

Chatbot based on clinical literature for decision support

Irene Sánchez-Montejo¹ , Carlos Telleria-Oriols¹  and Raquel Trillo² 

¹ Instituto Aragonés de Ciencias de la Salud, Zaragoza, Spain
isanchez.iacs@aragon.es
ctelleria@aragon.es

² Universidad de Zaragoza, Zaragoza, Spain
raqueltl@unizar.es

Abstract. Clinical practice guidelines try to provide the state-of-the-art in diagnostic and treatment methods for each disease, by a systematic review of the scientific evidence, but it can be difficult to keep up to date in a context of healthcare in constant evolution. Improvements in Deep Learning and Natural Language Processing have allowed to perform multiple applications, such as conversational agents (chatbots or virtual assistants), that are designed to simulate a human conversation. Language models behind these systems are able to analyze a huge collection of documents with unstructured data and extract the essential information from each one, easing the fast consultation of guidelines by practitioners and patients. This article provides an approach of a thesis plan to analyze different techniques and language models, and develop a chatbot able to answer according to clinical practice guidelines and other high-quality biomedical literature in a real-time decision support system for healthcare professionals, patients, and caregivers.

Keywords: Clinical practice guidelines · Natural Language Processing · Chatbots · Decision Support System · Deep Learning · Large Language Models

1 Introduction

Lately, natural language processing (NLP) has experienced significant advances thanks to the progress in machine learning (ML) algorithms, cheaper and more powerful computing infrastructures and the availability of huge digitized corpus. This technology allows a computer to understand natural texts while keeping its semantic meaning to perform multiple tasks, such as: text translation [1], sentiment analysis [2] or text summarization [3] among others. This has also enabled the creation of chatbots (virtual assistants).

A chatbot is a computer program designed to simulate a human conversation, allowing users to interact with it through text or voice messages. Chatbots are based on the use of natural language processing (NLP), which allows them to

understand and answer intelligently to queries made by a user, thus reproducing a conversation. However, their capabilities are not limited to giving simple answers to frequently asked questions. They can also be used to process and analyze complex texts, such as large volumes of data (e.g. GPT-4 developed by OpenAI [4]). Chatbots have been applied in many different fields, including marketing, economy, medicine or education [5].

In the healthcare field, these systems can be applied to process medical contents and help in the task of extracting relevant information from large and complex documents. There are some chatbots yet available to give some help proposing a diagnostic to a patient according to his or her symptoms [6], thus reducing the communication gap between patients and clinicians, and providing specific treatments. Training chatbots with updated clinical practice guidelines (CPG) and relevant scientific texts can benefit both clinicians and patients as support for decision making in a clinical situation. Becker et al. [7] concluded that the use of NLP has a positive impact on improving the clinical decision support system, and translates into an improvement in the quality of care.

Even more, NLP techniques and Language Models (LM) can be a very useful tool for CPG writers, in the process of reading, processing and assessing scientific and clinical relevance of biomedical articles in the context of evidence based medicine. But all these uses of natural language technologies must be objectively evaluated and measured with regard to their efficiency and usefulness.

2 Our proposal

With the objective of evaluating and proposing specific solutions in the healthcare context based on NLP and Large Language Models (LLM), we have detected some challenges in the use of LLM in this field, which do not appear in other domains, such as: the use of abbreviations that can have multiple meanings, so it is necessary to study the context first; structured and unstructured data, as Electronic Health Record (EHR) contain some free fields, that need to be treated by NLP algorithms according to the common data model OMOP [8] to ensure structured data; temporal events where it is necessary to identify the temporal sequence of each event to determine a correct diagnosis or treatment; or the use of negative sentences, among others.

Some previous works have studied how to build a chatbot trained on a medical corpus to provide information to the final user, using the NLP architecture that best fits the problem to be solved.

We plan to train LLMs using CPG and other high-quality biomedical literature written or translated into Spanish. We will then choose the LLM that best fits according to the defined metrics. Afterward, we will build chatbots to be used in real clinical contexts to evaluate their performance in terms of speed, resource consumption, accuracy and usefulness to make better clinical decisions.

Kim et al. [9] used a pre-trained BERT model and performed fine-tuning to create a model capable of answering medical questions, to infer which type of medical specialty was in line with certain symptoms. Fine-tuning is a technique to leverage the knowledge of an existing model to maintain the conversational style, while adding new information about the topic to be exploited, which improves the performance of the model. But, this work was limited by the data their model was trained with

Many of the aforementioned models have been trained using English as their core language. Nevertheless, there are some available models trained with Spanish data, such as BETO [10], which was the first BERT model built with Spanish as its core language. Later, BSC³ researchers developed MarIA [11], a set of Spanish language models. Gutierrez-Fandiño et al. [11] fine-tuned two existing models, RoBERTa as an encoder model, and GPT2 as a generative models. Both models were trained with a Spanish corpus generated by themselves, limited by aspects such as the length of the context and the size of the dataset.

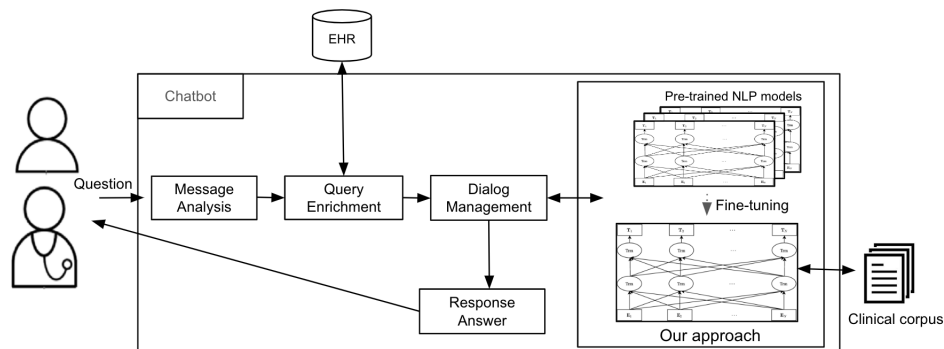


Fig. 1. Overview of our approach.

Since LLMs and NLP techniques are constantly evolving, the first goal of our work is to gather different LMs, analyze them and compare their performance, and try to adapt them according to the challenges mentioned above. In our thesis, we plan to use a large, complete clinical dataset to train a more robust chatbot and use fine-tuning on different models, such as RoBERTa [12], LLaMa [13] or OpenAI GPT [4], to retain their prior knowledge, or develop new models based on these architectures. The final models will be used to develop different chatbots that will be tested both by doctors, patients and caregivers, given that there are different guidelines issues for different targets, and all of them can take benefit from this system. The purpose of this chatbot is to answer various queries, including providing treatments or helping in the diagnosis. Besides, when

³ <https://www.bsc.es/es>

available, the model uses the patient's EHR to enrich the answer, promoting personalized medicine, which will be a novel approach, see Fig 1.

However, due to the inherent differences between the models, it is necessary to define the metrics that the models will be compared under, such as the response time, resource consumption, feasibility to be integrated in actual health-care systems, the quality and accuracy of the chatbot's answer, and whether it is appropriate according to the ground truth, which also needs to be defined. So, a second goal in our work is to define this metrics and develop a methodology to evaluate the quality and utility of chatbots in a medical context.

An additional open line in this thesis is to use LLM models to assess the evidence level for any new biomedical article according to scientific criteria, as to include that article and its conclusions into the corresponding CPG, enriching the guidelines in a continuous improvement process.

3 Conclusion

To sum up, starting from the well-known power of NLP techniques and LLMs in many real life domains, and specifically in healthcare domain, and the possibility of having documents of proven scientific quality, like CPG and the patient EHRs. We aim to analyze and compare different models in their application to clinical support decision systems based on chatbots, to define objective and reproducible metrics to evaluate the quality and utility of these chatbots, and explore the feasibility of using LM as a way to facilitate the work of guidelines writers.

Acknowledgements Project partially funded by PID2020-113037RB-I00 / AEI / 10.13039/501100011033.

References

1. Bird, S., Klein, E., and Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."
2. Yi, J., Nasukawa, T., Bunesco, R., and Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. The 3rd IEEE International Conference on data mining, IEEE.
3. Merchant, K., and Pande, Y. (2018). Nlp based latent semantic analysis for legal text summarization. In 2018 International conference on advances in computing, communications and informatics (ICACCI). IEEE.
4. OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.
5. Adamopoulou, E., and Moussiades, L. (2020). An overview of chatbot technology. In Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, 2020, Proceedings, Part II 16. Springer International Publishing.
6. Xu, L., Sanders, L., Li, K., and Chow, J. C. (2021). Chatbot for health care and oncology applications using artificial intelligence and machine learning: Systematic review. *JMIR cancer*, 7(4), e27850.
7. Becker, M., Kasper, S., Böckmann, B., Jöckel, K. H., and Virchow, I. (2019). Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. *International journal of medical informatics*.
8. Hripesak, G., Duke, JD., Shah, NH., Reich, CG., ... and Ryan, PB. (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform.*
9. Kim, Y., Kim, J. H., Kim, Y. M., ... and Joo, H. J. (2023). Predicting medical specialty from text based on a domain-specific pre-trained BERT. *International Journal of Medical Informatics*.
10. Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*.
11. Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., ... and Villegas, M. (2021). Maria: Spanish language models. arXiv preprint arXiv:2107.07253.
12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., ... and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
13. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... and Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

