Description de la façon dont le doctorant pense, aimerait utiliser l'informatique pour son projet

1er mars 2022

Dans le cadre de ce projet, il est prévu de constituer des données d'entraînement (*vérité terrain*), ainsi que d'entraîner un ou plusieurs modèles de segmentation et/ou de transcription à l'aide de l'interface eScriptorium [Kiessling et al., 2019]. Pour remédier aux particularités des mises en page complexes, caractérisées par la présence des dessins/photographies, des textes en marge, de plusieurs colonnes etc., il s'impose l'utilisation d'un vocabulaire contrôlé pour la description des pages, comme `SegmOnto` [Gabay et al., 2021], afin d'harmoniser les données requises par les analyseurs de mise en page. Ainsi, nous contribuerons à la production et au partage d'un corpus médical en français du XIXe s., à partir des documents imprimés/manuscrits (p. ex. notes de cours de Charcot) qui restent à numériser dans l'optique de notre projet. La correction automatique des transcriptions s'effectuerait grâce à un des outils basés sur les réseaux de neurones (idéalement BERT), comme `neuspell` [Jayanthi, Pruthi et Neubig, 2020].

Les analyses statistiques préliminaires des lexèmes spécifiques (calcul de spécificité de Lafon, cooccurrences etc.), seront effectuées dans le logiciel textométrique TXM [Heiden, 2010]. Ensuite, les visualisations des plongements des mots contextuels BERT faciliteront l'identification des champs lexicaux, des isotopies ou des marqueurs spécifiques (iconographiques, religieux, théâtraux...) de la circulation des idées de Charcot. Nous pourrions opter soit pour l'entraînement du modèle de langage BioBERT [Lee et al., 2020] pour le français, soit pour le réglage fin du modèle CamemBERT [Martin et al., 2019], en vue de l'extraction des concepts médicaux.

Il serait aussi envisageable de modéliser le réseau scientifique de Charcot créé à partir de l'influence de ses pratiques à l'aide de l'analyse des réseaux dans le logiciel comme Gephi [Bastian, Heymann et Jacomy, 2009] ou dans le langage de programmation `R`. Afin d'obtenir une perspective cartographique de ces échanges scientifiques, le réseau en question pourrait être géolocalisé en utilisant une base de données géographiques, puis projeté sur une carte interactive dans QGIS. L'étude des similarités des textes dans l'espace vectoriel, la visualisation de ces connexions ou des changements scientifiques entre auteurs, au fil du temps ou à travers les disciplines avec la librairie `stylo` [Eder, Rybicki et Kestemont, 2016], s'annonce potentiellement très intéressante.

Finalement, le repérage du transfert d'un champ disciplinaire à l'autre, étant littéral ou accompagné de modifications linguistiques (paraphrases, références...), nécessite d'aligner les textes de Charcot avec d'autres corpus, en utilisant les outils TextPAIR et OBVIE [Alrahabi, 2022] mis à disposition par le labex OBVIL/OBTIC.

Références

Alrahabi, Motasem (2022). "Obvie : interface web pour la fouille et la comparaison de textes". In : *Atelier DigitAl Humanities and cuLtural herItAge : data and knowledge management and analysis durant la conférence francophone sur l'Extraction et la Gestion des Connaissances (egc2022)* (cf. p. 1).

Bastian, Mathieu, Sebastien Heymann et Mathieu Jacomy (2009). "Gephi : an open source software for exploring and manipulating networks". In : *Proceedings of the international AAAI conference on web and social media*. T. 3. 1, p. 361-362 (cf. p. 1).

EDER, Maciej, Jan RYBICKI et Mike KESTEMONT (2016). Stylometry with R : a package for computational text analysis. *The R Journal* 8.1 (cf. p. 1).

GABAY, Simon et al. (2021). "SegmOnto : common vocabulary and practices for analysing the layout of manuscripts (and more)". In : *16th International Conference on Document Analysis and Recognition (ICDAR 2021)* (cf. p. 1).

HEIDEN, Serge (2010). "The TXM platform : Building open-source textual analysis software compatible with the TEI encoding scheme". In : *24th Pacific Asia conference on language, information and computation*. T. 2. 3. Institute for Digital Enhancement of Cognitive Development, Waseda University, p. 389-398 (cf. p. 1).

JAYANTHI, Sai Muralidhar, Danish PRUTHI et Graham NEUBIG (2020). Neuspell : A neural spelling correction toolkit. *arXiv preprint arXiv :2010.11085* (cf. p. 1).

KIESSLING, Benjamin et al. (2019). "eScriptorium : An open source platform for historical document analysis". In : *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. T. 2. IEEE, p. 19-19 (cf. p. 1).

LEE, Jinhyuk et al. (2020). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36.4, p. 1234-1240 (cf. p. 1).

MARTIN, Louis et al. (2019). CamemBERT : a tasty French language model. *arXiv preprint arXiv :1911.03894* (cf. p. 1).