Lemmatiser des textes et corriger l'annotation grâce à l'apprentissage profond avec Pyrrha

Thibault Clérice*[1,2], Ariane Pinche[3], Jean-Baptiste Camps*[1], Matthias Gille Levenson[4], Simon Gabay*[5], and Lucence Ing*[1]

[1]Centre Jean Mabillon – Université Paris sciences et lettres, Ecole Nationale des Chartes : EA3624 – France

[2]Histoire et Sources des Mondes antiques – Centre National de la Recherche Scientifique : UMR5189, Université Jean Monnet [Saint-Etienne], Université Jean Moulin - Lyon 3, Université de Lyon, Université Lumière - Lyon 2, École Normale Supérieure - Lyon – France

[3]Histoire, Archéologie et Littératures des mondes chrétiens et musulmans médiévaux – École Normale Supérieure - Lyon, Université Lumière - Lyon 2, École des Hautes Études en Sciences Sociales, Université Jean Moulin - Lyon 3, Université de Lyon, Avignon Université, Centre National de la Recherche Scientifique : UMR5648 – France

[4]Histoire, Archéologie et Littératures des mondes chrétiens et musulmans médiévaux – École Normale Supérieure - Lyon – France

[5]Université de Genève – Suisse

Résumé

L'analyse quantitative ou distante de données textuelles est centrale dans le champ des sciences humaines computationnelles, tout comme l'emploi de méthodes provenant du traitement automatique des langues. La lemmatisation et l'annotation morpho-syntaxique sont notamment utiles à de nombreux traitements très courants (modélisation de sujet, lexicométrie, etc.). Si les langues contemporaines sont la plupart du temps bien équipées, les langues anciennes ou médiévales, riches en variation (morphologique, graphique, etc.), et les langues peu dotées posent des difficultés réelles. La dernière génération de lemmatiseurs permet de remédier efficacement à ces problèmes, mais le passage d'outils à base de règles à des approches par apprentissage profond [1, 2] implique l'utilisation de grandes quantités de données d'entraînement, nécessitant un important travail de reprise et correction de données. Toutefois, la plupart des interfaces pour faire de l'annotation morpho-syntaxique et de la lemmatisation sont construites pour une saisie de première main plutôt que pour de la correction [3, 4]. Pyrrha [5] a été développée en vue de répondre aux pratiques existantes d'utilisateur-trice-s (correction dans des tableurs, par exemple), tout en s'inspirant de fonctionnalités proposées par des applications similaires de corrections d'OCR (corrections par lots [6], validation par dictionnaire, etc.).

Cet atelier présentera aussi l'utilisation de Pie [7]. Pie est un étiqueteur de séquence pour les langues riches en variations, hautement configurable et fondé sur des réseaux de neurones. Pyrrha a été construit pour compléter l'étiqueteur afin de post-corriger le texte balisé. Cela rend possible ensuite de le réutiliser afin de faire croître les données d'entraînements, dans

*Intervenant

un cercle vertueux, ou bien simplement de les analyser.

Dans cette perspective, nous proposons de réaliser un atelier de 4h, abordant les points suivants :

Utilisation du service de lemmatisation et d'annotation (langues actuellement traitées: différents états du français, du XIe au XVIIIe siècles; grec ancien et latin; moyen néerlandais) ;

Création de corpus, imports et exports en différents formats (tsv, XML-TEI) ;

Correction au fil du texte ou par lots ;

Moteur de recherche ;

Bonnes pratiques et travail collaboratif (gestion de référentiels communs, suivi des interventions des différents correcteurs et correctrices, ...) ;

Partage et archivage de données annotées.

Organisation de groupes de travail pour le développement des modèles

Il serait utile de prendre contact avec les instructeurs avant l'atelier afin de s'assurer que les langues préférées des stagiaires sont incluses dans la configuration du cours. La liste actuelle comprend le français classique, moderne, l'ancien français, le latin classique et médiéval, et le grec ancien. L'ajout de langues supplémentaires est envisageable, si prévu suffisamment en avance et si les données existent.

Bibliography

E. Manjavacas, A. Kádár et M. Kestemont, " Improving lemmatization of non-standard languages with joint learning ", arXiv preprint arxiv:1903.06939 (2019).

H. Schmid, " Deep learning-based morphological taggers and lemmatizers for annotating historical texts ", dans Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, 2019, p. 133–137.

S. M. Yimam, I. Gurevych, R. E. de Castilho, and C. Biemann, "WebAnno: A flexible, web-based and visually supported system for distributed annotations," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2013, pp. 1–6.

B. Almas and G. Höflechner, Arethusa: Annotation Environment. alpheios-project, 2018.

Thibault Clérice, Julien Pilla, and Jean-Baptiste-Camps, hipster-philology/pyrrha: 1.0.1. Zenodo, 2018.

T. Erjavec, "Architecture for Editing Complex Digital Documents," in Proceedings of the Conference on Digital Information and Heritage. Zagreb, 2007, pp. 105–114.

Enrique Manjavacas, Mike Kestemont, and Thibault Clérice, emanjavacas/pie v0.1.0. Zenodo, 2018.