Offre de stage (Master/Ingénieur)

Recherche et classification de sous-graphes dans un GNN (Graph Neural Network)
pour la reconnaissance d'entités nommées imbriquées

Laboratoire LIRIS – INSA Lyon

Contexte

Ce projet financé par la FIL (Fédération Informatique de Lyon) s'inscrit dans le cadre d'une collaboration entre les équipes DMD (Data Mining & Decision) du laboratoire ERIC et DM2L (Data Mining & Machine Learning) du laboratoire LIRIS. Nous nous intéressons à la caractérisation du discours selon les modalités spatiales et temporelles avec pour objectif de développer des méthodes capables d'apporter des réponses aux questions suivantes :

- Comment caractériser le discours en général ?
- Comment mesurer et interpréter des évolutions temporelles dans les caractéristiques du discours ?
- Comment spatialiser les résultats obtenus ?

Dans ce contexte, nous proposons deux offres de stage (master/ingénieur) autour de la problématique de la caractérisation du discours et de la recherche d'information spatio-temporelle :

1. Recherche et classification de sous-graphes dans un GNN pour la reconnaissance d'entités nommées imbriquées
2. Résumé extractif à l'aide de réseaux de neurones opérant sur des graphes (GNN).
   - Le descriptif de ce deuxième stage est disponible à l'adresse suivante : https://eric.msh-lse.fr/01-11-22-offre-de-stage-reseaux-de-neurones-operant-sur-des-graphes-pour-le-resume-automatique/

Objectifs du stage

Ce stage a pour objectif principal la conception d'une méthode d'annotation automatique d'entités nommées imbriquées. L'imbrication d'entités nommées présente un défi en Traitement Automatique du Langage (TAL) pour la tâche de Reconnaissance d'Entités Nommées (NER) et se rapproche de la tâche d'analyse syntaxique. Dans ce contexte, les entités nommées peuvent être considérées comme des arbres et non plus comme des séquences d'étiquettes. Conserver l'information de chaque entité imbriquée et englobante nous permet de considérer différents niveaux d'analyse. En effet, en fonction de la tâche considérée, chaque entité peut avoir son importance et fournir des informations cruciales (permettant par exemple d'améliorer la classification des entités identifiées). Dans le cas des entités de lieux, l'imbrication permet également de mettre en évidence certaines relations spatiales (topologiques notamment) entre les différentes entités (ex : [[ville du [comté de [Rouergue]]] en [France]]).

Dans ce travail, nous nous intéresserons en particulier à l'implémentation et l'expérimentation des GNN (Graph Neural Network) pour s'adapter au mieux à la structure hiérarchique des entités imbriquées. Le travail consistera

à développer une solution pour l'identification et la classification de sous-graphes pour la reconnaissance d'entités nommées étendues ou imbriquées (*nested named entities*). Cette tâche doit permettre la prise en compte du contexte d'évocation des entités nommées et a pour objectif de tirer parti de la structure syntaxique et de constructions linguistiques fréquentes modélisées au sein du graphe (et des sous-graphes).

Le ou la stagiaire pourra s'appuyer sur un précédent travail exploratoire mené dans le cadre du projet GEODE autour de la modélisation des articles encyclopédiques sous forme d'un graphe et de l'entraînement d'un GNN pour la reconnaissance d'entités nommées (classification de nœuds). L'objectif sera d'approfondir ce travail et de l'étendre pour la classification de sous-graphes.

Bibliographie

Carbonell, M., Riba, P., Villegas, M., Fornés, A. and Lladós, J. *Named Entity Recognition and Relation Extraction with Graph Neural Networks in Semi Structured Documents*. * 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 9622-9627

Finkel, J.-R., and Manning, C. 2009. *Nested named entity recognition*. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09). Association for Computational Linguistics, USA. pp.141–150.

Hanh, T.T.H., Doucet, A., Sidere, N., Moreno, J.G., Pollak, S. (2021). Named Entity Recognition Architecture Combining Contextual and Global Features. In: Ke, HR., Lee, C.S., Sugiyama, K. (eds) Towards Open and Trustworthy Digital Societies. ICADL 2021. Lecture Notes in Computer Science(), vol 13133. Springer, Cham.

Vigier, D., Moncla, L., Brenon, A., Mcdonough, K., & Joliveau, T. (2020) *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).* 7e Congrès Mondial de Linguistique Française (CMLF), Montpellier, France.

Wang, B., Lu, W., Wang, Y., Jin, H. *A Neural Transition-based Model for Nested Mention Recognition*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018. Brussels, Belgium. pp.1011–1017.

Xia, C., Zhang, C., Yang, T., Li, Y., Du, N,. Wu, X., Fan, W., Ma, F., Yu, P. *Multi-grained Named Entity Recognition*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, Florence, Italy. pp.1430–1440.

Déroulement du stage

Profils recherchés : Master 2 Informatique / Ingénieur
Des compétences sont attendues en programmation, en science des données (Machine Learning et Deep Learning) et en traitement automatique de la langue (TAL).

Rémunération : environ 570€ par mois
Lieu : Laboratoire LIRIS – INSA Lyon, Bâtiment Blaise Pascal, Campus La Doua, Villeurbanne.
Date de début : février/mars 2023
Durée : 5 à 6 mois

Candidature : Envoyer un mail présentant votre parcours et vos motivations ainsi que votre CV à :
ludovic.moncla@insa-lyon.fr