ici Kraken.

Le projet LectAuRep s'appuie donc depuis 2019 sur une application eScriptorium, qui est déployée par ALMAnaCH sur ses serveurs, d'abord sur une machine virtuelle d'INRIA, c'est-à-dire un serveur minimaliste avec peu de capacité, dont le but était de permettre à tous les membres du projet de travailler ensemble sur la même base de données. Puis, une première montée en charge en 2020, avec une migration sur le serveur « Traces6 » mieux équipé (notamment doté de cartes graphiques, indispensables pour un entraînement efficace des modèles). Enfin, LectAuRep fait partie des projets qui pourront tirer profit de la nouvelle montée en charge grâce au serveur CREMMA, financé par le DIM MAP : en plus d'être mieux équipé (il possède plus de GPU et plus de mémoire), il propose une architecture modulaire qui est compatible avec de futures améliorations. Sans cet environnement, la tâche de transcription n'est pas possible.

En 2021, les résultats de LectAuRep sont nombreux. Nous avons établi une méthode de production de modèles de transcription qui fonctionne bien : elle s'appuie sur l'utilisation de modèles dits « génériques » dont les taux d'erreur par caractère sont inférieurs à 10% (c'est-à-dire que le modèle fait une erreur pour moins d'une lettre sur 10). Ces modèles sont entraînés sur des lots de mains variés, généralement au moins une dizaine.

On en possède deux, entraînés sur différents ensembles de transcription qui sont plus ou moins parfaites. Ces modèles servent plusieurs objectifs : 1) produire une première passe de transcription qui permet, soit la publication d'une transcription

certes faussée, mais compatible avec une exploration des corpus dans le cadre d'une recherche floue, soit une pré-annotation des documents, ce qui fait gagner du temps lors de la transcription manuelle puisqu'au lieu de déchiffrer, on n'a qu'à corriger. 2) De plus, ces modèles servent de base pour affiner des modèles dits spécialisés, qui sont réentraînés sur des petits lots de données uniformes. De cette manière, on parvient rapidement à des taux d'erreur égaux, voire inférieurs à 5%, soit une faute tous les vingt caractères.

Bien entendu, dans le cadre d'une démarche ouverte, les conventions et pratiques de transcription élaborées sont documentés, de même que les expérimentations avec les données, les modèles et l'infrastructure.

Les données de transcription sont le nerf de la guerre ; elles sont la base pour entraîner des modèles, et LectAuRep en a produit beaucoup, soit en faisant la transcription entièrement à la main, soit en faisant de la reprise de transcription automatique

. Une partie de ces données est considéré comme « *gold* » : c'est-à-dire qu'elles ont été contrôlées et corrigées. Elles sont rendues publiques par l'intermédiaire de l'organisation HTR-United et pourront aussi faire l'objet d'une publication par le biais de data.culture.gouv.fr. Le reste des transcriptions nécessite encore des corrections de la part du DMC et intégrera le corpus *gold* progressivement.

D'autres livrables n'avaient pas été anticipés par le projet. On peut mentionner une contribution directe et continue au projet SCRIPTA PSL sous la forme de cas d'usage de retours d'utilisateurs, sous la forme de développement de fonctionnalités qui sont intégrées au code source de l'application (on mentionnera ici le travail d'Yves Tadjo, dont le contrat est financé par LectAuRep) et, globalement, sous la forme d'une documentation qui est mise à disposition de tous les utilisateurs de l'application.

Plus largement, les membres du projet sont engagés dans une démarche de partage d'expertise avec les utilisateurs porteurs de projets impliquant de l'HTR, avec des groupes de travail comme CREMMALab, et avec la communauté des GLAM.

Ce partage d'expertise prend aussi la forme de publications scientifiques sur les questions qui intéressent le projet, dont une partie fait l'objet de billets de blog sur un carnet hypothèses, ouvert à la faveur du confinement et du stage de Lucas Terriel en 2020.

Témoin d'une appropriation de la question de la mesure des performances des modèles « au contact du terrain », l'outil KaMI rend possible une meilleure évaluation de la réussite des modèles.

Il donne davantage de métriques, en combinant par exemple taux d'erreur par caractère (CER), taux d'erreur par mots (WER), distance de Levenshtein et opérations d'édition comme les substitutions, suppressions et ajouts. KaMI est agnostique : on peut s'en servir pour comparer deux chaînes de caractères, quel que soit le logiciel d'HTR utilisé pour les générer. Et enfin, il permet surtout de jouer avec des filtres afin de négocier la sévérité de l'évaluation en fonction de critères considérés comme importants : on peut par exemple ignorer les erreurs portant sur la reconnaissance des nombres dans un cas où on s'occuperait surtout de savoir si les lettres et les mots sont bien reconnus. Une telle évaluation est très utile pour anticiper la difficulté de la tâche de correction après l'application d'un modèle de transcription.

Le corpus de textes produits à l'occasion du projet LectAuRep fait émerger des défis qui peuvent intéresser les spécialistes du traitement automatique des langues car la langue utilisée dans les pages des répertoires est loin d'être naturelle : elle contient de nombreuses abréviations et entités nommées, et est faite de phrases non verbales.

Enfin, les documents qui intéressent le projet ont permis de s'interroger sur la manière de rendre accessibles des documents dont la mise en page est complexe. LectAuRep alimente une partie des exemples étudiés par l'équipe ALMAnaCH pour intégrer une application comme TEI Publisher dans une chaîne de traitement généraliste dédiée à l'HTR et reposant sur une utilisation plus systématique de la TEI.

Aurélia Rostaing :

Le projet LectAuRep a montré, à l'échelle d'une partie de l'échantillon initial, que des modèles d'HTR à large spectre fonctionnent suffisamment bien pour permettre une recherche floue sur des pages aux lignes d'écriture aérées (le corpus en noir et blanc du XIX^{e} siècle et une partie du corpus en couleurs du XXe siècle).

Les modèles de segmentation se heurtent pour le moment à un seuil quand les lignes sont trop serrées (soit une part importante du corpus du XXe siècle) ; il serait donc utile de disposer d'outils permettant d'évaluer la qualité des modèles de segmentation, dont celle de l'HTR dépend, afin de pouvoir affiner ces modèles en se fondant sur des métriques.

Le corpus cible de LectAuRep (évalué à plus d'un million d'images) est monumental. La mise en production de moins d'un pour cent de ce corpus (par exemple l'une des 122 études notariales, ou bien une tranche chronologique d'une année sur 140) requiert une logistique participative, des infrastructures et une ingénierie de projet, concernant notamment les flux des images et des données.

Pour affiner un échantillonnage, il peut être utile d'approfondir la connaissance diplomatique de nos sources physiques et numériques. Avoir une idée du nombre de

mains par registre, pouvoir préciser la répartition quantitative entre noir et blanc et couleurs grâce à une meilleure maîtrise des métadonnées permettrait peut-être d'optimiser et d'économiser les ressources nécessaires à l'intelligence artificielle.

Merci.