Découverte de relations insolites sur le patrimoine

Fabien Duchateau, Franck Favetta (prénom.nom@liris.cnrs.fr)

Contexte. De nombreuses connaissances sont actuellement stockées dans des documents textuels, ce qui les rend difficilement exploitables directement par les machines [1]. L'extraction d'information est un domaine de recherche visant à identifier, à partir d'un texte, des entités et leur(s) relation(s). Par exemple, la phrase *"Lyon est le chef-lieu de la région AURA"* permet d'extraire l'information *(Lyon, capitale, Auvergne-Rhône-Alpes)*. Plusieurs bases de connaissances ont été proposées comme YAGO, et différents outils d'extraction sont évalués sur des jeux de données.

Cependant, ces outils et bases de connaissances se concentrent sur les types de relations fréquents (*né·e à*, *capitale de*, *auteur·e de*, etc.) [2]. En effet, les systèmes fermés se basent sur une liste prédéfinie et donc limitée de types, tandis que les systèmes plus ouverts nécessitent un minimum de support (e.g., au moins K exemples de relations pour un type) afin de valider la pertinence de ce nouveau type de relation. Il est donc plus difficile de détecter des types de relation insolites. Par exemple, la basilique de Fourvière a inspiré l'architecture de l'église Notre Dame des Victoires de San Francisco, ou le quai de la pêcherie à Lyon a servi successivement de halle aux poissons, port de pêche puis sentier de promenade et marché aux livres. Dans ce projet, nous nous intéressons à ces relations insolites, avec pour domaine d'application le patrimoine local.

Objectifs. Ce projet cherche à atteindre les objectifs suivants :

- Établir une liste de types de relation (plus ou moins) insolites avec des exemples (e.g., en utilisant les sources de données patrimoine Lyon et patrimoine Auvergne Rhône-Alpes), les classifier et construire un jeu de données expertisé.
- Tester et comparer des outils d'extraction d'information (e.g., Ollie, ReVerb, Stanford OpenIE) sur le jeu de données produit.
- Rédiger un rapport décrivant les limites des outils actuels et les défis à relever.

Bibliographie

[1] X. Han, T. Gao, Y. Lin, H. Peng, Y. Yang, C. Xiao, Z. Liu, P. Li, M. Sun, and J. Zhou. More data, more relations, more context and more openness: A review and outlook for relation extraction. *JCNLP*, 2020. Lien JCNLP2020.

[2] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh. A survey on open information extraction. *Computational Linguistics*, 2018. Lien CL2018.