From Texts to Knowledge
6-month internship @ CEA List

Internship context

Based in Saclay (Essonne), the LIST is one of the two institutes of CEA Tech, the Technological Research Division of the CEA. Dedicated to intelligent digital systems, its mission is to carry out technological developments of excellence on behalf of industrial partners, in order to create value.

Within the LIST, the Laboratory of Textual and Visual Semantic Analysis (LASTI) conducts its research in the field of natural language processing and computer vision to extract, classify and generate information. The laboratory's research themes include learning with few data, trustworthiness and multimodality.

Missions

In real-world applications, companies may not always have access to a predefined target knowledge base tailored to their analytics needs. Automatically building and enriching a knowledge base (or database) from texts is feasible, but remains a challenging open problem. One of the main challenges is to link diverse entity mentions across different contexts and documents, and to consistently store the entities and their properties extracted from texts in a database. Addressing the problem involves coreference resolution to identify all mentions in documents and corpora that refer to the same entity. This internship may also be connected with entity linking task, which aims to disambiguate these mentions by linking them to a knowledge base.

The purpose of this internship is to study the principles, approaches, and methods of coreference resolution and entity linking. The intern will examine automatically extracted mentions from technical reports (in the building sector) or from public datasets, select a few promising methods (including Large Language Models or more frugal models) for disambiguating and linking them, then design strategies for automatically building and populating a database.

After an initial phase dedicated to bibliographic research, the objectives of the internship will be to:

- Prepare and/or create public and/or industrial datasets for evaluation and/or training on linking entity mentions within the same document and across a document corpus.
- Design and implement pipelines for automatic coreference resolution, validation, and comparison of methods.
- Build and populate a knowledge base from the extracted textual data.
- Evaluate the developed approaches on the identified datasets. The quality and consistency of both the database and the coreference resolution will be evaluated throughout the internship.

This internship provides the opportunity to contribute to a research area at the intersection of natural language processing, knowledge representation, and information extraction, with potential applications in both academia and industry.

Qualifications

- Students in their 4th or 5th year of studies (M1, M2 or gap year)
- Knowledge in Natural Language Processing
- Machine learning skills (deep learning, perception models, generative AI…)
- Python and Linux proficiency
- Basic knowledge of databases

Job-related benefits

Joining the CEA List and the LASTI as an intern means:

- Working in one of the most innovative research organizations in the world, addressing societal challenges to build the world of tomorrow.
- Discovering a rich ecosystem: privileged connections between the industrial and academic sectors.
- Conducting research autonomously and creatively: encouragement to disseminate and showcase results (scientific articles, patents, open-source codes...).
- Benefit from an internal computing infrastructure with more than 300 state-of-the-art GPUs
- Receive a stipend between €1300 and €1400 per month
- Have the opportunity to continue with a PhD or as a research engineer after the internship.
- Have the possibility of remote work, receive a 75% reimbursement on public transportation costs, and benefit from the “mobili-jeune” aid to reduce rent costs…

To apply, please send your CV, a cover letter, and the title of the internship to:
lastirecrute@cea.fr
If you are interested in more than one internship, please indicate your order of preference.