Internship - Question Answering With Open Knowledge Bases

Julien Romero - julien.romero@telecom-sudparis.eu

Oana Balalau - oana.balalau@inria.fr

Description Given a text, it is possible to extract from it knowledge in the form of subject-predicate-object triples, where all components of the triples can be found in the text. This is called Open Information Extraction (OpenIE). For example, from the sentence "The fish swims happily in the ocean", we can extract the triple (fish, swims, in the ocean). By gathering many of these statements, we obtain an Open Knowledge Base (OpenKB), with no constraints on the subjects, the predicates, and the objects.

Then, this OpenKB could be used for question answering (QA). There have been many approaches that target QA over non-open KBs. These approaches vary from crafting query templates that, once filled in, will be used to query the KB [4], to neural models, where the goal is to represent the question and the possible answers as latent vectors, where the correct answer should be close in the embedding space to the question [1]. In this project, we will focus on neural models, particularly knowledge graph embeddings, i.e., continuous representations for the entities and relations that can generally capture relevant information about the graph's structure.

The current way KB embeddings are computed raises two main challenges:

- Each entity and relation must be seen enough times during training so the system can learn relevant embeddings. The training is done taking edges information into account, so the entity or relation must be part of a sufficiently large number of edges.
- The textual representation of the verbal and noun phrases of the relations, subjects, and objects should be considered.

For example, a recent approach, MHGRN [2], computes embeddings by using a modified graph neural network architecture. This architecture, however, does not take into account the textual representation of relations. A better approach is CARE [3], that relies on two main ideas. First, it clusters the subjects and objects and creates an unlabelled edge between entities in the same cluster. That partially reduces the problem of the entities connected to a small number of edges, by leveraging the connection with better connected entities. Then, it computes embeddings for the relations using GLOVE (word embeddings) and GRUs (recurrent neural networks). We believe that the approach in CARE could be improved by considering more modern neural architectures using message-passing algorithms and integrating the textual representation of predicates, objects, and subjects. In addition, we will investigate if the clustering step is necessary, as it can bring a bias for one important downstream application of KB embeddings: canonicalization, the task of finding a representative for a set of nodes or edges.

In this project, we will improve open KB embedding methods by:

- Exploring state-of-the-art neural architectures and language models.

- Integrating textual representations of the subject, predicate, and object.
- Investigating if clustering before embedding computation is necessary.
- Integrating embeddings into question-answering models.

Planning The intern will start with a study of the state-of-the-art methods for OpenIE. First, they will get familiar with the traditional datasets and the primary baselines. Then, they will implement our new models and compare them with the previous works.

Prerequisites The intern should be involved in a master's program and have a good knowledge of machine learning, deep learning, natural language processing, and graphs. A good understanding of Python and the standard libraries used in data science (scikit-learn, PyTorch, pandas, transformers) is also expected. In addition, a previous experience with graph neural networks would be appreciated.

Work Environment The internship will take place at Telecom SudParis at Palaiseau and will be a collaboration with INRIA Saclay. The intern will join the computer science department. The internship is paid and will last six months.

If you are interested, please send us your resume, a transcript of your grades, and a cover letter (in French or English).

References

[1] Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 615–620, 2014.

[2] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. Scalable multi-hop relational reasoning for knowledge-aware question answering. *arXiv preprint arXiv:2005.00646*, 2020.

[3] Swapnil Gupta, Sreyash Kenkre, and Partha Talukdar. Care: Open knowledge graph embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 378–388, 2019.

[4] Christina Unger, Lorenz Bühmann, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano. Template-based question answering over RDF data. In *Proceedings of the 21st international conference on World Wide Web*, pages 639–648, 2012.