PROJET DE RECHERCHE MASTER

Master Icone 2020

Laboratoire L3i

Utilisation des ontologies et des bases de connaissances dans la désambiguïsation des entités

Chercheur(s) référent(s) :

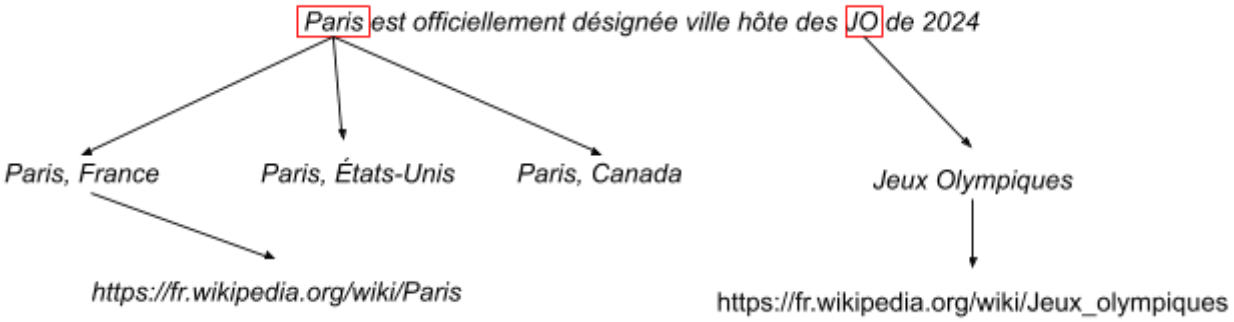
Elvys Linhares Pontes – elvys.linhares_pontes@univ-lr.fr
Luis Adrián Cabrera Diego – luis.cabrera_diego@univ-lr.fr

Contexte scientifique :

Ce projet de recherche se développera au sein des projets européens NewsEye et EMBEDDIA. L'étudiant(e) s'intégrera à l'équipe **« Images et contenus »** dans le thématique **« Documents et Contenus Numériques »** au sein du **Laboratoire L3i**. L'objectif de ce projet est de permettre l'amélioration de systèmes d'annotation sémantique du laboratoire L3i dans les projets NewsEye et EMBEDDIA.

Partie I : Etat de l'art :

L'annotation sémantique, appelé ainsi *Named Entity Linking (NEL)*, vise à reconnaître les mentions dans un document et à les relier à leurs entrées correspondantes dans une base de connaissances (BC), telle que Wikipedia et Freebase. Plus précisément, le NEL vise à localiser les mentions d'une entité nommée (EN) et à les relier avec précision à la bonne entrée d'une BC, un processus qui nécessite souvent d'une désambiguïsation. Par exemple:



Un système NEL effectue généralement deux tâches : la reconnaissance des entités nommées (REN) et la désambiguïsation des entités (DE). La REN extrait les entités dans un document, et l'DE relie ces entités à leurs entrées correspondantes dans une BC. Jusqu'à récemment, l'approche commune des systèmes populaires consistait à résoudre ces deux sous-problèmes indépendamment l'un de l'autre.

La plupart des travaux en DE sont basées sur trois modules : la génération des entités candidates, le classement des entités candidates et la prédiction des mentions non liées [1]. Plus précisément,

le premier module vise à récupérer les mentions d'entités apparentées dans les BCs qui font référence à une mention dans un document. Pour identifier les entités candidates, plusieurs travaux sont fondés sur l'utilisation des dictionnaires de noms [2], sur l'expansion de la forme de surface à partir du document local [3] et sur des méthodes basées sur un moteur de recherche [4]. Après la sélection des entités candidates, le deuxième module tente de déterminer le lien le plus probable dans la BC pour chaque mention. Les principales approches sont principalement basées sur des méthodes supervisées et non supervisées. Ces méthodes prennent en compte diverses techniques d'analyse et de classement des entités, par exemple, la comparaison des chaînes de noms [5], la popularité des entités [2], le type d'entité [6], le contexte textuel [7] et la cohérence entre les entités de la carte [8]. Enfin, ce dernier module valide si l'entité la mieux classée est l'entité cible pour une mention déterminée.

Le module de génération candidates est fondamental dans le processus de désambiguïsation car les modules suivants vont sélectionner le meilleur candidat parmi la liste de candidats générés. Par conséquent, l'objectif de ce projet de recherche est de se focaliser sur le premier module en ordre de mieux sélectionner cette liste de candidats et améliorer le processus de désambiguïsation.

Partie II : Réalisation pratique :

Le projet de recherche est composé de trois étapes :

1) Analyse de l'état de l'art et sélection des travaux les plus intéressants et pertinents.
2) Implémentation d'un algorithme pour récupérer les pages de Wikipédia/Wikidata liées à une mention.
3) Application de l'algorithme sur une collection de test et évaluer sa performance, ainsi que les point forts et faibles.

L'algorithme, ainsi que les résultats, peuvent donner lieu à la rédaction d'un article scientifique à la fin du stage.

Références :

[1] Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering 27(2), 443–460 (Feb 2015). https://doi.org/10.1109/TKDE.2014.2327028

[2] Guo, S., Chang, M.W., Kiciman, E.: To link or not to link? a study on end-to-end tweet entity linking. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1020–1030. Association for Computational Linguistics, Atlanta, Georgia (Jun 2013), https://www.aclweb.org/anthology/N13-1122

[3] Zhang, W., Sim, Y.C., Su, J., Tan, C.L.: Entity linking with effective acronym expansion, instance selection and topic modeling. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three. pp. 1909–1914. IJCAI'11, AAAI Press (2011). https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-319,

[4] Han, X., Zhao, J.: Nlpr kbp in tac 2009 kbp track: A two-stage method to entity linking. In: In Proceedings of Test Analysis Conference 2009 (TAC 09). MIT Press (1999)

[5] Zheng, Z., Li, F., Huang, M., Zhu, X.: Learning to link entities with knowledge base. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 483–491. HLT '10, Association for Computational

Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1857999.1858071

[6] Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 277–285. COLING '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1873781.1873813

[7] Li, Y., Wang, C., Han, F., Han, J., Roth, D., Yan, X.: Mining evidences for named entity disambiguation. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1070–1078. KDD '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2487575.2487681,

[8] Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). pp. 708–716. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), https://www.aclweb.org/anthology/D07-1074