Internship - Open Information Extraction With External Knowledge

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

Oana Balalau - oana.balalau@inria.fr

1 Description

Open Information Extraction (OpenIE) is the task of extracting subject-predicate-object triples from a text. For example, in the sentence "Barack Obama was born on August 4, 1961 and served as the 44th president of the United States", we could extract the triples *(Barack Obama, was born on, August 4, 1961)* and *(Barack Obama, served as, the 44th president of the United States).* This extraction requires understanding the sentence and how grammar works. It is generally done using deep language models such as BERT or GPT.

In this project, we will try to improve the existing systems by exploring two new directions. The first one consists of adding external knowledge during the extraction. For example, knowing that Barack Obama is a politician might give us more confidence in the extraction *(Barack Obama, served as, the 44th president of the United States).* To do so, we will have to select facts from an existing knowledge base and then encode this knowledge using a graph neural network.

The second direction consists in adding constraints for the extraction. By default, the system can choose anything as subject, predicate, and object. By constraining the extractions with rules such as "the words must be continuous" or "the subject must be a proper noun", we hope we can help the system to learn better extractions.

2 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.

3 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.

4 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).