Vers un traitement automatisé des commentaires classiques

Sven Najem-Meyer, Matteo Romanello

Vers un traitement automatisé des commentaires classiques

Sven Najem-Meyer
Ecole Polytechnique Fédérale de Lausanne
DHLAB
sven.najem-meyer@epfl.ch

Matteo Romanello
Université de Lausanne
IASA
matteo.romanello@unil.ch

Résumé

Bien qu'ils comptent parmi les publications les plus emblématiques du domaine, les commentaires se sont montrés particulièrement rétifs à la digitalisation des études classiques. Le projet *Ajax Multi-Commentary* (AjMC) accuse cet état de fait et vise à rendre l'ensemble de commentaires de l'*Ajax* accessibles et navigables au sein d'un « multi-commentaire ». Cet objectif requiert un traitement semi-automatique inédit, capable de transformer un ensemble de pages numérisées en un réseau de textes structurés. Passant de l'image au texte, puis du texte à l'information, nous décrirons ici les difficultés rencontrées et les solutions apportées à chaque étape du traitement.

1 Introduction

Les commentaires classiques comptent parmi les plus anciennes formes d'ouvrages critiques que nous connaissions. Si les éditions abondamment annotées des commentateurs modernes ont peu à peu supplanté les notules marginales des premiers scoliastes, l'exercice a gardé sa teneur originelle : il vise à reconstruire, à expliquer et à analyser un texte grec ou latin. La longévité du genre en fait aussi le témoin d'une pratique, dont de nombreuses études ont analysé l'histoire, les limites et les potentielles transformations à l'ère numérique (Most, 1999; Gibson et Kraus, 2002; Kraus et Stray, 2016).

Bien qu'elles aient suscité l'intérêt croissant du monde académique au cours des deux dernières décennies, ces recherches restent freinées par de nombreuses contraintes techniques. Le manque de métadonnées, la médiocrité des retranscriptions automatiques (OCR) et l'absence de système d'indexation entachent les commentaires numérisés et limitent les analyses historiques de grande échelle.

Accusant cette situation, le projet Ajax Multi-Commentary (AjMC) s'est donné pour but de rendre les commentaires de l'*Ajax* de Sophocle accessibles et navigables dans une interface dynamique, permettant de visualiser simultanément l'ensemble des scolies qui commentent un même extrait. Si ce « multi-commentaire » ouvre la voie à des approches comparatives inédites, il requiert aussi un traitement semi-automatique novateur, capable de transformer un corpus de commentaires numérisés en un réseau de textes structurés.

2 Vers un traitement automatisé des commentaires classiques

Cette section présente les défis rencontrés et les solutions envisagées à chaque étape du traitement (Fig. 1). Après une brève description du corpus, nous passerons d'abord de l'image au texte, à l'aide de l'OCR (2.2) et de l'analyse de la mise en page (AMP) (2.3). Nous passerons ensuite du texte à l'information structurée, à l'aide de la reconnaissance d'entités nommées (NER) (2.4) et de systèmes d'alignement (2.5).

2.1 Collecter les commentaires numérisés : le corpus.

Le projet réunit 17 commentaires de l'*Ajax*, publiés entre 1835 et 2011 en allemand, anglais, français, italien et latin. Au multilinguisme du corpus s'ajoute le multilinguisme des scolies elles-mêmes, qui citent quasi-systématiquement le texte commenté dans sa langue d'origine (grec polyphonique).

2.2 Retranscrire le texte : l'OCR

Cette première étape a pour but de convertir les pages numérisées en données textuelles. Si les modèles état de l'art frôlent la perfection sur des documents nets et monolingues, leurs performances tendent à se détériorer sur des documents historiques ou multilingues. Les résultats d'une étude préliminaire (Romanello et al., 2021) confirment malheureusement cette tendance. Toutefois, nos meilleurs modèles atteignent les seuils

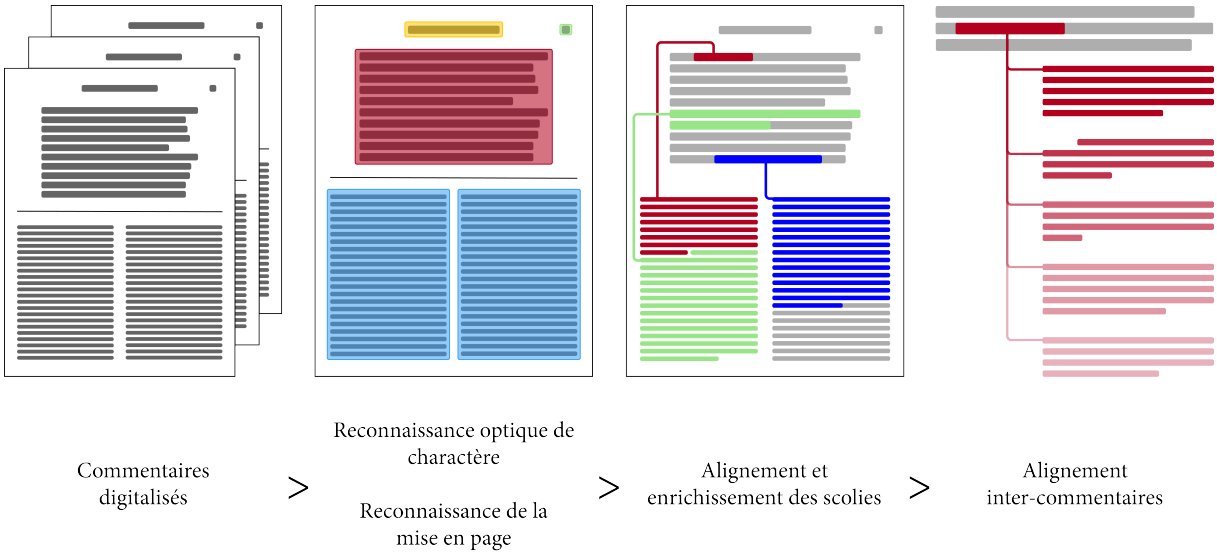

FIGURE 1 – Vue d'ensemble de la chaîne de traitement du projet AjMC.

minimaux recommandés pour la conduite d'analyses textuelles (Hill et Hengchen, 2019; van Strien et al., 2020). Si ce pronostic a pu être vérifié par le succès de la reconnaissance d'entités nommées sur les textes obtenus, la qualité de l'OCR reste déterminante pour l'alignement et la praticabilité du multi-commentaire. Son amélioration fait donc l'objet d'efforts continus.

2.3 Organiser le texte : l'AMP

Cette deuxième étape vise à segmenter une page en régions et à classer ces régions selon leur nature. La difficulté de la tâche naît de la complexité structurelle des commentaires : en-tête, texte, scolies et appareil critique se côtoient souvent sur une même page. Dans le but d'isoler les régions contenant des scolies, nous avons comparé (Najem-Meyer et Romanello, 2022) des modèles utilisant le texte (RoBERTa), l'image (YOLOv5) ou les deux modalités (LayoutLMv3). Nos expériences montrent un net avantage en faveur de l'approche visuelle, même si le modèle peine à généraliser les connaissances apprises lors de son entraînement. Si ces approches ne sont pas encore utilisables en production, elles peuvent déjà venir au soutien de procédures semi-automatiques en accélérant par exemple le travail d'annotation.

2.4 Enrichir le texte : la NER

La reconnaissance et la désambiguïsation d'entités nommées permet d'enrichir et de comparer plus facilement le contenu des scolies. Grâce à la mise à disposition des données d'AjMC lors de la shared-task de *HIPE 2022* (Ehrmann et al., 2022), nous avons pu en évaluer la difficulté. La tâche de reconnaissance vise à détecter et à classifier des entités selon une taxonomie donnée (auteur, oeuvre, personnage mythologique... etc). Elle affiche des résultats très prometteurs. *Fine-tuné* sur des données d'entraînement, un BERT multilingue et historique (Schweter et al., 2022) atteint un F-score de 91.3%, 84.2%, et 85.4% sur les commentaires allemands, français et anglais respectivement [1].

La tâche de désambiguïsation consiste quant à elle à lier les entités ainsi extraites à leurs entrées Wikidata. Les résultats sont cette fois moins encourageants. Les abréviations, très fréquentes dans les commentaires, en sont l'un des enjeux majeurs. Seuls 1.4% des relations correctement prédites par le meilleur modèle proviennent d'entités abrégées, bien que celles-ci constituent 47% du total des entités.

2.5 Superposer les textes : l'alignement

Pour superposer l'ensemble des commentaires et les explorer horizontalement, il est enfin nécessaire d'extraire les ancres textuelles (lemma) par lesquelles les scolies délimitent l'extrait qu'elles commentent. La difficulté de la tâche se mesure à la complexité des ancres. Souvent composées de numéros et de caractères grecs, elles concentrent un grand nombre d'erreur de transcription. Une campagne d'annotation initiée en décembre 2022 devra permettre une première série d'expériences dans le courant 2023.

1. F-score strict, sur la tâche NERC-coarse.

3 Conclusion

La chaîne de traitement que nous présentons vise à transformer semi-automatiquement des commentaires numérisés en textes structurés et enrichis. Il en résulte un « multi-commentaire » qui ouvre la voie à une analyse comparative des commentaires de l'*Ajax* de Sophocle. À plus long terme, les méthodes et les outils développés dans le cadre de ce projet pourront être généralisés à des corpus plus vastes et offrir de nouvelles perspectives de recherche.

Remerciements

Ce travail est soutenu par le Fonds National Suisse (subvention numéro PZ00P1_186033).

Bibliographie

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, et Simon Clematide. 2022. Extended overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, volume 3180. CEUR-WS.

Roy K. Gibson et Chris(tina) Shuttleworth Kraus, éditeurs. 2002. *The Classical Commentary: Histories, Practices, Theory*. Brill.

Mark J Hill et Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4) :825–843.

Christina S. Kraus et Christopher Stray, éditeurs. 2016. *Classical Commentaries: Explorations in a Scholarly Genre*. Oxford University Press.

Glenn W. Most, éditeur. 1999. *Commentaries = Kommentare*. Numéro 4 in Aporemata : Kritische Studien Zur Philologiegeschichte. Vandenhoeck & Ruprecht, Göttingen.

Sven Najem-Meyer et Matteo Romanello. 2022. Page Layout Analysis of Text-heavy Historical Documents : A Comparison of Textual and Visual Approaches. In *Proceedings of the Conference on Computational Humanities Research 2022*, pages 36–54, Antwerp. CEUR-WS.

Matteo Romanello, Najem-Meyer Sven, et Bruce Robertson. 2021. Optical Character Recognition of 19th Century Classical Commentaries: The Current State of Affairs. In *The 6th International Workshop on Historical Document Imaging and Processing (HIP '21)*, Lausanne. Association for Computing Machinery.

Stefan Schweter, Luisa März, Katharina Schmid, et Erion Çano. 2022. hmBERT: Historical Multilingual Language Models for Named Entity Recognition.

Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, et Giovanni Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 484–496, Valletta, Malta. SCITEPRESS - Science and Technology Publications.