Offre de stage M2 / PFE

IA hybride (neuro-symbolique) pour la recherche d'information géographique

Encadrants

- Ludovic Moncla, LIRIS - INSA Lyon (https://ludovicmoncla.github.io)
- Denis Vigier, ICAR - Université Lumière Lyon 2 (http://www.icar.cnrs.fr/membre/dvigier)

Contexte

Ce stage s'inscrit dans le projet interdisciplinaire GEODE [1] ayant comme thème principal une étude diachronique des discours géographiques au sein des encyclopédies. Il se positionne sur le volet informatique et a pour but la conception de méthodes d'intelligence artificielle innovantes et automatiques pour l'extraction d'information à partir d'articles encyclopédiques. Ce stage financé par le LabEx ASLAN [2] sera réalisé dans l'équipe DM2L (Data Mining et Machine Learning) du laboratoire LIRIS [3] en collaboration avec les membres du projet GEODE (laboratoires ICAR et EVS).

Objectifs du stage

Ce stage a pour objectif principal la conception d'une méthode hybride pour la reconnaissance et la classification d'entités nommées et de relations spatiales. La tâche de reconnaissance d'entités nommées (NER) est une tâche essentielle et très étudiée du TAL (Traitement Automatique du Langage). Elle permet d'extraire différents types d'entités et de structurer de l'information à partir de données non-structurées. Les types d'entités recherchées dépendent de la tâche et du corpus étudié. Dans le cadre du projet GEODE, les entités étudiées inclues les entités nommées (e.g., noms propres de lieux, de personnes, etc.), les entités nommées imbriquées ou étendues (e.g., [[ville du [comté de [Rouergue]]] en [France]]) et les entités spatiales nominales ou termes géographiques (occurrences non associées à un nom propre, ex : ville, village, lac, rivière, etc.). Les relations spatiales inclues les relations topologiques (e.g., proche de, dans, traverse, etc.), les relations d'orientations (e.g., au nord de, etc.) et les expressions de distances (e.g., à 6 lieues de, etc.).

La personne recrutée devra faire un état de l'art des différentes méthodologies de combinaison de méthodes d'IA symbolique et d'IA neuronale pour la conception d'une approche hybride d'extraction automatique d'information (i.e., NER et relations spatiales). Du côté des approches neuronales, on s'intéresse en particulier aux méthodes d'apprentissage supervisé, d'apprentissage profond, d'*active learning* et aux approches utilisant des grands modèles de langues pré-entrainés (LLMs). Du côté des approches symboliques on s'intéresse aux méthodes utilisant des heuristiques, des règles linguistiques, des dictionnaires et des graphes de connaissances. Nous nous intéresserons également à l'étude de la décomposition des tâches de la chaîne de traitement et à la comparaison des performances d'approches hybrides et d'approches « end-to-end ». D'un point de vue applicatif, les résultats du stage pourront venir enrichir la librairie Python PERDIDO [4] pour le geoparsing de textes.

Les expérimentations (entraînement et évaluation) seront menées sur les jeux de données du projet GEODE (tels que des ensembles d'articles de l'Encyclopédie de Diderot et d'Alembert (1751-1772)) ainsi que sur les benchmarks de l'état de l'art.

1. https://geode-project.github.io/
2. https://aslan.universite-lyon.fr
3. https://liris.cnrs.fr
4. https://github.com/ludovicmoncla/perdido/

Candidatures

Des compétences sont attendues en programmation et en science des données (Machine Learning et Deep Learning). Des connaissances en traitement automatique de la langue (TAL) seront appréciées.

Profil recherché Master 2 Informatique

Lieu du stage Laboratoire LIRIS – INSA Lyon, Campus La Doua, Villeurbanne.

Période de stage 5 à 6 mois entre février et juillet 2024

Candidature Envoyer un mail présentant votre parcours et vos motivations, votre CV et vos derniers relevés de notes à : ludovic.moncla@insa-lyon.fr et denis.vigier@univ-lyon2.fr

Références

[Brenon et al., 2022] Brenon, A., Moncla, L., and McDonough, K. (2022). Classifying encyclopedia articles : Comparing machine and deep learning methods and exploring their predictions. *Data & Knowledge Engineering*, 142 :102098.

[Cadorel et al., 2021] Cadorel, L., Blanchi, A., and Tettamanzi, A. G. (2021). Geospatial knowledge in housing advertisements : Capturing and extracting spatial information from text. In *Proceedings of the 11th on Knowledge Capture Conference*, pages 41–48.

[Carbonell et al., 2021] Carbonell, M., Riba, P., Villegas, M., Fornés, A., and Lladós, J. (2021). Named entity recognition and relation extraction with graph neural networks in semi structured documents. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9622–9627. IEEE.

[Hamilton et al., 2022] Hamilton, K., Nayak, A., Božić, B., and Longo, L. (2022). Is neuro-symbolic ai meeting its promises in natural language processing ? a structured review. *Semantic Web*, (Preprint) :1–42.

[McDonough et al., 2019] McDonough, K., Moncla, L., and Van de Camp, M. (2019). Named entity recognition goes to old regime france : geographic text analysis for early modern french corpora. *International Journal of Geographical Information Science*, 33(12) :2498–2522.

[Medad et al., 2020] Medad, A., Gaio, M., Moncla, L., Mustière, S., and Le Nir, Y. (2020). Comparing supervised learning algorithms for spatial nominal entity recognition. *AGILE : GIScience Series*, 1 :15.

[Moncla and Gaio, 2023] Moncla, L. and Gaio, M. (2023). Perdido : Python library for geoparsing and geocoding french texts. In *First International Workshop on Geographic Information Extraction from Texts (GeoExT)*.

[Moncla et al., 2021] Moncla, L., Vigier, D., Mcdonough, K., Brenon, A., and Joliveau, T. (2021). Combinaison d'approches qualitative et quantitative pour le repérage et la classification des entités nommées dans l'encyclopédie de diderot et d'alembert (1751-1772). In *Theoretical linguistics in the light of the interaction of qualitative and quantitative approaches*.

[Wang et al., 2022] Wang, Y., Tong, H., Zhu, Z., and Li, Y. (2022). Nested named entity recognition : a survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6) :1–29.

Master's Internship Offer

Hybrid (neuro-symbolic) AI method for geographical information retrieval

Supervisors

- Ludovic Moncla, LIRIS - INSA Lyon (https://ludovicmoncla.github.io)
- Denis Vigier, ICAR - Université Lumière Lyon 2 (http://www.icar.cnrs.fr/membre/dvigier)

Context

This internship is part of the interdisciplinary project GEODE [5], with its main theme being a diachronic study of geographical discourses within encyclopedias. It falls within the realm of computer science and aims to design innovative and automated artificial intelligence methods for information extraction from encyclopedia articles. This internship, funded by LabEx ASLAN [6], will be conducted within the DM2L (Data Mining and Machine Learning) team at the LIRIS laboratory [7], in collaboration with members of the GEODE project (ICAR and EVS laboratories).

Objectives

The main objective of this internship is the design of a hybrid method for the recognition and classification of named entities and spatial relationships. Named Entity Recognition (NER) is a fundamental and extensively studied task in Natural Language Processing (NLP). It allows for the extraction of various types of entities and the structuring of information from unstructured data. The types of entities sought depend on the task and the corpus being studied. Within the framework of the GEODE project, the entities under investigation include named entities (e.g., proper names of places, persons, etc.), nested or extended named entities (e.g., [[city in the [county of [Rouergue]]] in [France]]), and nominal spatial entities or geographical feature nouns (occurrences not associated with proper names, e.g., city, village, lake, river, etc.). Spatial relationships encompass topological relations (e.g., near, inside, crossing, etc.), directional relations (e.g., north of, etc.), and expressions of distance (e.g., 6 miles from, etc.).

The recruited individual will be responsible for conducting a state-of-the-art review of methodologies that combine symbolic AI and neural AI methods for the design of a hybrid approach to automatic information extraction (i.e., NER and spatial relationships). Regarding neural approaches, particular attention will be given to supervised learning methods, deep learning techniques, active learning, and approaches utilizing large pre-trained language models (LLMs). On the symbolic side, we are interested in methods employing heuristics, linguistic rules, dictionaries, and knowledge graphs. We will also delve into studying the task decomposition within the processing pipeline and comparing the performance of hybrid approaches against "end-to-end" approaches. From an application perspective, the results of this internship can contribute to enriching the Python library PERDIDO [8] for geoparsing text.

The experiments (training and evaluation) will be conducted using the datasets from the GEODE project (such as sets of articles from the Encyclopédie of Diderot and d'Alembert (1751-1772)), as well as state-of-the-art benchmarks.

5. https://geode-project.github.io/
6. https://aslan.universite-lyon.fr
7. https://liris.cnrs.fr
8. https://github.com/ludovicmoncla/perdido/

Applications

Skills in programming and data science (Machine Learning and Deep Learning) are expected. Knowledge in Natural Language Processing (NLP) would be appreciated.

Profile Sought Master 2 Informatique

Internship Location Laboratoire LIRIS – INSA Lyon, Campus La Doua, Villeurbanne.

Internship Duration 5 to 6 months between February and July 2024.

Application Send an email presenting your background and motivation, your CV, and your most recent transcripts to : ludovic.moncla@insa-lyon.fr and denis.vigier@univ-lyon2.fr

Références

[Brenon et al., 2022] Brenon, A., Moncla, L., and McDonough, K. (2022). Classifying encyclopedia articles : Comparing machine and deep learning methods and exploring their predictions. *Data & Knowledge Engineering*, 142 :102098.

[Cadorel et al., 2021] Cadorel, L., Blanchi, A., and Tettamanzi, A. G. (2021). Geospatial knowledge in housing advertisements : Capturing and extracting spatial information from text. In *Proceedings of the 11th on Knowledge Capture Conference*, pages 41–48.

[Carbonell et al., 2021] Carbonell, M., Riba, P., Villegas, M., Fornés, A., and Lladós, J. (2021). Named entity recognition and relation extraction with graph neural networks in semi structured documents. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9622–9627. IEEE.

[Hamilton et al., 2022] Hamilton, K., Nayak, A., Božić, B., and Longo, L. (2022). Is neuro-symbolic ai meeting its promises in natural language processing ? a structured review. *Semantic Web*, (Preprint) :1–42.

[McDonough et al., 2019] McDonough, K., Moncla, L., and Van de Camp, M. (2019). Named entity recognition goes to old regime france : geographic text analysis for early modern french corpora. *International Journal of Geographical Information Science*, 33(12) :2498–2522.

[Medad et al., 2020] Medad, A., Gaio, M., Moncla, L., Mustière, S., and Le Nir, Y. (2020). Comparing supervised learning algorithms for spatial nominal entity recognition. *AGILE : GIScience Series*, 1 :15.

[Moncla and Gaio, 2023] Moncla, L. and Gaio, M. (2023). Perdido : Python library for geoparsing and geocoding french texts. In *First International Workshop on Geographic Information Extraction from Texts (GeoExT)*.

[Moncla et al., 2021] Moncla, L., Vigier, D., Mcdonough, K., Brenon, A., and Joliveau, T. (2021). Combinaison d'approches qualitative et quantitative pour le repérage et la classification des entités nommées dans l'encyclopédie de diderot et d'alembert (1751-1772). In *Theoretical linguistics in the light of the interaction of qualitative and quantitative approaches*.

[Wang et al., 2022] Wang, Y., Tong, H., Zhu, Z., and Li, Y. (2022). Nested named entity recognition : a survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6) :1–29.