Sujet de la thèse :

Méthodes et outils pour l'étude diachronique des discours géographiques dans les encyclopédies françaises.

Contexte

Ce projet de thèse s'inscrit dans le cadre du projet GEODE ("Encyclopedic GEOgraphical DiscoursE: Writing about Geography in France from the Enlightenment to the Age of Wikipedia") financé par le LabEx ASLAN sur la période 2020-2024.
Ce projet interdisciplinaire réunit un consortium de chercheurs en informatique, linguistique, géographie et histoire appartenant aux laboratoires LIRIS, ICAR, EVS, LLF et LIDILEM ainsi qu'à l'Institut Alan Turing (Londres). GEODE s'appuie sur les résultats de projets précédents dans lesquels les différents partenaires ont pu collaborer [8, 9, 13, 14] et vise à en étendre les objectifs scientifiques. L'objectif principal visé est le développement de méthodes pour l'étude des changements majeurs survenus au sein des discours géographiques dans les encyclopédies françaises entre la seconde moitié du XVIII^{e} siècle (Encyclopédie de Diderot et d'Alembert) et nos jours (Wikipedia).

Mots-clés Extraction d'information, traitement automatique du langage, discours géographique, diachronie, encyclopédies

Encadrants

Frédérique Laforest, Professeur en Informatique (Laboratoire LIRIS, INSA Lyon)
Denis Vigier, Maître de conférences HDR en Langue et Linguistique Françaises (Laboratoire ICAR, Université Lumière Lyon 2)
Ludovic Moncla, Maître de conférences en Informatique (Laboratoire LIRIS, INSA Lyon)

Objectifs de la thèse

Le travail de thèse sera décomposé en plusieurs objectifs complémentaires.
Tout d'abord, le/la doctorant-e se concentrera sur la préparation des corpus (homogénéisation des formats, corrections, annotations) afin que le contenu de chaque encyclopédie puisse être traité par des méthodes automatiques. Ensuite, la proposition consistera à développer des algorithmes adaptés pour l'analyse automatique et la recherche d'informations géo-sémantiques et de routines discursives. Le/la doctorant-e s'intéressera en particulier au développement de modèles linguistiques adaptés à l'analyse diachronique du discours géographique. La méthodologie reposera sur la conception d'une chaîne de traitement nécessitant des ressources spécifiques pour le traitement de données géo-historiques (documents annotés, modèles linguistiques, ressources géographiques, etc.). Cette chaîne de traitement fera intervenir des méthodes de classification supervisée ou semi-supervisée pour la classification de textes et le repérage automatique de routines discursives ainsi que des méthodes d'apprentissage profond pour la génération de modèles de langue (tels que les word embeddings). Enfin, une étape du travail consistera également à proposer des méthodes d'interrogation et de visualisation adaptées pour l'analyse et la comparaison des différents corpus.

Une des originalités de cette thèse sera d'articuler approches quantitative et qualitative afin d'éclairer i) les stratégies sélectionnées pour la classification automatique des textes et la génération de modèles de langue ii) l'interprétation des résultats obtenus par ces méthodes.
L'objectif principal de cette thèse sera donc le développement et l'amélioration de méthodes de recherche et d'extraction automatique d'information géographiques pour l'analyse des discours géographiques. Parmi les résultats attendus, on peut citer la mise à disposition des données, ressources, résultats et algorithmes (préparation et correction des corpus, annotations morphosyntaxiques, annotations géo-sémantiques, modèles de langue, ressources géographiques) qui seront produits au cours de la thèse ainsi que la valorisation scientifique des méthodes développées et des résultats obtenus.

Organisation des travaux de recherche

Concernant la préparation et l'homogénéisation des corpus, des travaux devront, par exemple, être conduits afin d'automatiser l'encodage des structures textuelles et péritextuelles ainsi que des métadonnées associées au texte de *La Grande Encyclopédie* (1882-1905). Un accent devra aussi être mis sur l'amélioration des chaînes de traitement PERDIDO [5] et PRESTO [3], tout particulièrement en ce qui concerne le traitement de la variation graphique pour les textes du XVIII^{e} siècle et l'annotation (POS, lemme) des noms propres [4, 12]. Afin de mettre en œuvre et de combiner des approches quantitative et qualitative, ce travail nécessitera des compétences tant en linguistique qu'en informatique pour la phase d'amélioration de l'annotation des POS (morphosyntaxe) dans une perspective diachronique, l'identification et la compréhension des routines discursives ainsi que pour l'implémentation de solutions automatisées.

Le/la doctorant-e devra dès le début de sa thèse, s'employer à entrer dans les textes par une lecture fréquente d'articles des quatre encyclopédies du corpus (*Encyclopédie ou Dictionnaire raisonné des sciences des arts et des métiers* (1751-1772), *La Grande Encyclopédie* (1885-1902), *Encyclopædia Universalis* et *Wikipedia*), afin de s'en forger une connaissance personnelle. Il/elle devra en outre s'attacher à situer ces textes dans le contexte historique, politique et philosophique où ils ont vu le jour, et s'intéresser à leur processus de rédaction (identité des contributeurs, dates de parution des volumes, état de la "discipline" géographique vis-à-vis des autres disciplines à l'époque considérée, etc.). Ces connaissances progressivement élaborées seront réinvesties dans les choix de stratégies et de méthodes pour le traitement informatique des textes et dans l'interprétation des résultats obtenus. De même, une relative connaissance des évolutions propres à la géographie (en tant que pratique puis comme discipline académique) entre le XVIII^{e} siècle et nos jours [10, 2, 11] sera utile pour un traitement éclairé des textes et des résultats.

L'insertion du/de la doctorant-e dans un séminaire autour du discours géographique organisé conjointement par les laboratoires ICAR, LIRIS, et EVS lui permettra à la fois de se former sur ces aspects et de communiquer sur ses résultats.

Profil et compétence recherchées

Diplôme de master ou école d'ingénieur avec des compétences en informatique, traitement automatique du langage (TAL) et analyses de corpus. Des connaissances en intelligence artificielle (machine learning) et humanités numériques seront appréciées.

Conditions

La thèse sera effectuée en partie au laboratoire LIRIS à l'INSA Lyon et au laboratoire ICAR à l'ENS Lyon.
Démarrage de la thèse : février 2021

Candidature

Envoyer votre CV, vos derniers relevés de notes et votre lettre de motivation (en un seul pdf) par mail à frederique.laforest@insa-lyon.fr, denis.vigier@ens-lyon.fr et ludovic.moncla@insa-lyon.fr
Candidatures acceptées jusqu'au 20 décembre 2020

Bibliographie

[1] Barroux, G. et Pépin F. (dir.), *Le Chevalier de Jaucourt. L'homme au dix-sept mille articles*, Société Diderot, collection "L'Atelier, autour de Diderot et de l'Encyclopédie", 2015.

[2] Blais, H.; Laboulais-Lesage, I. *Géographies plurielles*. Les *sciences géographiques* au *moment* de l'*émergence* des *sciences humaines* (*1750-1850*), L'Harmattan, 2006, Histoire des *Sciences Humaines*

[3] Blumenthal, P., Vigier, D. *Du quantitatif au qualitatif en diachronie.* Présentation, Langages 2017/2 (N° 206), p. 5-20, 2017

[4] Diwersy S., Falaise A., Lay, M-H & Souvay G. Ressources et méthodes pour l'analyse diachronique in Blumenthal & Vigier (eds.), "Du quantitatif au qualitatif en diachronie. Prépositions françaises", Langages 206, 21-44, 2017

[5] Gaio, M. and Moncla, L. Geoparsing and geocoding places in a dynamic space context: The case of hiking descriptions. In: Aurnague, M and Stosic, D. (Eds), The semantics of dynamic space in French. Descriptive, experimental and formal studies on motion expressions, John Benjamins, Human Cognitive Processing, 66, pp.353-386, 2019

[6] Laramée, F. D. *La production de l'espace dans l'Encyclopédie. Portraits d'une géographie imaginée*, *Document numérique*, vol. vol. 20, no. 2, pp. 159-177, 2017

[7] Leca-Tsiomis, M. L'Encyclopédie selon Jaucourt, in *Le Chevalier de Jaucourt. L'homme au dix-sept mille articles*, Gilles Barroux et François Pépin (dir.), pp. 43-50, 2015

[8] McDonough, K., Moncla, L. & Van de Camp, M. *Named entity recognition goes to old regime France : geographic text analysis for early modern French corpora*. International Journal of Geographical Information Science (IJGIS), 33 (12), 25 pages, 2019

[9] Moncla, L., McDonough, K., Vigier, D., Joliveau T., & Brenon, A. *Toponym Disambiguation in Historical Documents Using Network Analysis of Qualitative Relationships*. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities, 4 pages, Chicago, IL, USA, November 5th–8th, 2019

[10] Numa *Broc* : *La Géographie des Philosophes, géographes et voyageurs. français au 18e siècle*. Editions Ophrys, Paris, 1975, 600 p.

[11] Peaud, L. (2016), *La géographie, émergence d'un champ scientifique*. France, Prusse et Grande-Bretagne (1780-1860)

[12] Rossari, C., Vigier, D. Enonciation et polyphonie dans le discours encyclopédique. TRANEL. Travaux Neuchâtelois de Linguistique, 69, 123 p. 2019

[13] Vigier, D., Moncla, L., Brenon, A., McDonough, K., & Joliveau T. *Classification des entités nommées dans l'Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers par une société de gens de lettres* (1751-1772). In: 7th Congrès Mondial de Linguistique Française (CMLF), Montpellier, France, July 6th–10th, 2020

[14] Vigier, D., Moncla, L., Joliveau T., McDonough, K., & Brenon, A. (2019), *GeoDISCO: Encyclopedic Geographical Discourse in France from the Enlightenment to Wikipedia*. In: 13th Workshop on Geographical Information Retrieval, Lyon, France, November 28th, 2019