Dynamic control of LLM for data augmentation in Information Extraction

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

The development of generative AI models based on large generative language models (LLMs) has led to significant advances in zero or few-shot approaches for a range of natural language processing tasks. However, directly using these models to perform target tasks is not always the best solution, whether from the perspective of performance, speed, or computational cost, and by extension, environmental impact. Using LLMs to annotate corpora that are then used to pretrain smaller zero or few-shot models, such as encoder models, frequently provides an interesting alternative.

The internship we propose takes place in this context, focusing on the automatic annotation of corpora by LLMs for Information Extraction tasks (named entity recognition and extraction of relations between these entities). More specifically, this internship will focus on improving text generation approaches to produce synthetic annotations, particularly by defining dynamic control methods to constrain this generation. Therefore, the intern will work on the following main tasks:

- Conduct a literature review on methods for controlled text generation and synthetic data generation;
- Define criteria for controlling the generation of synthetic annotations for Information Extraction tasks;
- Define methods for taking into account these criteria dynamically in the text generation process, including Monte Carlo search tree methods, the exploitation of attention maps or the structure of the context of the LLM;
- Conduct experiments and performance evaluations concerning the capacities of the defined models to follow the generation constraints;
- Conduct experiments and performance evaluations concerning the interest of the generated annotations for the pretraining of Information Extraction models.

This internship may be seen as an introduction to research. It may lead to the publication of a scientific paper if the results are convincing. Successful outcomes may also lead to the extension of this work through a PhD thesis, within the broader scope of the LASTI team's AI research.

Qualifications

- Students in their 4th or 5th year of studies (M2 or last year of engineering school)
- Natural Language Processing, especially text generation and information extraction
- Machine learning skills (deep learning, neural language models, generative AI…)
- Python proficiency in a deep learning framework (especially PyTorch or Tensorflow)

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 promote 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.