M2 Internship

Biomedical named entity normalization through contextual learning using LLMs

Named entity normalization (NEN) is a key step in biomedical information extraction. Its goal is to link an identified textual mention (e.g., a gene or protein mention) to a concept in a knowledge base (KB) (e.g., UniProt or NCBI Gene). NEN is challenging due to high terminological variation: the same biomedical entity can be referred to using many surface forms (e.g., abbreviations, synonyms, orthographic variants), which often differ greatly from the labels found in KBs. In addition, there may be further difficulties depending on the type of mentions and the writing style of the texts being analyzed. For example, the normalization of gene or protein mentions is particularly challenging due to the ambiguity of short abbreviations (e.g., identical abbreviations used for different species), typographic variations that carry biological meaning (e.g., capitalization to distinguish human versus animal genes), as well as the fact that the same surface form may refer interchangeably to a gene or to the protein it encodes. In all cases, named entity normalization aims precisely to ensure the unique and consistent identification of the concepts mentioned in texts, thereby guaranteeing the reliable integration and comparison of data from heterogeneous sources.

The classical approaches consist of fine-tuning a language model on annotated examples, then measuring the similarity between the representation of a mention to be normalized and the representations of the concepts [Sung et al., 2020]. These methods yield good results but often require large training datasets and domain-dependent tuning. By contrast, recent in-context learning (ICL) approaches directly leverage the reasoning capabilities and prior knowledge of large language models (LLMs) by expressing the task in natural language within a prompt, which makes it possible to use them without training or fine-tuning the model.

However, the application of in-context learning (ICL) to biomedical entity normalization remains largely underexplored. This is primarily due to a major challenge: concept descriptions and their large numbers within knowledge bases are often extensive, frequently exceeding model token limits and thereby hindering their inclusion in a single prompt.

The internship directions:

- To reproduce and analyze the results of the article by Luo et al. (2025) on the contribution of context for the normalization of biomedical entities;
- To experiment with different context-based learning strategies for entity normalization, particularly for gene and protein mentions;
- To propose and evaluate strategies to overcome the token limitations of LLMs, for example, via high-recall candidate filtering methods;
- To explore approaches aimed at reducing hallucinations in LLMs, in particular through the integration of Retrieval-Augmented Generation (RAG) methods;

- to evaluate the performance of the proposed approaches on benchmark datasets in the biomedical domain, notably BioCreative VI Bio-ID and JNLPBA.
- The results of this work may contribute to a better understanding of the potential uses of LLMs for biomedical entity normalization and may lead to the preparation of a poster or a scientific paper.

Skills:

– Programming skills, particularly in Python.
– Knowledge of natural language processing and machine learning.
– Interest in large language models (LLMs) and prompting methods.
– An interest in bioinformatics or biomedical applications would be a plus.
– Ability to read, analyze, and write scientific literature.
– Strong interest in research.

Information:

– Location: primarily at LISN (CNRS & Université Paris-Saclay) in Gif-sur-Yvette/Orsay, and at MaIAGE (INRAE & Université Paris-Saclay) in Jouy-en-Josas.
– Duration: 6 months.
– Period: starting no earlier than mid-March 2026.
– Compensation: according to the official stipend scale (approximately €660 per month).

Advisors:

Nona Naderi (LISN, CNRS, Université Paris-Saclay)
Louise Deléger (MaIAGE, INRAE, Université Paris-Saclay)
Arnaud Ferré (MaIAGE, INRAE, Université Paris-Saclay)

How to apply:

Please send a CV and a short cover letter to the contact persons.

References:

- Sung, M., Jeon, H., Lee, J., & Kang, J. (2020). Biomedical entity representations with synonym marginalization. Association for Computational Linguistics (ACL).
- Luo, G., Shi, N., Wang, G., & Tang, B. (2025). Contextual information contributes to biomedical named entity normalization. Journal of Biomedical Informatics
- Pérez-Pérez, M., ... & Krallinger, M. (2017). Evaluation of chemical and gene/protein entity recognition systems at BioCreative V. 5: the CEMP and GPRO patents tracks.
- Huang, M. S., Lai, P. T., Tsai, R. T. H., & Hsu, W. L. (2019). Revised JNLPBA corpus: a revised version of biomedical ner corpus for relation extraction task. arXiv:1901.10219