Building a French pipeline for information extraction and automated knowledge graph construction

Research intern position at Inria on information extraction and automated knowledge graph construction for French

Place of work: Inria's center in Rocquencourt (Paris area)

Duration: 6 months

Starting date: Anytime in 2022

Keywords: artificial intelligence, natural language processing, information extraction, knowledge graph, French language

Context

This internship fits within the roadmap activities of Inria's Defense & Security Department.

Analysts in geopolitical crises are employed by the French Ministry of Armed Forces to better identify emerging or ongoing conflicts throughout the world. These analysts are typically overwhelmed by continuous streams of plain-text information. The goal is to structure that information, so that it can be manipulated as graphs and therefore better formalized, cross-referenced and corroborated; such form would in turn enable more advanced visualizations such as automatically generating reports or various indicators on escalating tensions, with the perspective of better anticipation.

The field of Natural Language Processing (NLP) offers numerous tools and algorithms for information extraction, but they face several limitations. First, many of those are disparate isolated tools, with few comprehensive and consistent pipelines. While the first steps of information structuring (extraction) are extensively studied, fewer works reach the deeper stage of knowledge graph construction. And when they do, they are often developed for English only, whereas here the information stream would be in French.

Inria's Defense & Security Department develops and maintains a serious game platform which can simulate the activity of a crisis monitoring cell. Within that platform there will be the opportunity to experiment with the NLP tools developed by the intern, in order to provide the intern with practical feedback from players.

The intern will be supervised by Dr Lauriane Aufrant, who is the lead NLP researcher within Inria's Defense & Security Department.

Candidate profile

- Pursuing a master's degree in Natural Language Processing, Computational Linguistics or Computer Science with a specialization in Machine Learning

- Theoretical and practical knowledge of deep learning, as well as traditional machine learning and knowledge-driven AI

- Strong programming skills (at least Python, git, Linux environment, command line and scripting)

- Fluency in English. Knowledge or interest for the French language. Knowledge of a second foreign language would be appreciated.

How to apply

Send a CV and a cover letter to lauriane.aufrant@inria.fr and frederique.segond@inria.fr

Indications of referees or reference letters would be appreciated but are not mandatory.

Description

Building a knowledge graph from text involves a number of diverse NLP tasks, such as named entity recognition, named entity disambiguation, coreference resolution, open relation extraction, relation clustering, document-level event extraction, slot filling, etc.

The intern's work will touch upon the whole panel of tasks, but in varying depth. Considering the large amount of open source code releases in NLP research, priority is set on leveraging existing code and models. When French models are not readily available, open source code will need to be retrained on French corpora. And in some cases, it will be necessary to re-implement an algorithm from its published paper.

While deep learning approaches will be ubiquitous in that work, the intern will need to remain open to alternate solutions, as the large-scale datasets required by deep learning will not be available in French for all these tasks.

The first research focus will be to study the best combination scheme for these various tasks. For instance, relation clustering can inform named entity disambiguation by providing more comprehensive and consistent information on the named entities to disambiguate; and named entity disambiguation can inform relation clustering by enabling access to structured information on the arguments of those relations. To leverage such interactions within the pipeline, several approaches are possible: run one before the other, or vice-versa, iterate between both, or setup joint predictions. These choices will be made based on both theoretical and empirical analyses.

The second research focus will be a fine-grained evaluation of the performance all over the pipeline, in order to identify where in the pipeline the information is lost the most, and how errors propagate throughout the pipeline. A thorough analysis will lead to identify where to put the most research efforts in the future to improve qualitative and quantitative performance. Depending on the outcomes of that analysis, the opportunity to pursue with a PhD position will be discussed with the intern.