CLINICAL INFORMATION EXTRACTION FROM MEDICAL REPORTS

PhD position - ANR IMAGE-TEXTE-AVC

Laboratoire ERIC - Université Lumière Lyon 2

Context

Ischemic stroke, responsible for approximately 87% of all strokes, is a major cause of mortality and long-term disability worldwide. In Europe, 1.1 million people suffer a stroke annually, with projections estimating an increase to 1.5 million cases per year by 2040 due to aging populations (Wafa, 2020). Patient care and monitoring generates numerous medical reports.

The "IMAGE-TEXTE-AVC" project, funded by the French national research agency, focuses on merging imaging and text data with the aim to improve stroke functional recovery prediction. It involves several research laboraties, namely CREATIS, ERIC, S2HEP, as well as a university hospital, HCL.

This PhD position is offered by the ERIC laboratory at Lyon 2 University. It is expected to start in January 2026.

Objectives

This PhD thesis aims to develop a system for extracting structured clinical information from medical reports. Key challenges include structural variability, temporal dependencies between documents, and distant supervision. We propose to leverage recent foundation models that require extensive pre-training and fine-tuning on French medical data. The main experiments will be based on data from the HIBISCUS-STROKE cohort, hosted at the Centre d'Investigation Clinique (CIC) de Lyon.

Tasks and planning

Corpus construction and preprocessing (T0-T8)

Raw clinical documents will be made available securely at CIC starting at T0. ERIC will construct a standardized corpus. Medical terms will be normalized using SNOMED CT and ICD-10 (Rodrigues, 2014), and named entities (e.g., diseases, treatments) will be detected and marked to preserve them during anonymization. To this end, we plan to leverage pre-trained, encoder-only, language models, e.g. fine-tuned versions of CamemBERT-bio (Touchent, 2024) or DrBERT (Labrak, 2023). Note that these models appear particularly suited, both of them being designed to understand French biomedical language.

Corpus anonymization (T6-T16)

To detect and remove identifying information in documents, we plan to improve upon eds- pseudo, a French clinical anonymization system developed and used at the AP-HP, the Greater Paris University Hospitals (Tannier, 2024). It is a hybrid system, combining a fine-tuned version of CamemBERT (Martin, 2020) with manual rules. One of the paths for improvement we've identified relates to the preservation of important clinical information, as mentioned earlier. Also, to improve precision, we intend to adopt multi-level sensitivity labeling to classify document sections based on confidentiality. CamemBERT will then apply adaptive masking, fully anonymizing patient identifiers, selectively masking clinical data while preserving medical terms, and leaving general content unchanged (Chaoui, 2024). This work will be validated through automated benchmarking against standard de-identification datasets, such as i2b2 2014 De-identification Challenge. This ensures effective de-identification while maintaining essential clinical information, balancing privacy compliance and data usability.

Structured clinical information extraction (T12-T36)

Manually extracted patient clinical data are reported in a table, which does not allow us to trace their origin in the documents. This distant supervision makes it impossible to train a classical extractive system (which would require token-level annotation). Moreover, inferring some clinical information actually requires cross-checking several documents. We therefore propose an original formulation for this task, in the form of a retrieval-augmented generation (RAG) task (Lewis, 2020). This means developing two mains components:

- a system capable of retrieving documents that are likely to mention a given clinical variable, accounting for the temporal dependency between documents;
- a system capable of generating the target clinical information from the retrieved documents in a precise and factual manner.

Clinical document retrieval

Retrieval systems usually assume documents are independent and rely on a siamese encoder, e.g. SBERT (Reimers, 2019), to measure a score between queries and documents. The independence assumption makes computing document embeddings straightforward, as they can be processed independently by the encoder. However, in our case this assumption doesn't hold because each document related to a patient should be viewed in light of previous documents. We will develop a system that combines a recurrent design with an encoder to sequentially encode documents while capturing the temporal relationship between them. What's more, rather than relying on some off-the-shelf pre-trained siamese encoder, we plan on adapting more suited pre-trained encoders such as CamemBERT-bio (Touchent, 2024) or DrBERT (Labrak, 2023).

Structured clinical information generation

Retrieved documents will serve as context for generating structured clinical information with a decoder-only language model. There are many available pre-trained, decoder-only, language models for language generation. On the one hand, there are a few decoders specialized for the English biomedical language, e.g. Meditron (Chen, 2023) or Med42 (Christophe, 2024). On the other hand, there are a few decoders specialized to answer to instructions in French, e.g. Vigogne (Huang, 2023). However, there is yet to be released a model to answer instruction in French biomedical language. The aforementioned models are available in two sizes, 7B and 70B, and are actually fine-tuned versions of either Llama-2 7B or Llama-2 70B (Touvron, 2023). Thanks to this, we'll be able to combine their competences by performing model merging following the "soup of models" approach (Wortsman, 2022) or more advanced approaches like DARE (Yu, 2024) to come up with a model suited to our needs.

Deliverables

- D1 – T6 Pipeline for corpus construction and preprocessing
- D2 – T16 Pipeline for anonymization, corpus ready for D1.3 in WP1 and WP2
- D3 – T28 Preliminary version of the RAG system for clinical information extraction (to be finalized in D4)
- D4 – T36 RAG system for clinical information extraction

To apply

The candidate will preferably hold a Master's degree (or equivalent) in Computer Science. To apply, send the following documents to adrien.guille@univ-lyon2.fr and julien.jacques@univ-lyon2.fr:

- Cover letter explaining the connection between your experiences and this subject
- Curriculum vitae
- Academic transcripts (for the last 3 years)

References

- Chen, Z. et al. (2023). MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. arXiv preprint arXiv:2311.16079.
- Christophe, C. et al. (2024). Med42-v2: A Suite of Clinical LLMs. arXiv preprint arXiv:2408.06142.
- Huang, B. (2023). Vigogne: French Instruction-following and Chat Models. GitHub repository.
- Labrak, Y. et al (2023). DrBERT: A robust pre-trained model in French for biomedical and clinical domains. 61st Annual meeting of the association for computational linguistics.
- Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.
- Martin, L. et al. (2020). CamemBERT: a Tasty French Language Model. 58th Annual Meeting of the

Association for Computational Linguistics.

- Reimers, N. et al. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. International Joint Conference on Natural Language Processing.
- Rodrigues J.M. et al. (2014). ICD-11 and SNOMED CT Common Ontology: circulatory system. Stud Health Technol Inform. 205:1043-7.
- Tannier, X. et al. (2024). Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse. Methods of Information in Medicine, 63(1-02), 21-34.
- Touchent, R. et al. (2024). CamemBERT-bio: Leveraging Continual Pre-training for Cost-Effective Models on French Biomedical Data. Joint International Conference on Computational Linguistics, Language Resources and Evaluation
- Touvron, H. et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288.
- Wortsman, M. (2022). Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. International Conference on Machine Learning.
- Yu, L. et al. (2024). Language models are super mario: Absorbing abilities from homologous models as a free lunch. In 41rst International Conference on Machine Learning.