Towards a unified NLP framework for pipelining information extraction from raw text to knowledge graphs

Research intern position (with the goal of pursuing in PhD) at Inria on information extraction and automated knowledge graph construction for French

Place of work: Inria center in Paris area (Rocquencourt / Saclay / Paris)

Duration: 6 months internship + 3 years PhD

Starting date: Anytime in 2023

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.

Taking as example the current situation in Ukraine, one practical application can be to get an overview of where in Ukraine there are Russian tanks at present (and how many of them), based on reported sightings posted by locals on Twitter. Another example is the cross-referencing of live reports from online newspapers, to identify which transport infrastructures (e.g. train stations, bridges…) have been damaged all over the country, and thereby estimate remaining options for evacuation of civilians.

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. PhD supervision will be done jointly with Dr Frédérique Segond (Inria's Defense & Security Director) or with a researcher from another Inria team, depending on the exact chosen PhD topic (to be discussed, see below).

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 and frederique.segond (both at inria.fr)

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

Internship description

Building a knowledge graph from text involves a number of diverse NLP tasks, such as named entity recognition, named entity disambiguation (aka entity linking), 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.

PhD follow-up

The proposed internship is meant to serve as a proof of concept for a broader project on building a unified framework for information extraction. The goal is to implement a framework that is flexible

enough to integrate and evaluate any method from the state of the art (with the prospect of becoming a new standard for the community), but also experiment more deeply with pipeline design, considering innovative combination schemes or extra preprocessing (e.g. parsing, to enable syntax-aware models).

It is therefore proposed to pursue the internship with a PhD, whose exact topic will be written in coordination with the intern, to fit their primary interests within that broad objective.

In any case, the PhD topic will include at least working on the framework itself (which will include extensive survey work to build an accurate view of the diversity of existing approaches, and their commonalities), and extensions of particular interest for the team are:

- to pursue the work on the combination scheme, with the twofold goal of limiting error propagation and increasing the amount of available information in inputs,
- to propose new task-specific models that better leverage the existing third-party information produced along the pipeline,
- to work on algorithmic methods to speed up the building of a pipeline under that framework, for a new language that does not have as many datasets and existing models as English (including, but not limited to, transfer learning approaches),
- to extend the framework with domain adaptation capabilities, to facilitate the application of the framework to a new domain (for a language in which a full pipeline already exists),
- to focus on some specific tasks to improve in the French pipeline, according to the bottleneck analysis produced at the end of the internship.

Since the PhD application processes are early in the year (February-April), the intern will be asked to commit early to that PhD follow-up, possibly even before the internship begins, and to be ready to devote some time for writing the application over that period.