Natural Language Processing Models for Knowledge Discovery in Medical Texts

Eduardo Grande

Department of Software and Computing Systems, University of Alicante, Spain

Abstract

Electronic Health Records (EHRs) contain vast amounts of valuable information about patients' diseases, diagnoses, or medications, mostly in an unstructured format. Recently, Large Language Models (LLMs), particularly generative models, have gained popularity due to their remarkable capabilities. This PhD thesis aims to harness the power of these models for the medical field, specifically for knowledge extraction tasks. The goal is to adjust NLP models to extract critical insights from EHRs and other medical texts. However, one of the main challenges is the limited availability of publicly accessible medical data, especially annotated datasets in languages other than English. In order to adjust the models, the thesis explores various adaptation techniques, including prompt-tuning or continual pre-training to enhance the models' ability to process medical information effectively. Additionally, it evaluates different LLM architectures to determine the most suitable for medical knowledge extraction. Innovative strategies like LoRA (Low-Rank Adaptation) and QLoRA (Quantized LoRA) are also investigated to try to improve efficiency. The outcomes of this research hold the potential to significantly enhance healthcare delivery and and help practitioners quickly understand patient data.

Keywords

Natural Language Processing, Knowledge Discovery, Large Language Models, Electronic Health Records

1. Justification of the proposed research

Electronic Health Records (EHRs) are text documents that contain all medical information about patients. They maintain extensive records regarding patients' health, including biographical information, disease history, symptoms, diagnoses, medication prescriptions, and other relevant patient information.

The information contained in EHRs is often unstructured and written in free form, meaning there are no standardized formats for recording patient data. Only a few documents or pieces of information, such as lab tests that include blood measurements, can be considered structured. This lack of structure in the majority of documents makes it challenging to detect, extract, and classify the highly valuable information they contain.

The aim of the present doctoral work is to create Natural Language Processing (NLP) models that can automatically extract knowledge, not only from EHRs but also from other unstructured medical texts, by training them using collected publicly available corpora. The extracted knowledge can be extremely valuable for predicting patients' diseases, conducting population-based public health studies, or analyzing the impact of a disease on a population group.

Creating such models requires annotated data. In the medical field, the scarcity of these resources is a significant challenge. Most published works use private datasets obtained through specific agreements between researchers and medical institutions.

As explained by Li et al. [1], most works use private datasets, and those using public ones mainly rely on the MIMIC datasets [2, 3, 4] or the i2b2/n2c2 datasets. Both datasets contain annotated health documents in English that have been extensively used for developing medical NLP models.

When seeking data in other languages, such as Spanish, the difficulty increases significantly. A good example demonstrating this scarcity is the work done by Carrino et al. [5], which aimed to create a biomedical and medical language model for Spanish. The data they used primarily consisted of scientific

Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.

eduardo.grande@ua.es (E. Grande)
0000-0002-4894-3943 (E. Grande)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

publications, patents, or Wikipedia articles rather than EHRs. The only corpus composed of medical cases was not publicly available.

The strategy of using a mixture of documents, rather than exclusively using medical cases for training NLP models, is common. Many recent works resort to this approach to mitigate the scarcity of valuable data, such as EHRs. These documents often include publications available in PubMed, medical-related documents crawled from the internet, or even the UMLS ontology, as explained by Wornow et al. [6].

Currently, NLP researchers predominantly use Large Language Models (LLMs) to solve common field tasks. As shown in various reviews [1, 6, 7], most research works focus on using BERT models, or regarding the newest LLMs, most of them use GPT models (via the OpenAI API).

To date, almost no work has been done using the newest LLM models, such as LLaMa-3 or Mistral. Exploring these newly released models could yield promising results for knowledge extraction tasks.

In the following sections, the details of the proposal for this doctoral work will be described.

2. Background and related work

In the last two years, several surveys have been published concerning the use of LLMs in the medical field.

- Li et al. [1] presented a survey based on 329 related papers, explaining the evolution of LLMs, their different architecture types, what an EHR is, publicly available EHR datasets, methods of fine-tuning the models, and research trends in different NLP tasks, such as Named Entity Recognition (NER) and Information Extraction.
- Wornow et al. [6] conducted a survey about the use of Functional Models (FMs—models capable of performing many different tasks) in the medical field. They explained the different FMs specifically created for medical purposes (medical models, EHR models, etc.), their benefits, available public medical data, how the models have been evaluated, and future trends.
- Huang et al. [7] provided a comprehensive survey presenting various aspects of using LLMs in the medical field. They showed the different applications for which LLMs have been used, such as data processing (including the NER task) and various models. They also extensively explained the metrics and benchmarks used to measure the performance of these models for specific tasks.

Regarding related work, several specific relevant works are highlighted for this research:

- Guevara et al. [8] used LLMs to extract Social Determinants of Health (SDoH), which are conditions surrounding patients that affect their health. Their work is particularly interesting because they employed different approaches and models to achieve the goal of extracting SDoH from EHRs. They also used techniques such as LoRA [9] and PEFT [10] for efficiently adjusting the models.
- de la Iglesia et al. [11] created a corpus composed of 1038 Electronic Clinical Narratives (ECNs) written in Spanish. It contains annotations of seven different types, referring to illnesses, medications, or treatments. This corpus could be used as a base to adjust and test an LLM for detecting information and classifying it into different categories.
- Ahsan et al. [12] extracted evidence from EHRs using LLMs. By prompting models like Flan-T5 XXL or Mistral multiple times, they obtained evidence and corresponding explanations from the models, where the prompts contained information from given EHRs. This approach of directly prompting LLMs without fine-tuning could be explored as a baseline to compare with results obtained from fine-tuned models.
- Agrawal et al. [13] aimed to extract information from medical texts without fine-tuning models. They used GPT-3 for information extraction, focusing on prompting the model and parsing the output to get structured results, such as arrays of strings. Parsing the results of generative models is crucial for extracting complex information accurately.

- Gallego et al. [14] presented a pipeline for medical entity linking, emphasizing the use of standardized terminologies like UMLS or SNOMED-CT. These ontologies could be considered when performing knowledge extraction from medical documents, as linking information with the corresponding codes would normalize the data, facilitating integration with information retrieved by other systems.

These works have been instrumental in understanding the current trends in the field of knowledge extraction from medical texts. Given the recent popularity of LLMs, most of the surveys are quite recent, indicating that this research field is currently booming.

In the following sections, the proposed research and planned experiments will be described. These plans have been formulated after reviewing the mentioned surveys and similar experiments conducted by other researchers.

3. Description of the proposed research

This work plans to use LLMs for knowledge extraction, primarily exploring the NER task, in the medical field. Figure 1 illustrates how field-related data, such as EHRs, and knowledge bases, such as SNOMED-CT or UMLS, can be used to train an LLM for extracting knowledge from plain text. This enables the generation of structured information, such as text annotations for disease or pharmacological substance mentions.

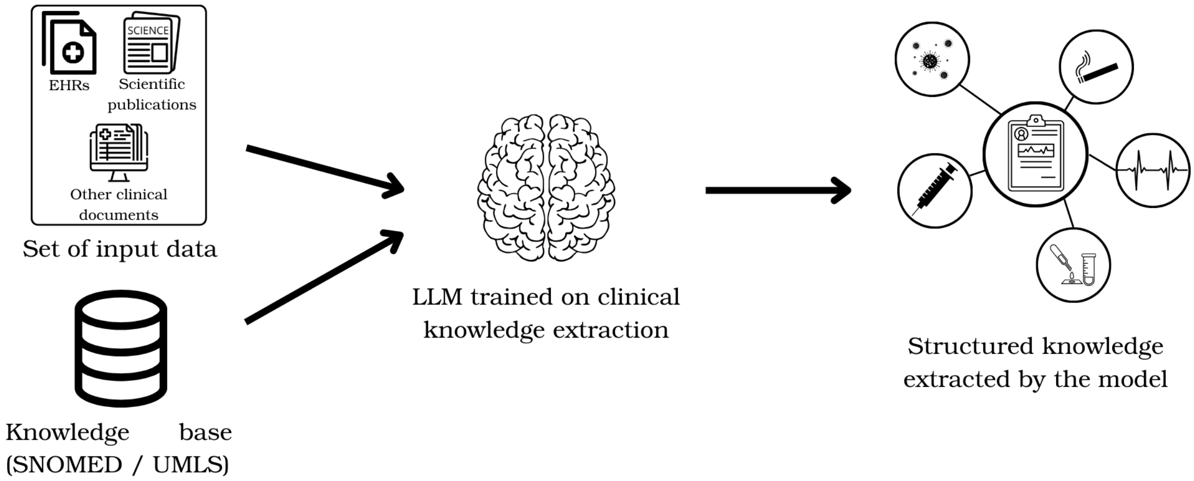
SCIENCE
EHRs
Scientific publications
Other clinical documents
Set of input data
Knowledge base (SNOMED / UMLS)
LLM trained on clinical knowledge extraction
Structured knowledge extracted by the model

Figure 1: Overview of the main goal of the doctoral work

As explained previously, almost no work has explored the use of currently published LLMs such as LLaMA-3, Mistral, or Claude 3. The main research will be exploring how to adapt these models to the knowledge extraction task. As outlined in the justification section, the majority of the available data is in English, with a limited number of corpora in Spanish. Therefore, the languages used for data extraction will primarily depend on the availability of data.

Many adaptation techniques can be explored, but the main one to explore is prompt-tuning or adapter-tuning. A set of templates can be defined, which can be applied over several corpus, similar to how Google does with FLAN [15]. By doing this, the models may learn specifically the task we want them to, transitioning from a generalist position towards a specific one.

Not only these adapting techniques are going to be explored. Continual pre-training techniques could also be explored. Several plain text datasets from the medical field could be collected and then continue the training of an LLM to incorporate new target vocabulary.

Regarding the data needed to undertake the work, we can distinguish between the necessity of data for continual pre-training (this process just needs plain text data) and the data for adjusting the models (annotated data).

For the non-annotated data, an exploration of available datasets and corpus can be done. Collecting new not-annotated data is easier than the annotated one, so collecting new data by crawling websites can be done, always taking in mind the rights and licenses of the crawled pages to avoid violating any rules or applicable laws.

On the other side, for the annotated data needed, in a first step the public available datasets are going to be explored. There's a need to explore if their data is within the desired scope and if their annotations are significant. Foreseeably, these data will be scarce, so more data will have to be collected. If more health records want to be obtained, alliances between medical institutions such as public hospitals could be arranged. Other sources of these data could be asking researches who have already use datasets of EHRs for sharing them under license and terms of use.

4. Methodology and Proposed Experiments

The methodology to achieve the established goals is based on achieving milestones planned for the three years of the doctoral research.

4.1. Year 1

- **State-of-the-art Study**: Begin with an extensive review of current techniques used for creating LLMs specialized in knowledge extraction.
- **Initial Experiments**: Start experimenting with some of these LLMs. This involves identifying and obtaining publicly available medical data to apply and test these models.

4.2. Year 2

- **Continued Research on LLMs**: Deepen the research on knowledge extraction LLMs and continue searching for medical data. This may involve forming agreements with hospitals or other healthcare institutions to access non-public datasets.
- **Data Storage and Representation**: Once the extracted knowledge is stored, representation methods should be researched to illustrate the knowledge extracted by the created models.

4.3. Year 3

- **Completion of Research**: Finalize the research activities initiated in the previous years.
- **Publications and Thesis Writing**: Complete pending scientific publications and start writing the doctoral thesis.

Regarding the proposed experiments, they will evolve based on the results obtained from initial trials. Some of the planned experiments include:

Prompt-Tuning LLM: Choose an LLM and perform prompt-tuning. This involves exploring different tuning methods, optimal hyperparameters, optimizers, and evaluation metrics.

In-Context Learning: Apply the prompt-tuning technique to an LLM. This involves creating and testing various sets of prompts. In-context learning uses textual inputs (prompts) to fine-tune an LLM. Different prompt construction strategies, such as one-shot, few-shot, and chain-of-thought, will be tested.

LoRA and QLoRA: Test the LoRA (Low-Rank Adaptation) and QLoRA (Quantized Low-Rank Adaptation) techniques for prompt-tuning. LoRA [9] involves training only a reduced set of model parameters, specifically low-rank representations, to reduce training time and GPU usage. QLoRA [16] extends LoRA by representing model weights in 4-bit precision, further reducing memory usage and improving efficiency.

Continual Pre-training: Investigate how the performance of an LLM improves with increased exposure to medical vocabulary. Conduct continual pre-training of an LLM with medical-specific vocabulary to enhance its performance in knowledge extraction tasks.

Use of Synthetic Generated Data: Address the scarcity of publicly available medical datasets by generating synthetic annotated data. Combine synthetic data with real data to increase the number of training examples, potentially improving model performance.

As the research progresses and new state-of-the-art publications are reviewed, additional experiments will be conceived. These may involve employing new techniques, testing new models, and exploring different methods for model adjustment. This dynamic approach ensures the research remains at the cutting edge of advancements in NLP and medical data extraction.

5. Specific Issues of Research to be Discussed

In this section, and once the doctoral work has been explained, various questions are posed for further discussion.

Q1. Sources of medical data To achieve the objectives of the present project, medical data is needed. Across the document, several ways of obtaining these data have been explained. Are these ways of obtaining the data useful? Could other ways be explored? Are there public medical datasets available that have not been mentioned?

Q2. Which LLM architecture is better to use? In recent years, most state-of-the-art LLMs have a decoder-only architecture, while the most-used LLM in the medical field has been BERT. Which architecture is better for knowledge extraction in the medical field? How much can the architecture of a model influence the resultant performance?

Q3. What is the better way of adjusting the LLMs to the knowledge extraction task? Throughout the work, different ways of adjusting models have been presented, with prompt-tuning being the most common. What is the best way of performing prompt-tuning on an LLM? Are there any other techniques besides LoRA and QLoRA?

The resultant discussion produced by the presented questions, as well as other aspects that could arise, would be enriching for the PhD thesis.

6. Acknowledgements

The publication is part of the grant PRE2022-101573, funded by MCIN/AEI/10.13039/501100011033 and the ESF+.

References

[1] L. Li, J. Zhou, Z. Gao, W. Hua, L. Fan, H. Yu, L. Hagen, Y. Zhang, T. L. Assimes, L. Hemphill, et al., A scoping review of using large language models (llms) to investigate electronic health records (ehrs), arXiv preprint arXiv:2405.03066 (2024).

[2] M. Saeed, C. Lieu, G. Raber, R. G. Mark, Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring, in: Computers in cardiology, IEEE, 2002, pp. 641–644.

[3] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific data 3 (2016) 1–9.

[4] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al., Mimic-iv, a freely accessible electronic health record dataset, Scientific data 10 (2023) 1.

[5] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: https://aclanthology.org/2022.bionlp-1.19. doi:10.18653/v1/2022.bionlp-1.19.

[6] M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M. A. Pfeffer, J. Fries, N. H. Shah, The shaky foundations of clinical foundation models: a survey of large language models and foundation models for emrs, arXiv preprint arXiv:2303.12961 (2023).

[7] Y. Huang, K. Tang, M. Chen, A comprehensive survey on evaluating large language model applications in the medical industry, arXiv preprint arXiv:2404.15777 (2024).

[8] M. Guevara, S. Chen, S. Thomas, T. L. Chaunzwa, I. Franco, B. H. Kann, S. Moningi, J. M. Qian, M. Goldstein, S. Harper, et al., Large language models to identify social determinants of health in electronic health records, NPJ digital medicine 7 (2024) 6.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[10] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, Peft: State-of-the-art parameter-efficient fine-tuning methods, https://github.com/huggingface/peft, 2022.

[11] I. de la Iglesia, M. Vivó, P. Chocrón, G. de Maeztu, K. Gojenola, A. Atutxa, An open source corpus and automatic tool for section identification in spanish health records, Journal of Biomedical Informatics 145 (2023) 104461.

[12] H. Ahsan, D. J. McInerney, J. Kim, C. Potter, G. Young, S. Amir, B. C. Wallace, Retrieving evidence from ehrs with llms: Possibilities and challenges, arXiv preprint arXiv:2309.04550 (2023).

[13] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, D. Sontag, Large language models are few-shot clinical information extractors, arXiv preprint arXiv:2205.12689 (2022).

[14] F. Gallego, G. López-García, L. Gasco-Sánchez, M. Krallinger, F. J. Veredas, Clinlinker: Medical entity linking of clinical concept mentions in spanish, arXiv preprint arXiv:2404.06367 (2024).

[15] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, arXiv preprint arXiv:2109.01652 (2021).

[16] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, Advances in Neural Information Processing Systems 36 (2024).