# Stage de master / Graduate internship
# Does translation lead to better recognition of biomedical named entities from original French Electronic Health Records ?

**Keywords**: natural language processing (NLP), named entity recognition (NER), biomedical concepts
**Ideal starting date**: March/April 2022
**Duration**: 4-6 months
**Advisors**: Xavier Tannier (LIMICS, Professor) xavier.tannier@sorbonne-universite.fr, Fabrice Carrat (IPLESP, Professor), Christel Gérardin (IPLESP, MD, PhD student)
christel.ducroz-gerardin@iplesp.upmc.fr
**Location**: LIMICS/ IPLESP

## Context (General presentation of the topic)

Named-entity recognition (NER) is an important step in Natural Language Processing (NLP), especially for processing specialized textual documents such as medical reports in order to extract key information. Major improvements have been made in this area, especially in English since a very large amount of data is accessible.
Modern NLP makes extensive use of pre-trained language models, which allow efficient semantic representation of texts. The development of algorithms such as transformers [Vaswani2017,Devlin2018] has allowed significant improvements in this field, and these algorithms are now used in a vast range of NLP applications: question-answering, neural machine translation, named entity recognition and sequence classification.
Transformers language models need to be trained on a very large amount of data (in the biomedical field, this led to models such as *BioBERT* [Lee2020], *ClinicalBERT* [Huang2019]), enabling significant improvement in medical information retrieval. Two main types of data are available in the biomedical field to train any language model: public articles (e.g. Pubmed) and clinical Electronic Health Records databases (e.g. MIMIC III). In many languages other than English, efforts still need to be made to obtain such interesting results, in particular due to a much smaller amount of accessible data [Neveol2018].
At the same time, machine translation has also gained in performance thanks to the same type of language models based on transformers, and the last few years have seen the emergence of high-quality automatic translation.
These last two observations led several research teams to add a translation step in order to analyze medical texts, for instance to extract relevant mentions in ultrasonography reports [Campos2017,Suarez2021] or to perform medical concept normalization [Wajsbürt2021].

## Objective of the internship

In this work, we want to address the question of adding an English translation stage to improve medical concept extraction from French electronic health records. Hence, we will compare two approaches : :

1. translation-oriented : first, the French hospital record is translated to English, then a NER stage is performed in English (with arguably better language models and resources, overcoming the loss of quality brought by the automatic translation).
2. a classical, monolingual French NER performed on the hospital records.

In order to be able to compare the two pipelines, both systems will have a final mapping step from extracted terms to Unified Medical Language System (UMLS®) “Concept Unique Identifier” (CUI).
a baseline will be annotated directly with the Unified Medical Language System (UMLS®) “Concept Unique Identifier” (CUI). The UMLS corresponds to several knowledge sources including a metathesaurus where all the medical terms (drugs, symptoms, acts, etc.) are mapped to a “concept unique identifier” (CUI). CUIs are language-independent, which will allow a fair evaluation of the two

pipelines. In this approach, we will only focus on the *signs and symptoms* (e.g. fever, anuria, vomiting, etc.) and *diseases* (e.g. Crohn's disease, myocardial infarction, etc.) categories. The annotations of 200 texts were achieved by a medical doctor before the internship.

The student will have to compare different algorithms and their performances during her/his internship.

### Expected skills of the student

Background in programming science and programming skills, having a good knowledge in deep learning and/or natural language processing.

The intern will be given a “bonus” (around 600€/month) and a contribution to transport costs.

### Bibliographic references

[Vaswani2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

[Devlin2018] Delvin J., Changm. W., Leek. & Toutanovak (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human LanguageTechnologies - Proceedings of the Conference,1, 4171–4186.

[Lee2020] Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.

[ClinicalBERT] Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath. "Clinicalbert: Modeling clinical notes and predicting hospital readmission." *arXiv preprint arXiv:1904.05342* (2019).

[Neveol2018] Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. (2018). Clinical natural language processing in languages other than english: opportunities and challenges. Journal of biomedical semantics,9(1):1-3.

[Campos2017]Campos, L, Pedro,V., Couto, F. Impact of translation on named-entity recognition in radiology texts. Database (2017) Vol. 2017: article ID bax064;

[Suarez2021] Suárez-Paniagua, Víctor, Hang Dong, and Arlene Casey. "A multi-BERT hybrid system for Named Entity Recognition in Spanish radiology reports." *CLEF eHealth* (2021)

[Wajsbürt2021] Wajsbürt, Perceval, Arnaud Sarfati, and Xavier Tannier. "Medical concept normalization in French using multilingual terminologies and contextual embeddings." *Journal of Biomedical Informatics* 114 (2021): 103684.