Lighthouse: focused search of entities and relations on the Web

Oana Balalau[1], Ioana Manolescu[1], and Fabian Suchanek[2]

[1]Inria Saclay, firstname.lastname@inria.fr
[2]Télécom Paris, firstname@lastname.name

Keywords: information retrieval, machine learning
Team: CEDAR team, Inria SIF and LIX (CNRS UMR 7161 and Ecole polytechnique); DIG team, Télécom Paris.

Background: Digital data is produced and shared at break-neck speed today – whether as text (news articles), as semi-structured data (tweets, other social media content) or as structured data (RDF or CSV files). Real-world events such as elections, public demonstrations, disclosures of illegal or surprising activities, etc. are mirrored in new data items being created and added to the global corpus of available information. Making sense of this wealth of data and being able to use them in an *integrated* fashion is a highly prized goal.

In collaboration with Le Monde, the team CEDAR has collected public data about French politicians and has organized this information as a heterogeneous graph, where nodes represent entities, i.e. person, location or organization, and edges represent connections between these entities. Such connections can be explored using *keyword search*. Specifically, given a set of $k \geq 2$ search terms such as, e.g., "Assemblée Nationale" and "Russia", we can find all the paths that lead from a node matching the former term to another that matches the second term. This enables finding that "Assemblée Nationale" and "Russia" are connected through the contract that the wife of a member of the Assemblée has with a Russian state-owned company [1].

Internship Goal: Potentially useful entities or useful relations between entities may be absent from our dataset. The goal is to enrich our heterogeneous graph with more nodes and links, by tapping into data-sources which might be unreliable, such as websites retrieved via an Internet search. A potential approach requires two main steps: *i*) predicting or proposing entities and relations that should be added to the graph and *ii*) verifying if the prediction is correct via a focused search for external information . First, from a triple (e_1, r, e_2), we should either predict a missing entity e_1 or e_2 or a missing relation r, using for example a graph embedding approach [3]. Second, a predicted triple can be seen as a claim, i.e. "Caroline Abadie is a member of the French National Assembly", and to verify the correctness of the prediction we can use external information [2, 4].

Practical information. The internship will take place in the Inria team CEDAR, at Inria Saclay. A successful internship may provide opportunities for a funded Ph.D. on a follow-up subject.

References

[1] C. Chanial, R. Dziri, H. Galhardas, J. Leblay, M.-H. Le Nguyen, and I. Manolescu. ConnectionLens: Finding Connections Across Heterogeneous Data Sources. *PVLDB*, 11:4, 2018.

[2] F. Li, X. L. Dong, A. Langen, and Y. Li. Knowledge verification for long-tail verticals. *PVLDB*, 10(11), Aug. 2017.

[3] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[4] K. Popat, S. Mukherjee, A. Yates, and G. Weikum. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, 2018.